Yuhang Yang
Maode Ma
Baoxiang Liu (Eds.)

Information Computing and Applications

4th International Conference, ICICA 2013
Singapore, August 2013
Revised Selected Papers, Part I

Part 1

Springer

Communications
in Computer and Information Science 391

Editorial Board

Yuhang Yang   Maode Ma   Baoxiang Liu (Eds.)

Information Computing and Applications

4th International Conference, ICICA 2013
Singapore, August 16-18, 2013
Revised Selected Papers, Part I

Springer

Volume Editors

Yuhang Yang
Shanghai Jiao Tong University
800 Dongchuan Road, Shanghai 200240, China
E-mail: yhyangsjtu@gmail.com

Maode Ma
Nanyang Technological University
School of Electrical and Electronic Engineering
Nanyang Avenue, Singapore 639798
E-mail: emdma@ntu.edu.sg

Baoxiang Liu
Hebei United University, College of Science
46 West Xinhua Road, Tangshan, China
E-mail: liubx5888@126.com

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

Printed on acid-free paper

Preface

Welcome to the proceedings of the 4th International Conference on Information Computing and Applications (ICICA 2013), which was held during August 16–18, 2013, in Singapore.

As future-generation information technology, information computing and applications become specialized, information computing and applications including hardware, software, communications, and networks are growing with ever increasing scale and heterogeneity, and becoming overly complex. The complexity is becoming more critical along with the growing applications. To cope with the growing and computing complexity, information computing and applications focus on intelligent, self-manageable, scalable computing systems and applications to the maximum extent possible without human intervention or guidance.

With the rapid development of information science and technology, information computing has become the third approach of scientific research. Information computing and applications is the field of study concerned with constructing intelligent computing, mathematical models, numerical solution techniques, and using computers to analyze and solve natural scientific, social scientific, and engineering problems. In practical use, it is typically the application of computer simulation, intelligent computing, Internet computing, pervasive computing, scalable computing, trusted computing, autonomy-oriented computing, evolutionary computing, mobile computing, applications and other forms of computation to problems in various scientific disciplines and engineering. Information computing and applications is an important underpinning for techniques used in information and computational science and there are many unresolved problems worth studying.

The ICICA 2013 conference provided a forum for engineers and scientists in academia, industry, and government to address the most innovative research and development including technical challenges and social, legal, political, and economic issues, and to present and discuss their ideas, results, work in progress and experience on all aspects of information computing and applications.

There was a very large number of paper submissions (665). All submissions were reviewed by at least three Program or Technical Committee members or external reviewers. It was extremely difficult to select the presentations for the conference because there were so many excellent and interesting submissions. In order to allocate as many papers as possible and keep the high quality of the conference, we finally decided to accept 126 papers for presentations, reflecting an 18.9% acceptance rate; 63 papers were included in this volume. We believe that all of these papers and topics not only provided novel ideas, new results, work in progress and state-of-the-art techniques in this field, but also stimulated future research activities in the area of information computing and applications.

The exciting program for this conference was the result of the hard and excellent work of many individuals, such as the Program and Technical Committee members, external reviewers, and publication chairs under a very tight schedule. We are also grateful to the members of the local Organizing Committee for supporting us in handling so many organizational tasks, and to the keynote speakers for accepting to come to the conference with enthusiasm. Last but not least, we hope you enjoy the conference proceedings.

September 2013
Yuhang Yang
Maode Ma
Baoxiang Liu

Organization

ICICA 2013 was organized by Nanyang Technological University, Hebei Scene Statistical Society, IEEE Education Chapter of Singapore, and sponsored by the National Science Foundation of China, Republic Polytechnic of Singapore, Hunan Institute of Engineering, Yanshan University, and Northeastern University at Qinhuangdao. It was held in cooperation with *Communications in Computer and Information Science* (CCIS) of Springer.

Executive Committee

Honorary Chair

Jun Li	Hebei United University, China

General Chairs

Yanchun Zhang	University of Victoria, Australia
Baoxiang Liu	Hebei United University, China
Maode Ma	Nanyang Technological University, Singapore

Program Chairs

Chunfeng Liu	Hebei United University, China
Fengbo Hou	Chengde Petroleum College, China
Leizhen Wang	Northeastern University at Qinhuangdao, China
Jiang Lijun David	IEEE Education Chapter of Singapore, Republic Polytechnic, Singapore

Local Arrangements Chairs

Jincai Chang	Hebei United University, China
Aimin Yang	Hebei United University, China

Steering Committee

Qun Lin	Chinese Academy of Sciences, China
Lorna Uden	Staffordshire University, UK
Nadia Nedjah	State University of Rio de Janeiro, Brazil
Rafal Kapelko	Wroclaw University of Technology, Poland
Wende Zhong	Nanyang Technological University, Singapore
Nadia Nedjah	State University of Rio de Janeiro, Brazil

Gaoxi Xiao	Nanyang Technological University, Singapore
Bingsheng He	Nanyang Technological University, Singapore
Lorna Uden	Staffordshire University, UK
Yiming Chen	Yanshan University, China
Xiangdong Song	Yanshan University, China
Changcun Li	Hebei United University, China
Zhijiang Wang	Hebei United University, China
Rongbo Zhu	South-Central University for Nationalities, China
Guohuan Lou	Hebei United University, China
Xiaoqi Li	Northeastern University at Qinhuangdao, China
Dianchuan Jin	Hebei United University, China
Xiaosheng Wang	Hebei University of Engineering, China
Shenggui Duan	Shijiazhuang University of Economics, China
Xiangdong Song	Yanshan University, China
Fengbo Hou	Chengde Petroleum College, China
Xiaomin Wang	Northeastern University at Qinhuangdao, China

Publicity Chairs

Aimin Yang	Hebei United University, China
Xilong Qu	Hunan Institute of Engineering, China

Publication Chairs

Maode Ma	Nanyang Technological University, Singapore

Financial Chair

Jincai Chang	Hebei United University, China

Local Arrangements Committee

Lihong Li	Hebei United University, China
Shaohong Yan	Hebei United University, China
Yamian Peng	Hebei United University, China
Lichao Feng	Hebei United University, China
Yuhuan Cui	Hebei United University, China

Secretariat

Kaili Wang	Hebei United University, China
Jingguo Qu	Hebei United University, China
Yafeng Yang	Hebei United University, China

Program/Technical Committee

Yuan Lin	Norwegian University of Science and Technology, Norway
Yajun Li	Shanghai Jiao Tong University, China
Yanliang Jin	Shanghai University, China
Mingyi Gao	National Institute of AIST, Japan
Yajun Guo	Huazhong Normal University, China
Haibing Yin	Peking University, China
Jianxin Chen	University of Vigo, Spain
Miche Rossi	University of Padova, Italy
Ven Prasad	Delft University of Technology, The Netherlands
Mina Gui	Texas State University, USA
Nils Asc	University of Bonn, Germany
Ragip Kur	Nokia Research, USA
On Altintas	Toyota InfoTechnology Center, Japan
Suresh Subra	George Washington University, USA
Xiyin Wang	Hebei United University, China
Dianxuan Gong	Hebei United University, China
Chunxiao Yu	Yanshan University, China
Yanbin Sun	Beijing University of Posts and Telecommunications, China
Guofu Gui	CMC Corporation, China
Haiyong Bao	NTT Co., Ltd., Japan
Xiwen Hu	Wuhan University of Technology, China
Mengze Liao	Cisco China R&D Center, China
Yangwen Zou	Apple China Co., Ltd., China
Liang Zhou	ENSTA-ParisTech, France
Zhanguo Wei	Beijing Forestry University, China
Hao Chen	Hu'nan University, China
Lilei Wang	Beijing University of Posts and Telecommunications, China
Xilong Qu	Hunan Institute of Engineering, China
Duolin Liu	ShenYang Ligong University, China
Xiaozhu Liu	Wuhan University, China
Yanbing Sun	Beijing University of Posts and Telecommunications, China
Yiming Chen	Yanshan University, China
Hui Wang	University of Evry, France
Shuang Cong	University of Science and Technology of China, China
Haining Wang	College of William and Marry, USA
Zengqiang Chen	Nankai University, China
Dumisa Wellington Ngwenya	Illinois State University, USA

Hu Changhua   Xi'an Research Institute of Hi-Tech, China
Juntao Fei   Hohai University, China
Zhao-Hui Jiang   Hiroshima Institute of Technology, Japan
Michael Watts   Lincoln University, New Zealand
Tai-hon Kim   Defense Security Command, Korea
Muhammad Khan   Southwest Jiaotong University, China
Seong Kong   The University of Tennessee USA
Worap Kreesuradej   King Mongkuts Institute of Technology
                               Ladkrabang, Thailand
Uwe Kuger   Queen's University of Belfast, UK
Xiao Li   CINVESTAV-IPN, Mexico
Stefa Lindstaedt   Division Manager Knowledge Management,
                               Austria
Paolo Li   Polytechnic of Bari, Italy
Tashi Kuremoto   Yamaguchi University, Japan
Chun Lee   Howon University, Korea
Zheng Liu   Nagasaki Institute of Applied Science, Japan
Michiharu Kurume   National College of Technology, Japan
Sean McLoo   National University of Ireland, Ireland
R. McMenemy   Queen's University Belfast, UK
Xiang Mei   The University of Leeds, UK
Cheol Moon   Gwangju University, Korea
Veli Mumcu   Technical University of Yildiz, Turkey
Nin Pang   Auckland University of Technology,
                               New Zealand
Jian-Xin Peng   Queen's University of Belfast, UK
Lui Piroddi   Technical University of Milan, Italy
Girij Prasad   University of Ulster, UK
Cent Leung   Victoria University of Technology, Australia
Jams Li   University of Birmingham, UK
Liang Li   University of Sheffield, UK
Hai Qi   University of Tennessee, USA
Wi Richert   University of Paderborn, Germany
Meh shafiei   Dalhousie University, Canada
Sa Sharma   University of Plymouth, UK
Dong Yue   Huazhong University of Science and
                               Technology, China
YongSheng Ding   Donghua University, China
Yuezhi Zhou   Tsinghua University, China
Yongning Tang   Illinois State University, USA
Jun Cai   University of Manitoba, Canada
Sunil Maharaj Sentech   University of Pretoria, South Africa
Mei Yu   Simula Research Laboratory, Norway
Gui-Rong Xue   Shanghai Jiao Tong University, China
Zhichun Li   Northwestern University, China

Lisong Xu	University of Nebraska-Lincoln, USA
Wang Bin	Chinese Academy of Sciences, China
Yan Zhang	Simula Research Laboratory and University of Oslo, Norway
Ruichun Tang	Ocean University of China, China
Wenbin Jiang	Huazhong University of Science and Technology, China
Xingang Zhang	Nanyang Normal University, China
Qishi Wu	University of Memphis, USA
Jalel Ben-Othman	University of Versailles, France

Table of Contents – Part I

Internet Computing and Applications

Engineering Management and Applications

Intelligent Computing and Applications

Control Engineering and Applications

Cloud and Evolutionary Computing

Knowledge Management and Applications

Computational Statistics and Applications

Table of Contents – Part II

Internet Computing and Applications

Engineering Management and Applications

Intelligent Computing and Applications

Business Intelligence and Applications

Knowledge Management and Applications

Information Management System

Computational Statistics and Applications

Game-Based Spectrum Allocation Algorithm in Cognitive Radio Networks

Qiufen Ni, Rongbo Zhu[*], Zhenguo Wu, and Yongli Sun

College of Computer Science, South-Central University for Nationalities 708 Minyuan Road,
Wuhan 430074, China
`rongbozhu@gmail.com`

Abstract. In order to solve the severe waste of spectrum resources due to the way of fixed spectrum allocation and the situation that the scarcity of spectrum restricts the development of wireless communication system, this paper proposed an efficient cognitive users spectrum allocation plan based on the evolutionary game, which depicts the dynamic behaviors choosing to purchase spectrum of primary users among a group of cognitive users in a group a, which focuses on the case has two primary users and thirty cognitive users in cognitive radio networks. Detailed simulation results show that the proposed scheme can improve the spectrum utilization and maximize the utility of the cognitive users and primary users.

1 Introduction

With the constantly emerging of a variety of new wireless communications technologies, and the rise of all kinds of wireless business, the scarcity of the spectrum resource has become a bottleneck constraining the development of the next generation wireless communication system [1]. Dynamic spectrum allocation is a key technology in cognitive radio, which adjusts the occupied band and transmission power of cognitive users dynamically through the perception results of the state of the channel, under the premise of without affecting the communication of the primary users to improve the spectrum utilization rate as much as possible [2].

Currently there are a lot of researches on spectrum allocation. The self-organizing Dynamic Spectrum Management (SO-DSM) was proposed in ad hoc cognitive radio network [3], which was inspired by the human brain and designed a channel activity patterns that cognitive radio networks to extract and remember the primary network and other cognitive radio network and the sub channels without the use of primary users in the recent to use the self-organizing map (SOM) technology to assign the channel temporarily. Simulation results showed that, SO-DSM greatly reduces the probability of collision with primary users and the probability of interruption of CR link. A novel spectrum allocation model of capacity consciousness was proposed in [4] for cognitive radio networks, which modeled the interference restrictions based on the concept of interference temperature. Hence the secondary users increased their transmission power until one of their neighbor interference temperatures exceeded the interference

---

[*] Corresponding author.

temperature threshold value. A learning process which allocates spectrum to the cognitive users dynamically was explored in [5], which also took the interference problems into account increasing the allocated power of the throughput. Q-learning algorithm was introduced into the cognitive radio system [6], and it presented a spectrum allocation algorithm based on Q-learning. Simulation showed that Q-learning algorithm could achieve reasonable spectrum allocation tasks and improve the convergence speed and the efficiency of spectrum allocation.

Game theory has widely applied into the resource management of wireless communication network, among which there are many methods study the spectrum allocation using game theory. In the mathematical model of game theory, the players always try their best to get the largest interests with minimal cost. The effectiveness of these two kinds of players be measured by the utility function or cost function. As a result, the behavior that every player strives for the max profits can be equivalent to maximize the utility function or minimize the cost function [7]. Reference [8] summed up the matching game, the cournot game, the bertrand game, the stackelberg game, the repeated game, the supermodel game, the potential game, the auction game and the evolutionary game these nine game theory model for the researches of spectrum allocation from the angle of cooperative game and non-cooperative game, and also analyzed the suitable problems for research of each game model, and the advantages and disadvantages of the spectrum allocation when used to study. Evolutionary game theory is a new development of game theory in recent years. Its basic idea is the biological evolution and genetic theory [9].It gradually adapts the environment's changes though test, imitation, dynamic adjustment process.

The process of spectrum allocation in cognitive radio is equivalent to a leasing market of spectrums; the authorized user can lease the free spectrum to unauthorized users. Among the lessors of spectrum namely authorized users have a competitive relationship for renting spectrums. Among the renters of spectrum namely unauthorized users is also a competitive relationship for loaning spectrum. And the biggest characteristic of game theory is to be able to find the Nash equilibrium for the corresponding game process, sometimes the point of Nash equilibrium is just the optimal point, thus, game theory can guide and analyze the spectrum allocation algorithms, and help us to find the optimal strategy of spectrum allocation in cognitive radio [10].

Dynamic spectrum leased problem was considered of the spectrum secondary market of the cognitive radio network that the secondary service provider leased spectrum from the spectrum brokers and then provided services to the secondary users [11]. The optimal decision of the secondary providers and the secondary users decided dynamically under the competition. Since the secondary users could adapt to the service selection strategy according to the received service qualities and prices. Modeling the dynamic service selection to a evolutionary game in a lower level. Using dynamic service choice, competitive secondary providers could dynamically lease spectrums to provide services to secondary users. In a higher layer making a spectrum leased differential game to simulate the competition. The evolutionary game service chose distributed to describe the state of differential game. Open loop and closed loop Nash equilibrium both had solutions as differential dynamic control game. Numerical comparison showed that it was better than the static control in the profit and convergence speed.

Evolutionary game theory is employed to study cooperative spectrum allocation problem in the cognitive radio [12], which pointed out that the cooperative spectrum sharing among a primary users (PU) and multiple secondary users (SUs) helped to improve the whole system's throughput. They proposed a kind of two-tier game in the sharing of spectrum in which secondary users decided whether to cooperate under replicator dynamics and the primary user adjusted its strategy to allocate time slots for the cooperative secondary users' transmission. In addition, the author also designed a distributed algorithm to describe the secondary user's learning process, proved that the dynamics could effectively converge to the evolutionary stable strategy (ESS), this was also the optimal strategy of the primary and the secondary users. Simulation results showed that the proposed mechanism converged to the ESS automatically, at this time all the secondary users would keep their strategies. In addition, it showed that the mechanism could help secondary users to share information and obtained higher transmission rate than that in fully cooperative or non-cooperative scene.

The existed methods using evolutionary game theory to study the spectrum allocation of cognitive radio network are mainly divided into two kinds: one kind is spectrum allocation between multiple primary users and multiple cognitive users, one kind is the situation that a primary user allocates spectrum for multiple cognitive users. However this paper uses the method of evolutionary game theory to analyze the behavior of spectrum allocation in a group which has two primary users and 30 cognitive users in the cognitive networks. The special condition will make people more visual, intuitive understand the evolutionary process of cognitive users for the reason that maximizing their own benefits when choosing spectrums, and the behavior of primary users reasonably decided the sale prices in order to maximize their own profits.

2   Proposed Algorithm

2.1   System Model

Figure 1 shows the spectrum trading system model among multiple primary users and multiple cognitive user groups: there are N cognitive users groups: SUs group 1 to SUs group N, and also there are a number of cognitive users in each group. The cognitive user group 1 has i cognitive users, group N has k cognitive users. There M authorized user frequency: spectrum frequency f_1 to spectrum frequency f_M.

Cognitive users located within the group 1 to N can freely access to the frequencies f_1 to f_1. Primary user PU 1 to PU M can obtain certain profit through leasing spectrum to the cognitive users. At the same time, cognitive users also expect to achieve its own data communication by buying the available spectrums, and get larger data transmission revenues at smaller pay. Among the cognitive users in group 1 to N, they communicate and exchange information with each other by base stations, and complete their own evolution at the same time when competing with limited spectrum resources (i.e., strategy adjusting). At the same time, based on the spectrum selection strategy of different cognitive users, among authorized users PU 1 to PU M also need to competitively adjust their sharing spectrum bandwidth and the corresponding spectrum price, in order to win more cognitive users and maximize its own interests at the same time.

This article mainly research the spectrum trading behaviors in a group which has two primary users and 30 cognitive user among cognitive users and primary users, the model as shown in figure 2, there are 30 cognitive users in Fig. 2, any one of the cognitive user SU i ($1 < i < 30$) among the 30 cognitive users can buy the idle spectrums from the primary user PU 1 and PU 2.
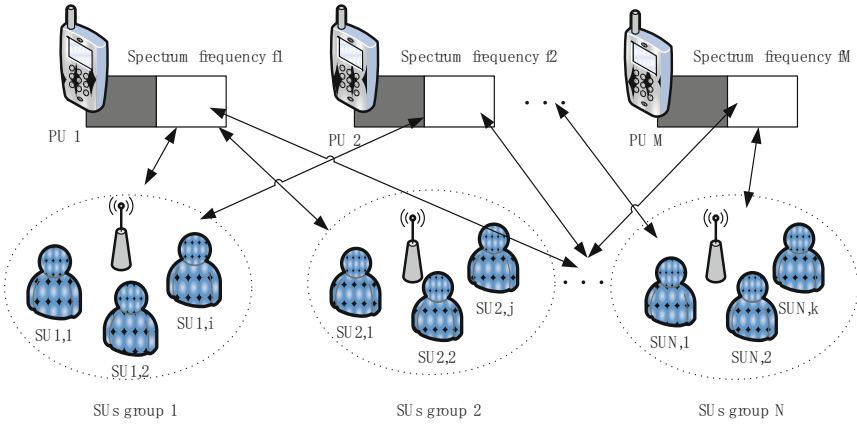
Fig. 1. Spectrum trading model among multiple primary users and multiple groups of cognitive users
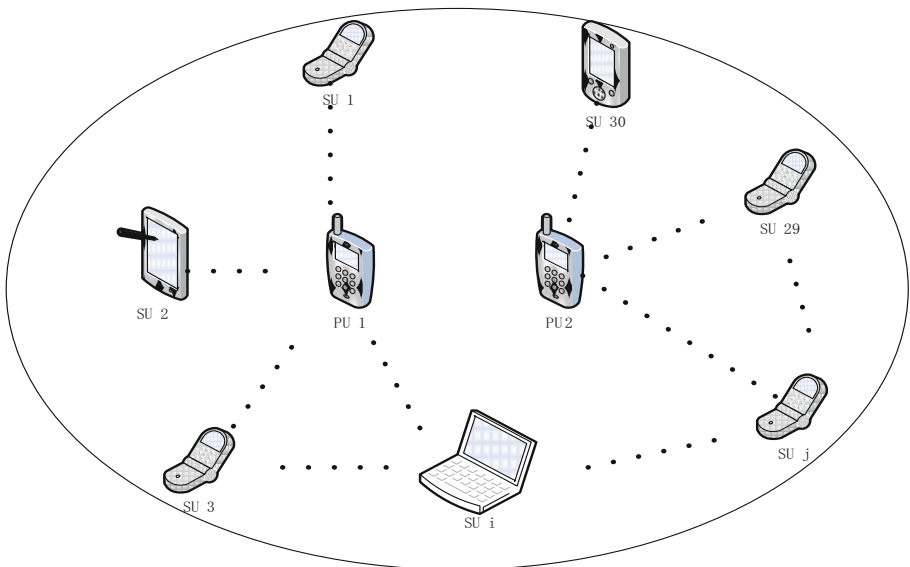
Fig. 2. The spectrum trading model in the group that has two primary users and 30 cognitive users

Related elements of Fig. 2 shows in the game are as follows:

Player: In the elliptical area (groups), the cognitive users compete the limited spectrum resources of authorized user 1 and authorized user2 .

Population: All the users in the elliptical area constitute a group.

Strategy: Cognitive users in the elliptical area buy free spectrum of authorized user 1 and authorized user 2.

Payoff: The net utility of game participants.

In the evolutionary game theory, in order to more accurately describe the dynamic properties of a system, people put forward the replicator dynamics model [13] to describe the dynamic adjustment process of the whole system.

In the dynamic evolutionary game, in a group, individual strategies can be copied through learning by other individuals. Replicator dynamics from separated secondary user's groups, using n_k^a to express the number of secondary users who buy spectrums from primary user k in the group a. The total number of secondary users in the group a is:

$$N = \sum_{k=1}^{M} n_k^a \,, \tag{1}$$

where M is the total number of primary users who provide spectrum opportunities, the proportion of secondary users who purchasing opportunities from the primary user k.

$$x_k^a = n_k^a / N \,. \tag{2}$$

The state of the group uses vector $\mathbf{x}^a$:

$$\mathbf{x}^a = [x_1^a, \cdots, x_k^a, \cdots x_M^a \,]. \tag{3}$$

Continuous time imitator dynamic is defined as follows [13]:

$$\frac{\partial x_k^a(t)}{\partial t} = \dot{x}_k^a(t) = \rho x_k^a(t) \cdot [\pi_k^a - \bar{\pi}^a] \,, \tag{4}$$

$$\bar{\pi}^a = \sum_{k=1}^{M} x_k^a \pi_k^a \,, \tag{5}$$

here $\bar{\pi}^a$ is the average pay of secondary users in group a. The changing speed of the spectrum selection of the secondary users is controlled by the parameter ρ.

It can be seen that if the secondary users' revenue in the group a is lower than the average, the secondary users who buy spectrums from the primary user k will reduce as the evolution of time. The group of players will dynamically evolve based on the imitator over time; it will converge to the evolutionary equilibrium. When reaching the evolutionary equilibrium, the revenues of all players are equal to the average gain, i.e.

$$\frac{\partial x_k^a(t)}{\partial t} = \dot{x}_k^a(t) = 0 . \tag{6}$$

In order to gain the benefits of cognitive users, the paper used logarithmic utility function to quantify the actual accessible data transmission throughput to meet the needs of users. Defining the revenue of cognitive users who choose primary user i in group a as that the difference of price between the obtained transmission utility of cognitive users through accessing to the free spectrum of primary user i and the payment to primary user:

$$\pi_i^a = \phi\left(\frac{\mu b_i d_i}{\sum\limits_{a \in A_i} x_i^a N^a}\right) - p_i = \ln\left(\frac{\mu b_i d_i}{\sum\limits_{a \in A_i} x_i^a N^a}\right) - p_i, \tag{7}$$

where the meaning of the parameters represent as table 1:

Table 1. Description of parameters

Parameters	Description
b_i	Spectrum bandwidth that primary user i can share, the unit is Hz
d_i	Spectrum transmission efficiency based on adaptive modulation, the unit is bps / Hz
x_i^a	The percentage of cognitive users who choose the cognitive i in the group a
N^a	The total number of cognitive users in the group a .
p_i	The rental price that cognitive users pay to primary user i .
μ	Weight constant of utility that cognitive user obtained.

The change of the percentage of cognitive users who choose primary $i(i = 1, 2)$ in the elliptical area of Fig. 2 can be expressed as follows:

$$\dot{x}_i^a = \rho x_i^a (\pi_i^a - \bar{\pi}^a) , \tag{8}$$

where $\bar{\pi}^a$ is the average profit of cognitive user in group a .

$$\bar{\pi}^a = x_1^a \pi_1^a + x_2^a \pi_2^a . \tag{9}$$

In Fig. 2, replicator dynamics can be calculated as:

$$\dot{x}_1^a = \rho x_1^a \left(\ln\left(\frac{\mu b_1 d_1}{x_1^a N^a}\right) - p_1 - x_1^a \pi_1^a - x_2^a \pi_2^a \right). \tag{10}$$

$$\dot{x}_2^a = \rho x_2^a \left(\ln\left(\frac{\mu b_2 d_2}{x_2^a N^a}\right) - p_2 - x_1^a \pi_1^a - x_2^a \pi_2^a \right). \tag{11}$$

Then we can get:

$$\begin{cases} \dot{x}_1^a = 0, \\ \dot{x}_2^a = 0, \\ x_1^a + x_2^a = 1, \end{cases} \quad , \tag{12}$$

and get the solution of evolutionary equilibrium as:

$$x_1^a = \frac{1}{\dfrac{b_2 d_2}{b_1 d_1} e^{p_2 - p_1} + 1}. \tag{13}$$

But the profits which primary 1 and 2 gain through providing spectrums for cognitive users can be expressed as:

$$\pi_1 = p_1 * x_1^a N^a, \tag{14}$$

$$\pi_2 = p_2 * x_2^a N^a. \tag{15}$$

Solving the derivative of each primary user reward to spectrum price can get the optimal reaction function of the primary users:

$$p_i^* = f(p_j) = 1 + LambertW\left(\frac{b_i d_i}{b_j d_j} e^{p_j - 1}\right), \tag{16}$$

where $i = 1, 2, i \neq j$. From the expression (16), we can know that each primary user's optimal reaction has nothing to do with cognitive users of the group, it is a increasing function of other primary users' strategies. According to the reference [14], we can know that pricing game of the primary users of is a supermodel game, and there is at least a Nash equilibrium.

2.2 Spectrum Selecting Algorithm

We assume that every cognitive user within the group a based on a centralized controller (such as base station) to obtain average utility information $\overline{\pi}^a$ of all users within the group. Every cognitive user's spectrum selection decision is based on their own current rewards and the average reward of all users in the area a. As a result, the specific steps of spectrum selection algorithm of the cognitive users based on evolutionary game are as follows:

Step 1. Every cognitive user randomly selects primary user $i(i = 1, 2)$;

Step 2. All the cognitive users in group a calculate π_i^a and send it to the base station;

Sept 3. Base station calculates the average utility $\overline{\pi}^a$ of group a and broadcast it to all the cognitive users;

Setp 4. If $\pi_i^a < \bar{\pi}^a$, then change the selective primary i with the probability $(\pi_i^a < \bar{\pi}^a)/\bar{\pi}^a$;

Setp 5. The cognitive user converts to a primary network $j(j \neq i, \pi_j^a > \pi_i^a)$ to get a higher reward;

Setp 6. Repeating the steps 2~5.

All cognitive users select different spectrums in the group a have the same earnings, which reach the evolutionary equilibrium.
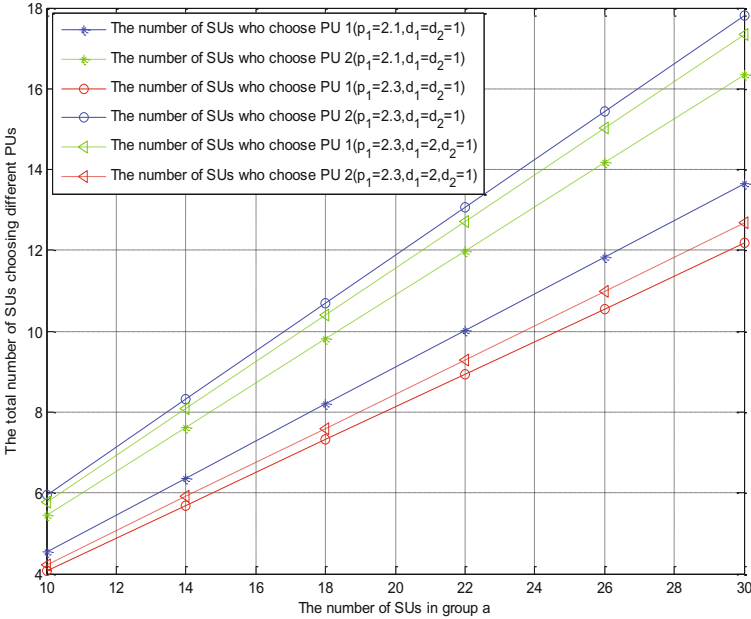
Fig. 3. Total number of cognitive users choosing different spectrum of primary users

3   Simulation Results

Matlab is used to verify the validity of the proposed scheme. Assume that there are 30 cognitive users with the same conditions to compete the idle spectrum of authorized user 1 and authorized user 2 together. We will analyze the dynamic evolution behavior of cognitive users when choose to buy different primary users from the adaptive evolutionary equilibrium and the convergence of evolutionary algorithm of the cognitive population these two aspects, and the behavior that reasonably adjusting the selling price of the spectral from the optimal reaction of primary users and Nash equilibrium, the primary user's income two aspects.

Simulation parameter Settings: $b_1 = 240, b_2 = 260, \mu = 10$.

1). Self-adaption of evolutionary equilibrium

Fig. 3 shows the change of the number of cognitive users who choose the primary 1 and 2 in the process that the number of total cognitive users in the group increased from 10 to 30 when $p_2 = 2.0$ is fixed, p_1 and d_1, d_2 take different values.

We can see from the figure that though the transmission efficiency of the two primary users is the same: $d_1 = d_2 = 1$ and the spectrum price of primary 2 is higher than primary 1: $p_1 > p_2$, the increasing amplitude of cognitive users who choose primary 2 is greater than that of selecting primary 1 because primary user 2 can provide more spectrum bandwidth to cognitive users than primary user 1. When the transmission efficiency of the primary user 1 raise to $d_1 = 2$ .We can see from the figure that the price of primary user 1 is higher than that of primary user 2, but the increasing amplitude of cognitive users who choose primary 1 is greater than that of selecting primary 1 obviously.

The Fig. 3 embodies the adaptive behavior of cognitive users in choosing to buy spectrums; cognitive users would adjust their own strategies when buying spectrums according to the price and quality of spectrums provided by primary users. So as to choose the spectrum with low price or good quality to improve their income.

2). Convergence of group evolutionary algorithm:

Fig. 4 shows the times of iterations which cognitive users required for stability of returns under different spectrum transmission efficiencies. When reaching the evolutionary equilibrium, the income of cognitive user is equal to average revenue.
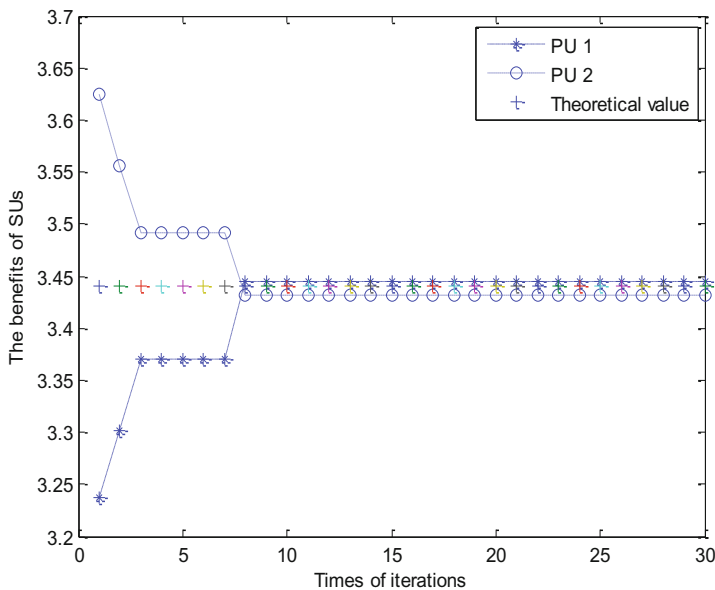
Fig. 4. The times of iteration which cognitive user required for steady returns

In Fig. 4, Because the spectrums that secondary users can share provided by primary user 2 are more than that of primary user 1, and the spectrum transmission efficiency of primary user 2 is much better than that of primary user 1, so more cognitive users will choose the spectrum of primary user 2 to access to, so in the beginning, the income of cognitive user who choose the primary user 2 is 3.63.It is greater than the income 3.23 of cognitive users who choose the primary user 1. In addition, the profit of the cognitive users who choose primary user 1 is lower than the group's average income, and the profit of the cognitive users who choose primary user 2 is higher than the group's average income. Cognitive users would make a comparison with the group's average revenue, and then adjusting their strategies, and constantly evolving, until their income is equal to the group's average revenue. It can be seen from Fig. 4, the algorithm converged in step 8.The cognitive users in the group evolve to that the average revenue gained from the primary user 1 is 3.445, and it is 3.430 from the primary user 2. They are very close to the theoretical value 3.440.

It is also can be seen from these figures that, the evolutionary equilibrium that the spectrum selection process of cognitive user reaches is a dynamic balance, and it will constantly update the state of evolution equilibrium as the changing of external environment.

3). Optimal reaction and Nash equilibrium

Fig. 5 shows the change of the optimal response curve and Nash equilibrium of the primary users under different spectrum transmission efficiency of the primary users 1 and 2.
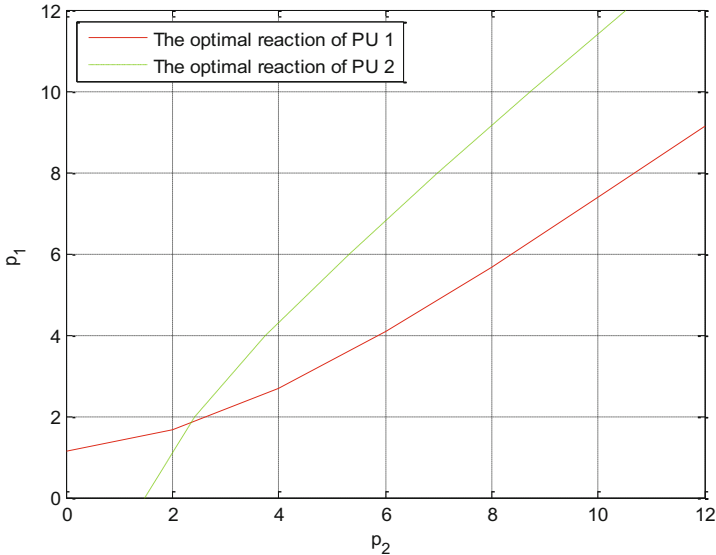
Fig. 5. Optimal reaction of primary users

In the figure, when $d_1 = 1, d_2 = 2$, the Nash equilibrium of primary user 1and 2 are $p_1 = 1.774, p_2 = 2.292$. Because the spectrum bandwidth provided by primary users 2 is greater than that of primary user 1, it can also accommodate more cognitive users, spectrum transmission efficiency of primary user 2 is greater than that of primary user 1, so the primary user 2 obtain greater profits by raising the price of spectrums. While due to the limited spectrum bandwidth resources and low transmission efficiency of the primary user 1, it has to reduce the price of spectrums to attract and retain more cognitive users, to ensure that their incomes do not decrease. It will affect the optimal reaction the price of spectrum of the primary user 1.It makes the reduction of the optimal price of the primary user 1 and the increasing of the optimal price of the primary user 2.
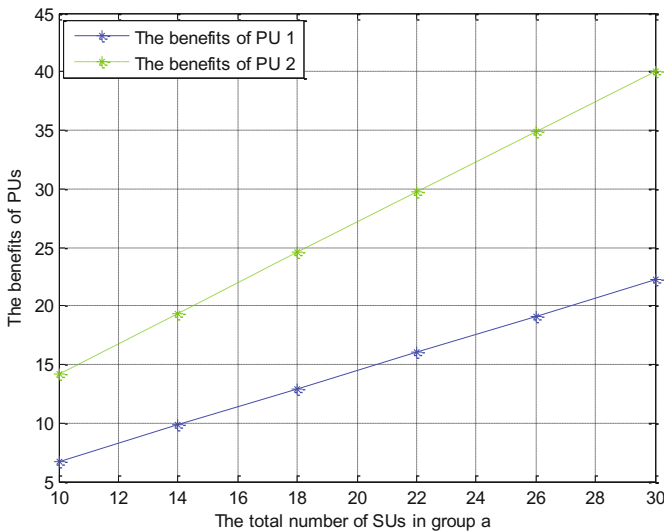
Fig. 6. The increasing of benefits of primary users as the increasing of cognitive users in group *a*

We can see from the figure, the optimal reaction function of every primary user is nonlinear function of other users' strategies, and the point of Nash equilibrium lies in the cross node of the optimal reaction function of the two users. In addition, the slope of the optimal reaction function the primary user 2 is always greater than 1, and gradually tending to be 1 with the increasing of p_2. The slope of the optimal reaction function of the authorized user 1 is always less than 1, and gradually tending to be 1 with the increasing of p_1.

4). The benefits of primary users

The Fig. 6 show the changing curve of primary users in the process that the cognitive users increase from 10 to 30 in group *a* under the situation that the transmission efficiency of spectrums of primary users is changeable.

In Fig. 6, $d_1 = 1, d_2 = 2$. When the number of cognitive users increase, by the reason that the primary user 1 can provide less spectrums than the primary user 2, and the spectrum transmission efficiency of primary user 2 is greater than primary user 1. While good transmission efficiency and high network capacity can attract more cognitive users. So we can see from the simulations: when the number of cognitive users in group a is 10, the revenue of primary user 1 is 5.5, while the revenue of primary user 2 increased to 11; When the total number of cognitive users increase to 26, the profits of primary user 1 increase to 19, while the profits of primary user 2 increase to 35. The revenues of primary user 1 increase 7 than that the total number of cognitive users is 18, while the revenue of primary user 2 increase 10. It is visible that, though the revenues of primary user 1 and 2 both are increased linearly. But the increasing margin of revenue of primary user 2 is much greater than that of primary user 1.

We can see from the figures 6, the quality and capacity of the free spectrums which provided by primary users will affect directly the choices of cognitive users. The primary users who own the spectrums of good quality or big capacity will be more favored by cognitive users. Moreover, the primary users with high spectrum transmission efficiency and big capacity can also raise the price of spectrums, so they will have higher benefits.

4   Conclusion

The paper first introduced the evolutionary game theory and its imitators dynamic model. Then modeling the process that cognitive users buy the idle spectrums of primary users as the dynamic evolutionary game; And designing a scene that there are two primary users 1 and 2 and 30 cognitive users in the group a .And designing a scene that there are two primary users 1 and 2 and 30 cognitive users in the group , supposing that collecting, processing, and broadcasting the user's earning information, We proposed spectrum allocation algorithm based on game theory to get the solution of evolutionary equilibrium. Analyzing and revealing the process of evolutionary game and the evolutionary equilibrium of cognitive users under different parameter settings and the change of primary user's earnings and the change of the Nash equilibrium point from the adaptive evolutionary equilibrium of cognitive users, the convergence of evolutionary algorithm of the cognitive users' group a , the optimal reaction of primary users and Nash equilibrium, the primary user's income four aspects. When the spectrum those cognitive users have bought reach the evolutionary equilibrium, the number and earnings of cognitive users that access to the primary user 1 and 1 and 2 both reach a stable state, so as to maximize the utility of the cognitive network, while the effectiveness of primary user is also maximized. The future work will focus on more general and actual case with relaxing the assumptions to optimize spectrum allocation in cognitive radio networks.

References

1. Federal Communications Commission, "Spectrum policy task force report", FCC Document ET Docket No.02-135 (November 2002)
2. Ji, Z., Liu, K.J.R.: Dynamic spectrum sharing: a game theoretical overview. IEEE Communications Magazine 45(5), 88–94 (2007)
3. Khozeimeh, F., Haykin, S.: Brain-Inspired Dynamic Spectrum Management for Cognitive Radio Ad Hoc Networks. IEEE Transaction on Wireless Communications 11(10), 3509–3517 (2012)
4. Yousefvand, M., Khorsandi, S., Mohammadi, A.: Interference-Constraint Spectrum allocation Model for Cognitive Radio Networks, pp. 357–362. IEEE Conference Publications (2012)
5. Maulik, S., Roy, R., De, A., Bhatttacharya, A.: Online dynamic Resource allocation in interference temperature constrained cognitive radio network using reinforcement learning, pp. 1–5 (2012)
6. Wu, S., Jiang, H., Xu, W.: Research of Cognitive Radio Spectrum Allocation Based on Improved Q-learning Algorithm. In: International Conference on Oxide Materials for Electronic Engineering (OMEE), pp. 13–16. IEEE (2012)
7. Osborne, M.J.: An Introduction to Game Theory. Oxford Univ. Press (2003)
8. Ni, Q., Zhu, R., Wu, Z.: Spectrum Allocation Based on Game Theory in Cognitive Radio Networks. Journal of Networks 8(3), 712–722 (2013)
9. Vincent, T.L., Brown, J.S.: Evolutionary Game Theory, Natural Selection, and Darwinian Dynamics. Cambridge Univ. Press (July 2005)
10. Vincent, T.L., Brown, J.S.: Evolutionary game theory. Cambridge University Press, London (2005)
11. Zhu, K., Niyato, D., Wang, P., Han, Z.: Dynamic Spectrum Leasing and Service Selection in Spectrum Secondary Market of Cognitive Radio Networks. IEEE Transactions on Wireless Communication 11(3), 1136–1145 (2012)
12. Wu, Z., Cheng, P., Wang, X., Gan, X., Yu, H., Wang, H.: Cooperative Spectrum Allocation for Cognitive Radio Network: An Evolutionary Approach. In: International Conference on Communications (ICC), pp. 1–5. IEEE (2011)
13. Niyato, D., Hossain, E., Han, Z.: Dynamics of Multiple Seller and Multiple Buyer Spectrum Trading in Cognitive Radio Networks: A Game-Theoretic Modeling Approach. IEEE Trans. on Mobile Computing 8(8), 1009–1022 (2009)
14. Topkis, D.M.: Supermodularity and complementarity. Princeton University Press, Princeton (1998)

Web Service Resources Aggregation
Based on the Customer Behavior

Zhong Wu[1,2], Gui-hua Nie[1], Shang-ying Xu[1], and Ji Luo[1]

[1] Institute of E-Business, Wuhan University of Technology,
430070 Wuhan, China
[2] Economic Management Department, Wuhan University of Technology HuaXia College,
Wuhan 430223, China
{Zhong Wu,Gui-hua Nie,Shang-ying Xu,Ji Luo,7849800}@qq.com

Abstract. At present, some issues exist in web service resources aggregation studies, such as resource isomerism, ignoring personalized needs of customers. This article aims to study the overall model of web service resources aggregation based on the isomerism customer behavior information in the internet, thus to provide structured and semantic intelligent service resources aggregation. First, this article constructs the overall model of customer behavior and divides the service resources, then aggregates the isomerism resources and unstructured resource based on the mediator / rapper structure. As a result, it satisfies customers and improves the efficiency of service resources inquires. It has important theoretical and substantial empirical values.

Keywords: customer behavior, resources aggregation, resource isomerism, mode integration.

1   Introduction

With the rapid development of network technology, web service resource has become more and more abundant. However, it is often difficult to accurately position the messy information How to comprehensively utilize and integrate the isomerism resources, and meet customer's demand for personalized in the open and dynamic network environment are the key issues of service resources aggregation research.

Information aggregation needs to bring together and analyze related information from different data sources, it is to solve the isomerism resources in the semantic aspects, and provide aggregate function based on the relationship among data sources and business process [1]. At present, many experts and scholars have made a lot of research. Patrick and Bangyu Wu applied the workflow technology based on the service composition to the information integration, and studied the information aggregation approach based on business process [2][3]. P. Kokkinos established a grid resource information aggregation framework [4]. Zibin Zheng put forward the reliability information sharing mechanism among customers and constructed a public information sharing platform for web customers [5]. Jialang Seng proposed an intelligent information integration method, it provided a far intelligent information integration query for GAV and LAV [6]. Hongjiang Cao proposed an approach for

web service discovery based on context and reasoning [7]. Kai Ye proposed a user preferences-driven dynamic service aggregation model (UPDDSAM) and the corresponding aggregation algorithm. The model implemented runtime aggregation of service processes and service resources by adopting dynamic rules and latebinding technology [8]. Lei Chen introduced a semiautonomous methodology to provide for domain-specific information gathering and integration from multiple distributed sources [9]. Nair discussed the open management and security issues associated with the two models of cloud bursting and cloud brokerage [10].

However, the research and application in information aggregation still has deficiencies. On the one hand, the scope of application resources is narrow, information resources are in the isomerism environment. On the other hand, the personalized service application is little and its form is simple. This article is to make a key technology research on the integration of customer behavior mode and the isomerism resources mode, and also service resources aggregation. In this article, we use mediator/rapper structure and the aggregation mode to transform the corresponding isomerism mode of the underlying information sources into isomorphic mode. We also utilize the general structure and mode to rewrite structured and semi-structured information sources, use ontology reasoning engine and mediator/rapper structure to deal with unstructured information sources. Ontology reasoning may assist to inquire about the structure and semantic interoperability. Rapper may transform the data interface and manipulate the request between mediator and information source. Mediator may provide the underlying isomerism information element of integrated data for the client and process customers' requests.

2   Resources Divided Based on Customer Behavior

Based on the research by Hawkins, the overall customer behavior mode can be drawn and it is shown in Fig.1. The patterns of customers' behaviors are influenced by the social, cultural, psychological and personal factors, including personal and psychological factors demonstrated by specific situational factors.
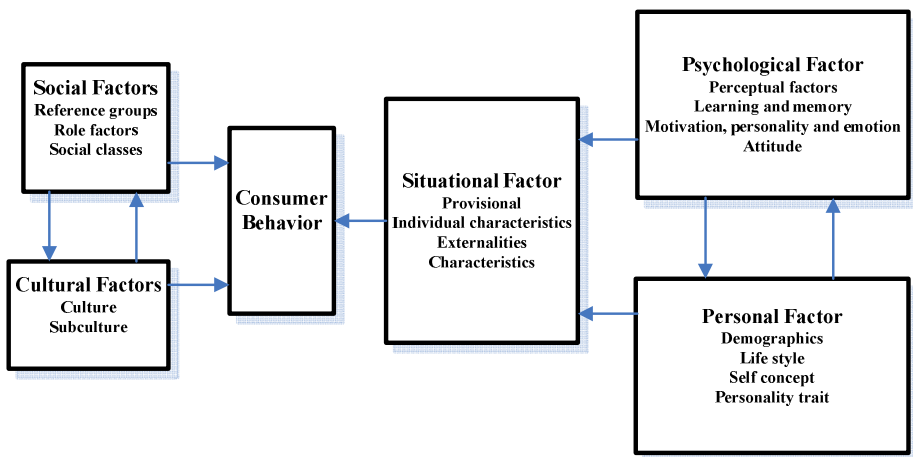
Fig. 1. Overall customer behavior mode

It can be seen from Fig. 1 that culture, subculture, reference groups, the role of factors and social classes are the cultural and social factors that influence customers' behaviors. Demographic factors, lifestyle, self-concept and personality characteristics, perceptual factors, learning and memory, motivation, personality and mood and attitude are the personal and psychological factors that affect customers' behaviors. These factors not only determine the customer's decision-making behavior to some extent, but also amplify or inhibit their impact on the external environment and the marketing stimulus.

We divide customers' service resources according to comprehensive analysis of various factors that affect customers' spending behaviors, then, provide customers with a variety of service resources on different consumers' behaviors influencing factors.

2.1   Divided by Consumption Levels

Social attributes of the basic customers' information need to be collected, such as social class, cultural background, income and family background. After a comprehensive survey, we divide the customers into three groups such as high-end consumer group, midrange consumer group and low-end consumer group. The demand-oriented products of the high-end consumer group are luxury products, high-end technology products, as well as a variety of top services. The demand-oriented products of the midrange consumer group are economic products, cost-effective products and mid-range services. The demand-oriented products of the low-end consumer group are discount service information such as cheap products, low-end consumer goods and electronic coupons.

2.2   Divided by Demographic Characteristics

Taking into account the demographic characteristics, consumers' personal factors such as age, gender and personality need to be collected. We divide the customers groups into four groups according to age, those are children's groups, youth groups, middle-aged groups and older age groups. The needs of children groups orients books and pens, stationery, science and humanities exhibitions, educational courses, hobby classes and other service information. The youth groups demand personalized products, fashion products and cost-effective products. The middle-aged groups demand practical products, daily necessities and household appliances. The older age groups demand health care products, treatments, and so on.

Also we can divide the customers groups into two groups according to gender, those are women and men groups.

2.3   Divided by Situational Factors

The situational factors take into account the promotional activities that the various service providers and major shopping malls made on different festivals and different consumption periods. For example, travel service providers may provide dragon boat activity on the Dragon Boat Festival and moon cake activity on Mid-Autumn Festival,

the major shopping malls launches discount activities on the National Day, and the online shopping malls launches promotional discounts in the anniversary celebration. The participation effect of such service resources information is fine, and it applies to all levels of customer groups which can participate in different promotional activities.

In conclusion, due to the diverse cultural environment, social environment, situational factors for different customers, it is particularly necessary to collect customers' information. We can distinguish customers according to gender, age and social class, divide different customers into different customer markets, aggregate service resources that adapt to customers' demands to provide more effective and appropriate services.

3   Resource Aggregation Tools

With the development of internet and the information technology, intelligent information isomerism in e-commerce has become widespread and extremely important. The simple acceptance of intelligent information without aggregation may lead to a disordered information request, and also lead to the offset settings of e-commerce. The general approach to deal with isomerism problems is to establish a common data model. XML is the standard data document format for exchanging web information. However, XML only deals with structured isomerism information, and can barely handle the semantic isomerism situation. Ontology is treated as an important and natural method to express clear semantics and relationships of the real world, and has played an important role in semantic integration of unstructured information.

Tsimmis is a system which improves information integration rapidly and ensures the information steady. It is based on the structure of mediator/ rapper, and the information of mediator and rapper is automatically generated for the underlying structured and unstructured data mapping information of object exchange model (OEM). Information manifold (IM) is an isomerism information integration system for querying and browsing multiple network sources of information. It uses the knowledge representation technology to develop the same domain model, and makes information retrieval and organization from isomerism information sources. DISCO system is used to provide scaling access to isomerism data sources. It involves with solving three problems of isomerism database data sources including weak data source, fragile demodulation and the data sources which can't be obtained get-failed problems. Garlic system is a united database management system, which integrates data from a variety of relational and non-relational data plan. It has the mediator / rapper structure, utilizes Garlic Data Language (GDL) to describe the local data source and uses the grammatical form of the Garlic data model. Clio system is used to manage and improve isomerism data integration, conversion and evolution of complex tasks. Its integration engine consists of three engines. In order to produce the desired mapping, they interact with the internal mapped knowledge base and customers. Yat model is a semi-structured data model based on an ordered number and mark node. It can generate to improve any description of the data and allow different levels of description of the reasoning. The Mix system using XML likes its predecessor Tsimmis, it relies on the well-known demodulation structure to provide customers with the integration of the underlying data source query. It uses XML as a

common data model deposited into a unified and flexible description of any source data, and it is not used to deal with semantic isomerism. Agora system is used to survey the flexibility and availability of the system based on relational technology for processing XML query. It uses the object model to the performance of the data source and uses virtual universal relational mode which is constructed in accordance with a fully normalized relationship between the version of the XML document hierarchy for the connection to a data source and XML global mode.

The above data integration tools are different in isomerism information sources, data model and information integration strategy. Wherein, on the aspect of isomerism information sources, IM is the multiple network information, Disco is distributed isomerism database, Garlic and Clio is a variety of relational and non-relational data sources, Agora is relational and XML data sources, several others are applicable to any data source. On the aspect of information integration strategy, there are two integration patterns: GAV and LAV. We must consider the trade-off between different strategies when executing the information integration. IM and Agora use local view and other tools use the global view. We can use different tools to different data sources for data aggregation.

4   Mode Integration of Isomerism Resources

When intelligent resource integration is performed, it will encounter the problem of resource description. For the establishment of a uniform view of the existing information resources integration mode, the isomerism mode of the underlying information sources needs to try to translate into isomorphic mode [6]. This article uses XML as a common data model and builds a global model, the global integration process is shown in Fig. 2.
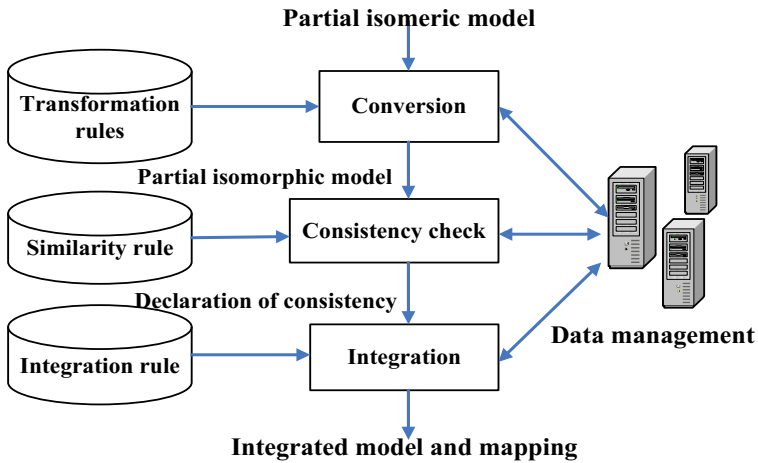
Fig. 2. Global integration process

To change the description of isomerism information sources data model into an isomorphic one, it's necessary to create XML and local data model for rewriting rules. Isomerism information sources relates to structured information sources, semi-structured information sources and unstructured information sources. Common structure and pattern can be applied to rewrite structured and semi-structured information sources, and special processing can be applied to at a later stage.

In order to transform structured data model and semi-structured information sources into XML, firstly, we must specify that the rewrite rule is a one-to-one correspondence with a local data model, secondly, we must determine the link between XML and ontology data model. Table 1 shows the corresponding relationship between relational model structure and XML model structure.

Table 1. Relationship between relational and XML model

Relational model Structure	XML model Structure
Relationship	Elements (Complex Type)
Attribute	Elements
Data Types	Data Types (Primitive/ Simple Type)
Cardinal Number	Multiplicity (Minimum / Maximum Number of Occurrences)
Primary Key	Keyword
Foreign Key	Keyword

Table 2 shows the correspondence between object model structure and XML model structure, it is similar to rewrite the relational model.

Table 2. Object and XML mode structure

Object Mode Structure	XML Mode Structure
Class	Elements (Complex Type)
Attributes (Simple)	Elements
Primitive Types	Data Types (Primitive Type)
Structure (Customer-defined Type)	Data Types (Primitive /Simple Type)
Keyword	Keyword
Expansion (Inherited)	Supports Only Single Inheritance (Extension / Restriction)
Relationship / Opposite	Does Not Support
Extent	Does Not Support
Prevention	Does Not Support

Before the integrated mode, we need to determine the commonalities and distinguish the relative relationship between the different models. Mode integration is to find a similar structure to isomerism mode by corresponding relationship. The corresponding relationship is the integration point. In order to get the reasons of

Table 3. The reasons for structural isomerism

Reasons	Explanation
Naming Conflicts	The command conflict has two kind of types. One is synonyms, the other one is homonyms.
Field Conflict	Two properties is semantic similarity, they may have different fields or data types.
Default Values Conflict	This conflict depends on the relevant attribute domain definition.
Identifier Conflict	Since the identifier records used by the primary key of the two entities in the two data sources is different in the semantics.
Integrity Constraints Conflict	Two similar semantic attributes may be inconsistent with each other because of being restricted by the constraint.
Missing Data Items Conflict	When the entities descriptions are similar to the semantic entities, in which an entity lost attribute, this conflict will occur.
Aggregation Conflict	When aggregation used for a data source to determine an entity (or attribute), the conflict will occur in other data sources.
Attributes-Entities Conflict	When the same object in one data source is modeled as an attribute, while in another is modeled as an entity, this conflict will occur.
Data value - Entities Conflict	When the value of an attribute in a data source is corresponded to the entity in another data source, this conflict will occur.
Data values - Attributes Conflict	When the value of an attribute in a data source is corresponded to another attribute in another data source, this conflict will occur.

structural isomerism, we must ensure the interoperability among the underlying data sources to solve the isomerism problem. The reasons for structural isomerism summarized below in Table 3.

The unstructured information sources are just like static web pages and multimedia files and have no model, so they must be processed as a special case and provided special solutions by ontology. We can create an index to these data sources without transformation. We wait till the global mode is created, then we specify mappings between them.

Dealing with benefit relationship in the field of application by ontology which contains facts and relationships to capture the knowledge of the real world is a promising approach. However, the majority of the ontology creation is based on the manual way. The exploitation of ontology has our main stages, including determining the objectives and scope, establishing ontology (which includes ontology acquisition, ontology encoding and ontology integration) , evaluation, and criteria guidance.

Establish ontology is to achieve a realistic semantic interpretation. Not only semantic isomerism but also the definition of the data items and relationships need to be taken into account in the establishment of ontology. The reasons for semantic isomeric are listed in Table 4.

Table 4. The reasons for semantic isomeric

Reasons	Explanations
Scope and Current Value Conflicts	The two attributes with similar semantic may be represented by different units and methods.
Representation Conflicts	The two attributes with similar semantic may be represented by different forms. For example, school grade is represented as {1,2,3,4,5}or {A,B,C,D,E}.
Subjective mapping conflicts	From the subjective point of view, the two attributes is the same, but they have their own representations. For example, Germany grade is {15, 14, …, 0}, while the American grade is {A,B,C,D,E}.
Subclasses Conflicts	The contents of an attribute is a subclass of another. For example, hotels contain congress-hotels, but the content of the latter is smaller, it is only a portion of the former.
Overlapping Conflicts	Part of the two attributes are the same, but they are not equal. For example, hotel and campsites inn.
Incompatibilities Conflicts	The two attributes have the same concept, but in fact their meanings are still slightly different. For example, hotels and inns are the places for travelers to live in when traveling, but the inns are cheaper, some are youth inns. On the contrary, hotels are more expensive .
Aggregation Conflicts	The concepts of the two attributes are different, but the concept in one attributeis an aggregation concept of another's. For example, the hotel enterprises and hotels. Hotel enterprises refer to the enterprises that operate the hotels, and hotels are the places for travelers to live in when traveling.

In consideration of the above reasons, we must establish a relationship between the global mode and the model which has been established. Since ontology uses KM to define the data items and relationship, we consider the elements in global mode as the instance of defined resource. That is to say, we can use the ontology to define the relationships among the elements in global mode. For example, the employee which is the defined resource by ontology means the relation between staff and employee, employees is staff's parent class, and all staff are employees. So we can use this definition to define the staff in elements global mode.

According to the relationship among the different data models, semantic definition of ontology can be appended to the elements of the global model. We use the ontology to define the complete relationship in the field of application, and use

ontology instances to examine global model and define the elements of the global model. When creating a relationship between global model and ontology, we can make ontology reasoning to assist in inquiring and arriving the interoperability of structure and semantic. After creating integrated mode, we need to consider the mapping between the global mode and local mode, and use GAV method in accordance with the query reproduction process to determine this mapping and aggregate the unstructured information modes.

5   Service Resources Aggregation Based on Customer Behavior

Service resources aggregation architecture based on customer behavior is shown in Fig. 3. It shows the overall process of service resources acquisition for the customers. The whole process is from the bottom to the top. First acquired various service resources of the information sources by the rapper, then transformed and integrated the isomerism informations by the mediator. The isomerism informations are transformed into standardized XML documents and then are back to the client as a result of customers' inquiries.
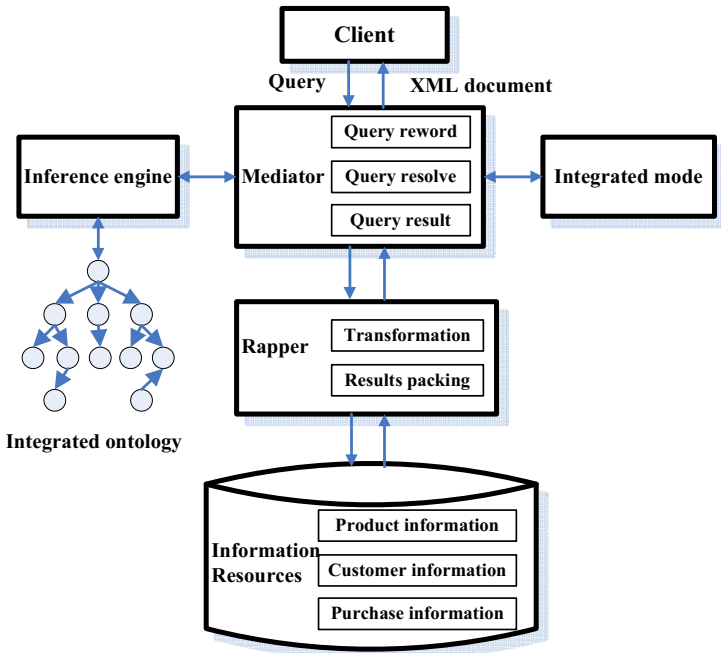
Fig. 3. Service resources aggregation architecture based on customer behavior

Through the above resources aggregation process, we can uniformly aggregate the service resources such as products information that the customers required for, purchasing information and customers' information. We use global mode integration, mediator / rapper and ontology reasoning to solve the problem of isomerism information resources and unstructured information sources, and unified transform isomerism information and unstructured information into structured documents, then return the documents' data required by the customers, thereby improving the quality of web services.

References

1. Liu, M., Zhang, Z., Zhang, X.: Study on Methods of Internet Resources Aggregation. Mechanical Management and Development 23, 170–172 (2008)
2. Hung, P.C.K., Chiu, D.K.W.: Developing work-flow-based information integration with exception support in a Web services environment. In: Proceedings of IEEE Thirty-Seventh Hawaii International Conference on System Sciences, pp. 210–218. IEEE Press, Hawaii (2004)
3. Wu, B., Chi, C., Chen, Z., Gu, M., Sun, J.: Workflow-based resource allocation to optimize overall performance of composite services. Future Generation Computer Systems 25, 199–212 (2009)
4. Kokkinos, P., Varvarigos, E.A.: Scheduling efficiency of resource information aggregation in grid networks. Future Generation Computer Systems 28, 9–23 (2012)
5. Zheng, Z., Lyu, M.R.: WS-DREAM: A Distributed Reliability Assessment Mechanism for Web Services. In: 38th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, pp. 392–397. IEEE Press, Anchorage (2008)
6. Seng, J., Kong, I.L.: A schema and ontology-aided intelligent information integration. Expert Systems with Applications 36, 10538–10550 (2009)
7. Cao, H., Chen, D., Fu, K.: Semantic Web Service Discovery Based on Context Reasoning Method. In: The 10th Wuhan International Conference on E-Business, pp. 485–491. Press, Wuhan (2011)
8. Ye, K., Li, X., Qiao, X.: User preferences-driven dynamic service aggregation model in convergent network. In: 3rd IEEE International Conference on Broadband Network and Multimedia Technology, pp. 933–938. IEEE Press, Beijing (2010)
9. Chen, L., Feng, X.: Intelligent Information Aggregation Service Based on Multi-Agent. In: 2011 International Conference on Information and Industrial Electronics, pp. 330–334. Press, Chengdu (2011)
10. Nair, S.K., Porwal, S., Dimitrakos, T.: Towards Secure Cloud Bursting, Brokerage and Aggregation. In: 8th IEEE European Conference on Web Services, pp. 189–196. IEEE Press, Eindhoven (2010)

Adaptive Consensus of Multi-agents in Jointly Connected Networks

Gaoyang Liu, Hui Yu, and Yi Zhang

College of Science, China Three Gorges University, Yichang,
443002, China
yuhui@ctgu.edu.cn

Abstract. This paper investigates the adaptive consensus problem for multi-agent systems with unknown nonlinear dynamics. The topologies of the networks are jointly connected. An adaptive consensus algorithm is proposed by using linear parameterizations of unknown nonlinear dynamics of all agents. By stability analysis and parameter convergence analysis of the proposed algorithm, adaptive consensus can be realized based on neighboring graphs. The stability analysis is based on algebraic graph theory and Lyapunov theory, the PE condition plays a key role in parameter convergence analysis. The simulation results are provided to demonstrate the effectiveness of our theoretical results.

Keywords: Multi-agent, Adaptive Consensus, Parameter Convergence, Jointly Connected.

1   Introduction

Recently, distributed coordinated control of multi-agents has attracted a great deal of attention from many fields such as physics, biology, robotics and control engineering. This is due to its broad applications in formation control [1], distributed control of unmanned air vehicles [2], flocking behavior [3], etc. One critical problem of distributed coordinated control of dynamical agents is to find a control law such that all agents reach an agreement on a common decision value. This problem is the so-called consensus problem.

   In the past decades, consensus problem of agents has been studied in different situations, e.g., networks with time-varying topologies [4], consensus over random networks [5], networks with time-delay [6], quantized consensus [7], high order consensus [8], to name just a few. In [4], the problem of information consensus among multiple agents system with dynamically changing interaction topologies were considered. In [5], the consensus problem for stochastic discrete-time linear dynamical systems was studied. In [6], the authors considered consensus problems in networks of continuous-time agents with diverse time-delays and jointly-connected topologies. In [7], the authors studied a discrete version of the distributed averaging problem that models such quantization. In [8], the consensus problems of multi-agent systems in the framework of high-dimensional state space were considered.

Recently, an interesting topic is the consensus problem when the information of agents in the system is unknown, and the models of agents are nonlinear. In this case, the consensus problem becomes challenging. Adaptive control design is one of effective methods to solve this problem [9-12]. In [9], the authors investigated the coordination problem where the objective is to steer a group of agents to a formation that translates with a prescribed reference velocity. In [10], the authors proposed an adaptive approach that guarantees tracking of the reference velocity by incorporating relative velocity feedback in addition to relative position feedback. In [11], a robust adaptive control approach is proposed to solve the consensus problem of multi-agent systems with uncertainties and external disturbances. In [12], the authors presented an adaptive synchronization controller for distributed systems with non-identical unknown nonlinear dynamics,

In this paper, we consider the adaptive consensus of multi-agent systems with non-identical unknown nonlinear dynamics and jointly connected topologies. The leader agent in the system is also assumed to be unknown and nonlinear. By parameterization of the unknown nonlinear dynamics by some basis functions, each agent estimating the unknown parameters, a neighbor-based adaptive consensus algorithm is proposed to solve the consensus problem based on algebraic graph theory and Lyapunov theory. By Lyapunov analysis and parameter convergence analysis, consensus errors and parameter estimate errors can be proved to be globally uniformly asymptotically convergent to zero.

2   Preliminaries

Information exchange among agents in a multi-agent system can be modeled by an interaction graph. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ be a weighted graph with a node set $\mathcal{V}$, and a weighted adjacency matrix $\mathcal{A} = [a_{ij}] \in R^{N \times N}$ with nonnegative elements. The edge of edge set $\mathcal{E}$ is denoted by (i, j). If there is an edge between two nodes, the two nodes are neighbors (or adjacent) to each other. The set of neighbors of node i is denoted by $\mathcal{N}_i = \{j \in V | (i, j) \in \mathcal{E}, j \neq i\}$. The elements of the adjacency matrix $\mathcal{A}$ are defined such that $a_{ij} > 0 \Leftrightarrow (i, j) \in \mathcal{E}$, while $a_{ij} = 0$ if $(i, j) \notin \mathcal{E}$. Its degree matrix $\mathcal{D} = \text{diag}\{d_1, d_2, \cdots, d_N\}$ is a diagonal matrix, where diagonal elements $d_i = \sum_{j=1}^{N} a_{ij}, \ i = 1, 2, \cdots, N$. Then the Laplacian associated with the adjacency matrix $\mathcal{A}$ is defined as $\mathcal{L} = \mathcal{D} - \mathcal{A}$. And it has the following property.

Lemma 1: Laplacian $\mathcal{L}$ of graph $\mathcal{G}$ has at least one zero eigenvalue with $1_N = (1, 1, \cdots, 1)^T \in R^N$ as its eigenvector, and all the non-zero eigenvalues of $\mathcal{L}$ are positive. Laplacian $\mathcal{L}$ has a simple zero eigenvalue if and only if graph $\mathcal{G}$ is connected.

A graph is undirected if edges $(i, j) \in \mathcal{E}$ are unordered pair (namely $a_{ij} = a_{ji}$). A graph is simple if it has no selfloops or repeated edges (namely $a_{ii} = 0$). A path is a sequence of interconnected edges in a graph. If there is a path between any two nodes of a graph $\mathcal{G}$, then $\mathcal{G}$ is said to be connected, otherwise disconnected. The union of a collection of graphs is a graph with node set and edge set being the union of node set

and edge set of all of the graphs in the collection. We say that a collection of graphs is jointly connected if the union of its members is connected.

In this paper, in order to describe the information transmission between N agents and a leader, we need define another graph $\bar{\mathcal{G}}$ on nodes $0, 1, 2, \cdots, N$ which consists of graph $\mathcal{G}$, node 0 which represent the leader and edges between the leader and its neighbors. Note that the graph describing the interaction topology can vary with time. To account this we need to consider all possible graphs $\{\bar{\mathcal{G}}_p | p \in \mathcal{P}\}$, where $\mathcal{P}$ is an index set for all graphs defined on vertices $\{0, 1, 2, \cdots, N\}$. We use $\{\mathcal{G}_p | p \in \mathcal{P}\}$ to denote subgraphs defined on vertices $\{1, 2, \cdots, N\}$. The dependence of the graphs upon time can be characterized by a switching signal $\sigma : [0, \infty) \to \mathcal{P}$, that is, at each time t the underlying graph is $\bar{\mathcal{G}}_{\sigma(t)}$. It is assumed in this paper that σ switches finite times in any bounded time interval.

Lemma 2: (Barbalat's lemma) If a scalar function $f(t)$ is uniformly continuous such that $\int_0^\infty f(t)dt$ exists and finite, then $\lim_{t \to \infty} f(t) = 0$.

Consider a multi-agent system consisting of N agents and a leader. The dynamics of each agent is

$$\dot{x}_i(t) = f_i(x_i, t) + u_i(t), i = 1, 2, \cdots, N, \tag{1}$$

where $x_i(t) \in R$ is the state of ith agent, $u_i(t) \in R$ the control input, and $f_i(x_i, t)$ the dynamics of agent i, which is assumed to be unknown and nonlinear. Standard assumptions for existence of unique solutions are made, i.e., $f_i(x_i, t)$ is continuous in t and Lipschitz in x_i. The dynamics of the leader is

$$\dot{x}_0(t) = v_0(t), \tag{2}$$

where $x_0(t) \in R$ is the state of the leader, $v_0(t) \in R$ its velocity and assumed to be unknown.

In this paper, we study the problem of designing controllers $u_i, i = 1, 2, \cdots, N$, such that all agents follow the leader.

Definition 1: The consensus of the multi-agent system (1) and (2) is said to be achieved, if for each agent $i \in \{1, 2, \cdots, N\}$ there is a controller u_i such that the system satisfies

$$\lim_{t \to \infty} \|x_i(t) - x_0(t)\| = 0, \tag{3}$$

for any initial condition $x_i(0), i = 1, 2, \cdots, N$.

In order to achieve the adaptive consensus, we assumed the unknown nonlinear dynamic $f_i(x_i, t)$, and the unknown velocity $v_0(t)$ of the leader satisfy the linearly parameterized assumption [14], [15], and can be parameterized as

$$f_i(x_i, t) = \phi_i^T(x_i, t)\theta_i, i = 1, 2, \cdots, N, \tag{4}$$

$$v_0(t) = \phi_0^T(t)\theta_0, \tag{5}$$

where $\phi_0(t)$ and $\phi_i(x_i, t)$ are basis function column vectors, $\theta_0, \theta_i \in R^m$ are true parameter column vectors which are to be estimated.

Because θ_0 is unavailable to each agent, so we use $\hat{\theta}_{0i}$, $\hat{v}_i(t)$ to estimates the unknown parameter vector θ_0 and $v_0(t)$ of the ith agent, respectively.

$$\hat{v}_i(t) = \phi_0^T(t)\hat{\theta}_{0i}, i = 1, 2, \cdots, N. \tag{6}$$

Similarly, the estimate of $f_i(x_i, t)$ can be described by

$$\hat{f}_i(x_i, t) = \phi_i^T(x_i, t)\hat{\theta}_i, i = 1, 2, \cdots, N. \tag{7}$$

In this work, we consider the problem of designing adaptive consensus controllers such that all agents follow the leader, that is, for any initial condition $x_0(0), x_i(0), \hat{\theta}_{0i}(0) \in R^m, \hat{\theta}_i(0) \in R^m, i = 1, 2, \cdots, N$.

$$\lim_{t \to \infty} \|x_i - x_0\| = 0, \lim_{t \to \infty} \|\hat{\theta}_{0i} - \theta_0\| = 0, \lim_{t \to \infty} \|\hat{\theta}_i - \theta_i\| = 0, \tag{8}$$

For parameter convergence analysis, we need following PE condition.

Definition 2: (PE Condition) Assume a matrix Φ is persistently exciting (PE), that is to say there are two positive real number δ_0 and α such that

$$\int_t^{t+\delta_0} \Phi\Phi^T d\tau \geq \alpha I > 0, \forall t \geq 0. \tag{9}$$

3   Main Results

In the case of jointly connected topologies, we assumed that there is an infinite sequence of bounded, non-overlapping, contiguous time intervals $[t_k, t_{k+1})$, and for some positive constant T_0, T, satisfy $t_0 = 0, T_0 \leq t_{k+1} - t_k \leq T, k = 0, 1, 2, \cdots$. The time-varying topology switches at time instants $t_k, k = 0, 1, 2, \cdots$, that is to say the time varying graph topology is time invariant in each time interval $[t_k, t_{k+1})$. Since the interconnection topology is variable, we define a piecewise-constant switching signal $\sigma : [0, \infty) \to \mathcal{P}$.

Denote $x = \text{col}(x_i), \bar{x} = x - \mathbf{1}_N x_0, u = \text{col}(u_i), \hat{\Theta}_0 = \text{col}(\hat{\theta}_{0i}), \hat{\Theta}_f = \text{col}(\hat{\theta}_i), i \in \{1, 2, \cdots, N\}$. We assume the adaptive consensus control scheme include two parts.

♦Neighbor-based feedback laws:

$$\begin{aligned} u_i(t) = &c\sum_{j \in N_i(t)} a_{ij}(t)(x_j - x_i) + cb_i(t)(x_0 - x_i) \\ &+ \phi_0^T(t)\hat{\theta}_{0i} - \phi_i^T(x_i, t)\hat{\theta}_i, i = 1, 2, \cdots, N. \end{aligned} \tag{10}$$

or in matrix form

$$u(t) = -cH_{\sigma(t)}\bar{x} + \Phi_0^T\hat{\Theta}_0 - \Phi_f^T\hat{\Theta}_f, \tag{11}$$

◆ The adaptive laws:

$$\begin{cases} \dot{\hat{\theta}}_{0i} = -\frac{c_0}{c}\phi_0(t)(x_i - x_0) \\ \dot{\hat{\theta}}_i = \frac{c_1}{c}\phi_i(t)(x_i - x_0) \\ i = 1, 2, \cdots N, \end{cases} \tag{12}$$

or in matrix form

$$\begin{cases} \dot{\hat{\Theta}}_0 = -\frac{c_0}{c}\Phi_0\bar{x} \\ \dot{\hat{\Theta}}_f = \frac{c_1}{c}\Phi_f\bar{x}, \end{cases} \tag{13}$$

where c, c_1, c_2 are constants, $H_{\sigma(t)} = L_{\sigma(t)} + B_{\sigma(t)}, B_{\sigma(t)}$ is a diagonal matrix with diagonal elements $b_1(t), b_2(t), \cdots, b_N(t), \Phi_f = \text{diag}\{\phi_1, \phi_2, \cdots, \phi_N\}, \Phi_0 = I_N \otimes \phi_0$, I_N is the $N \times N$ identity matrix.

Denote $\Theta_0 = \mathbf{1}_N \otimes \theta_0, \bar{\theta}_{0i} = \hat{\theta}_{0i} - \theta_0, \bar{\Theta}_0 = \hat{\Theta}_0 - \Theta_0 = \text{col}(\bar{\theta}_{0i}), \Theta_f = \text{col}(\theta_i), \bar{\theta}_i = \hat{\theta}_i - \theta_i, \bar{\Theta}_f = \hat{\Theta}_f - \Theta_f = \text{col}(\bar{\theta}_i)\bar{\Theta} = \text{col}(\bar{\Theta}_0, \bar{\Theta}_f), \Phi = \text{col}(\Phi_0, \Phi_f)$. Then we have the following error dynamics of system (1) and (2)

$$\begin{cases} \dot{\bar{x}} &= -cH_{\sigma(t)}\bar{x} + \Phi_0^T\bar{\Theta}_0 - \Phi_f^T\bar{\Theta}_f \\ \dot{\bar{\Theta}}_0 &= -\frac{c_0}{c}\Phi_0\bar{x} \\ \dot{\bar{\Theta}}_f &= \frac{c_1}{c}\Phi_f\bar{x}. \end{cases} \tag{14}$$

We consider the case when the interaction topology is jointly connected. Because of the requirement that the switching interconnected graph $\bar{G}_\sigma$ is connected through each time interval $[t_k, t_{k+1}), k = 0, 1, 2, \cdots$, is a strong condition. In order to achieve adaptive consensus under jointly connected topology, we suppose that in each time interval $[t_k, t_{k+1})$, there exists a sequence of nonoverlapping, contiguous subinterval $[t_k^0, t_k^1), \cdots, [t_k^l, t_k^{l+1}), \cdots, [t_k^{l_k-1}, t_k^{l_k})$ with $t_k = t_k^0, t_{k+1} = t_k^{l_k}$ for some integer $l_k \geq 0$ such that in each of such subintervals the switching graph $\bar{G}_{\sigma(t)}$ is time invariant. We assume that there is a constant number $\tau > 0$, which is often called dwell time, with $t_k^{l+1} - t_k^l \geq \tau, 0 \leq l \leq l_k - 1$. Note that in each of such subintervals the interconnected graph $\bar{G}_\sigma$ is permitted to be disconnected. A collection of switching graphs $\{\bar{G}_{\sigma(t)}|s \in [t, t + \triangle t], \triangle t > 0\}$ is said to be jointly connected across a time interval $[t, t + \triangle t]$ if its union is connected. $H_p, p \in \mathcal{P}$ has N eigenvalues denoted as $\lambda_p^1, \lambda_p^2, \cdots, \lambda_p^N$ based on some labeling rule [16]. Define $l(p) = \{k|\lambda_p^k \neq 0, k = 1, 2, \cdots, N\}$, we have following Lemma [16]:

Lemma 3: Graphs $G_p, p \in \mathcal{P}$ are jointly connected across $[t_k, t_{k+1})$, if and only if $\bigcup_{t \in [t_k, t_{k+1})} l(\sigma(t)) = 1, 2, \cdots, N$.

Theorem 1: Consider the multi-agent system (1)-(2). Assume that the switching interconnected graph $\bar{G}_\sigma(t)$ is jointly connected across each time interval $[t_k, t_{k+1}), k = 0, 1, 2, \cdots, \phi_i$ and $\dot{\phi}_i, i = 1, 2, \cdots, N$ are uniformly bounded and the PE condition defined in (9) is satisfied, then, by control law (10) and parameter adaptive law (12), $(\bar{x}, \bar{\Theta}) = 0$ is a globally uniformly asymptotically stable equilibrium, i.e., consensus is reached with parameter convergence globally uniformly asymptotically.

Proof: For the error dynamic system (14), consider the following Lyapunov function candidate

$$V(t) = \frac{1}{2c}\bar{x}^T\bar{x} + \frac{1}{2c_0}\bar{\Theta}_0^T\bar{\Theta}_0 + \frac{1}{2c_1}\bar{\Theta}_f^T\bar{\Theta}_f. \tag{15}$$

Obviously, $V(t)$ is continuously differentiable at any time except switching instants. We know that $H_p, p \in \mathcal{P}$ are symmetrical across each time interval $[t_k, t_{k+1})$, then there exists an orthogonal matrix U_p such that $U_p H_p U_p^T = \Lambda_p = \text{diag}\{\lambda_p^{i_1}, \lambda_p^{i_2}, \cdots, \lambda_p^{i_N}\}$, where $\lambda_p^{i_1}, \lambda_p^{i_2}, \cdots, \lambda_p^{i_N}$ are eigenvalues of $H_p, i_1, i_2, \cdots, i_N$ is a permutation of $1, 2, \cdots, N$.

At non-switching time t, we assume that subsystem $p \in \mathcal{P}$ is active, the time derivative of Lyapunov function candidate along the trajectory of the system (14) is

$$\dot{V}(t) = \frac{1}{c}\bar{x}^T\dot{\bar{x}} + \frac{1}{c_0}\bar{\Theta}_0^T\dot{\bar{\Theta}}_0 + \frac{1}{c_1}\bar{\Theta}_f^T\dot{\bar{\Theta}}_f = -\bar{x}^T H_p \bar{x}. \tag{16}$$

Let $\tilde{x} = U_p\bar{x}$, we have

$$\begin{aligned}
\dot{V}(t) &= -\bar{x}^T H_p \bar{x} = -\sum_{i \in l(p)} \lambda_p^i \tilde{x}_i^T \tilde{x}_i \\
&\leq -\delta_p \sum_{i \in l(p)} \tilde{x}_i^T \tilde{x}_i \leq -\delta_{\min} \sum_{i \in l(p)} \tilde{x}_i^T \tilde{x}_i \\
&\leq 0.
\end{aligned} \tag{17}$$

where $\delta_p = \min\{\lambda_p^i, i = 1, 2, \cdots, N\}, \delta_{\min} = \min\{\delta_p | p \in \mathcal{P}\}$. Thus $\lim_{t \to \infty} V(t) = V(\infty)$ exists.

Consider infinite sequence $\{V(t_k), k = 0, 1, 2, \cdots\}$, from Cauchy's convergence criteria, we have, for $\forall \epsilon > 0$, there exists a positive integer K such that, for$\forall k > K$, $|V(t_{k+1}) - V(t_k)| < \epsilon$, or equivalently $|\int_{t_k}^{t_{k+1}} \dot{V}(t)dt| < \epsilon$. The integral can be rewritten as $\sum_{l=0}^{l_k-1} \int_{t_k^l}^{t_k^{l+1}} \dot{V}(t)dt > -\epsilon$. From (17), we know

$$\sum_{l=0}^{l_k-1} \int_{t_k^l}^{t_k^{l+\tau}} \sum_{i \in l(\sigma(t_k^l))} \tilde{x}_i^T \tilde{x}_i dt \leq \sum_{l=0}^{l_k-1} \int_{t_k^l}^{t_k^{l+1}} \sum_{i \in l(\sigma(t_k^l))} \tilde{x}_i^T \tilde{x}_i dt \leq \frac{\epsilon}{\delta_{\min}}. \tag{18}$$

Because l_k is assumed to be finite in each time interval $[t_k, t_{k+1})$. Thus, for $\forall k > K$, we have

$$\int_{t_k^l}^{t_k^{l+\tau}} \sum_{i \in l(\sigma(t_k^l))} \tilde{x}_i^T \tilde{x}_i dt \leq \frac{\epsilon}{\delta_{\min}}, l = 0, 1, \cdots, l_k - 1, \tag{19}$$

which implies

$$\lim_{t \to \infty} \sum_{l=0}^{l_k-1} \int_t^{t+\tau} \sum_{i \in l(\sigma(t_k^l))} \tilde{x}_i^T(s)\tilde{x}_i(s)ds = 0. \tag{20}$$

From Lemma 3 $\bigcup\limits_{t\in[t_k,t_{k+1})} l(\sigma(t)) = 1, 2, \cdots, N$, due to the joint connectivity of the graphs through the time interval $[t_k, t_{k+1})$, we have

$$\lim_{t\to\infty} \int_t^{t+\tau} \sum_{i=1}^N a_i \tilde{x}_i^T(s)\tilde{x}_i(s)ds = 0, \tag{21}$$

where a_i are some positive integers. Moreover, from (15) and (17) it follows that both $\bar{x}$ and $\bar{\Theta}$ are uniformly bounded for any $t \geq 0$ and so is $\dot{\bar{x}}$ due to (14) and the assumption that both ϕ_i and $\dot{\phi}_i$ are uniformly bounded. Therefore $\sum\limits_{i=1}^N a_i \tilde{x}_i^T(s)\tilde{x}_i(s)$ is uniformly continuous. From Lemma 2, we have $\lim\limits_{t\to\infty} \sum\limits_{i=1}^N a_i \tilde{x}_i^T(t)\tilde{x}_i(t)dt = 0$, then $\lim\limits_{t\to\infty} \tilde{x}_i(t)dt = 0. \lim\limits_{t\to\infty} \bar{x}_i(t)dt = 0, i = 1, 2, \cdots, N.$ Thus

$$\lim_{t\to\infty} \|x_i(t) - x_0(t)\| = 0, i = 1, 2, \cdots, N. \tag{22}$$

Next, we show that for any initial condition

$$\lim_{t\to\infty} \|\bar{\Theta}(t)\| = 0, \tag{23}$$

i.e., for $\forall\epsilon > 0$ there exists $\mathcal{T}_\epsilon > 0, \forall t > \mathcal{T}_\epsilon$, such that $\|\bar{\Theta}(t)\| < \epsilon$. During showing the parameter convergence, we need the following claim.

Claim 1: Given any $\epsilon > 0$ and $\mathcal{T} > 0$, for any initial condition $\bar{x}(0), \bar{\Theta}(0)$, there exists $t > \mathcal{T}$ such that $\text{col}\|(\bar{\theta}_{0i}, \bar{\theta}_i)\| < \epsilon, i = 1, 2, \cdots, N$.

Proof: We equivalently show by contradiction that for any $\epsilon > 0$, and some $i \in \{1, 2, \cdots, N\}$, a time $\mathcal{T}_1$ such that

$$\text{col}\|(\bar{\theta}_{0i}, \bar{\theta}_i)\| > \epsilon, \forall t > \mathcal{T}_1 \tag{24}$$

does not exist.

Without loss of generality, for infinite sequence of time intervals $[t_k, t_{k+1})$, $k = 0, 1, \cdots$, consider an infinite subsequence of time intervals $[t_{k_j}, t_{k_j+1}), j = 0, 1, \cdots$ with identical length T_1 satisfying $T_0 \leq T_1 \leq T$, that is $t_{k_j+1} = t_{k_j} + T_1$.

Define function

$$\Psi(\bar{\Theta}(t), t) = \frac{1}{2}[\bar{\Theta}^T(t+T_1)\bar{\Theta}(t+T_1) - \bar{\Theta}^T(t)\bar{\Theta}(t)]. \tag{25}$$

Due to $\lim\limits_{t\to\infty} \bar{x}(t) = 0, \lim\limits_{t\to\infty} V(t) = V(\infty)$ exists and (15), we can obtain that $\lim\limits_{t\to\infty} \bar{\Theta}^T(t)\bar{\Theta}(t) = \eta V(\infty)$, where η is a positive constant number and so we can conclude that $\lim\limits_{t\to\infty} \Psi(\bar{\Theta}(t), t) = 0$ because of (25). From Cauchy's convergence criteria, we have, for $\forall\epsilon_1 > 0$, there exists a $t_{\epsilon_1} > 0$, such that

$$\|\Psi(\bar{\Theta}(t), t) - \Psi(\bar{\Theta}(t'), t')\| < \epsilon_1, \forall t, t' > t_{\epsilon_1}. \tag{26}$$

The time derivative of the function $\Psi(\bar{\Theta}(t), t)$ defined in (25) at time instant t_{k_j} is

$$
\begin{aligned}
&\dot{\Psi}(\bar{\Theta}(t_{k_j}), t_{k_j}) \\
&= \int_{t_{k_j}}^{t_{k_j}+T_1} \tfrac{d}{d\tau}[\bar{\Theta}^T(\tau)\dot{\bar{\Theta}}(\tau)]d\tau \\
&= \sum_{l=0}^{l_k-1} \int_{t_{k_j}^l}^{t_{k_j}^{l+1}} \tfrac{d}{d\tau}[\bar{\Theta}_0^T(-\tfrac{c_0}{c})\Phi_0\bar{x} + \bar{\Theta}_f^T(\tfrac{c_1}{c})\Phi_f\bar{x}]d\tau \\
&= -\tfrac{1}{c}\sum_{l=0}^{l_k-1} \int_{t_{k_j}^l}^{t_{k_j}^{l+1}} \tfrac{d}{d\tau}[(c_0\bar{\Theta}_0^T\Phi_0 - c_1\bar{\Theta}_f^T\Phi_f)\bar{x}]d\tau \\
&= \sum_{l=0}^{l_k-1} \int_{t_{k_j}^l}^{t_{k_j}^{l+1}} \{[\tfrac{c_0^2}{c^2}\bar{x}^T\Phi_0^T\Phi_0 + \tfrac{c_1^2}{c^2}\bar{x}^T\Phi_f^T\Phi_f - \tfrac{c_0}{c}\bar{\Theta}_0^T\dot{\Phi}_0 + \tfrac{c_1}{c}\bar{\Theta}_f^T\dot{\Phi}_f]\bar{x} \\
&\qquad + [c_0\bar{\Theta}_0^T\Phi_0 - c_1\bar{\Theta}_f^T\Phi_f]H_p\bar{x}\}d\tau \\
&\quad - \sum_{l=0}^{l_k-1} \int_{t_{k_j}^l}^{t_{k_j}^{l+1}} (\tfrac{c_0}{c}\bar{\Theta}_0^T, -\tfrac{c_1}{c}\bar{\Theta}_f^T)\Phi\Phi^T \begin{pmatrix} \bar{\Theta}_0 \\ -\tilde{\Theta}_f \end{pmatrix} d\tau \\
&\triangleq I_1 - I_2.
\end{aligned}
\tag{27}
$$

Because $\bar{x}$, $\bar{\Theta}$, ϕ_i, $\dot{\phi}_i$ are uniformly bounded, letting $M_{\bar{x}}, M_\Theta, M_\Phi > 0$ be such that $\|\bar{x}\| \le M_{\bar{x}}$, $\|\bar{\Theta}_0\| \le M_\Theta$, $\|\bar{\Theta}_f\| \le M_\Theta$, $\|\Phi_0\| \le M_\Phi$, $\|\Phi_f\| \le M_\Phi$, $\|\dot{\Phi}_0\| \le M_\Phi$, $\|\dot{\Phi}_f\| \le M_\Phi$, $\forall t \ge 0$. Then we can rewrite the first integral in (27) as

$$
I_1 \le M \sum_{l=0}^{l_k-1} \int_{t_{k_j}^l}^{t_{k_j}^{l+1}} \|\bar{x}\| d\tau,
\tag{28}
$$

where $M = [\tfrac{c_0^2+c_1^2}{c^2}M_{\bar{x}}M_\Phi + \tfrac{(c_0+c_1)(1+c\delta_{\max})}{c}M_\Theta]M_{\bar{x}}M_\Phi$ is a positive constant.

Besides, we know that l_k, $k = 1, 2, \cdots$, are assumed to be finite and $\lim_{t\to\infty} \|\bar{x}(t)\| = 0$, we can obtain

$$
I_1 \le \frac{c_0}{2c}\alpha\epsilon^2, \forall t_{k_j} \ge \mathcal{T}_2.
\tag{29}
$$

Without loss of generality, we select $c_0 = c_1$ and let $\tilde{\Theta} = \mathrm{col}(\bar{\Theta}_0, -\bar{\Theta}_f)$, then we have

$$
I_2 = \frac{c_0}{c} \sum_{l=0}^{l_k-1} \int_{t_{k_j}^l}^{t_{k_j}^{l+1}} \tilde{\Theta}^T\Phi\Phi^T\tilde{\Theta} d\tau.
\tag{30}
$$

Let $(\Phi^T\tilde{\Theta})_i$ express the ith component of $\Phi^T\tilde{\Theta}$, $\mathrm{col}(\bar{\theta}_{0i})$ and $\mathrm{col}(\bar{\theta}_i)$ be column vectors with components indexed by $i \in l(\sigma(t_{k_j}^l))$, $\tilde{\Theta}^p = \mathrm{col}[\mathrm{col}(\bar{\theta}_{0i})_{i \in l(\sigma(t_{k_j}^l))}$, $\mathrm{col}(\bar{\theta}_i)_{i \in l(\sigma(t_{k_j}^l))}]$, $M_{i,N+i}$ be the principle submatrix of $\Phi\Phi^T$ by retaining like-numbered rows and columns indexed by $\{i, N+i : i \in l(\sigma(t_{k_j}^l))\}$, then we have

$$
\begin{aligned}
I_2 &= \frac{c_0}{c} \sum_{l=0}^{l_k-1} \int_{t_{k_j}^l}^{t_{k_j}^{l+1}} \sum_{i \in l(\sigma(t_{k_j}^l))} (\Phi^T\tilde{\Theta})_i^2 d\tau \\
&= \frac{c_0}{c} \sum_{l=0}^{l_k-1} \int_{t_{k_j}^l}^{t_{k_j}^{l+1}} (\tilde{\Theta}^p)^T M_{i,N+i}\tilde{\Theta}^p d\tau.
\end{aligned}
\tag{31}
$$

From the PE condition (9), we have

$$\int_t^{t+\delta_0} M_{i,N+i}d\tau > \alpha I > 0, \forall t > 0. \tag{32}$$

Assume now by contradiction that exists a time $\mathcal{T}_1$ so that (30) holds, which implies that for some $p, \|\tilde{\Theta}^p\| > \epsilon$, because of Lemma 1. From (31), (32) and for some l,

$$
\begin{aligned}
I_2 &= \frac{c_0}{c} \int_{t_{k_j}^l}^{t_{k_j}^{l+\tau}} (\tilde{\Theta}^p)^T M_{i,N+i} \tilde{\Theta}^p ds \\
&\geq \frac{c_0}{c}\epsilon^2 \int_{t_{k_j}^l}^{t_{k_j}^{l+\tau}} \frac{(\tilde{\Theta}^p)^T}{\|\tilde{\Theta}^p\|} M_{i,N+i} \frac{\tilde{\Theta}^p}{\|\tilde{\Theta}^p\|} ds \\
&\geq \frac{c_0}{c}\alpha\epsilon^2, \forall t_{k_j} \geq \mathcal{T}_1.
\end{aligned}
\tag{33}
$$

From (27), (29) and (33), we obtain

$$\dot{\Psi}(\bar{\Theta}(t_{k_j}), t_{k_j}) \leq -\frac{c_0}{2c}\alpha\epsilon^2 < 0, \forall t_{k_j} \geq \mathcal{T}_3. \tag{34}$$

with $\mathcal{T}_3 = \max\{\mathcal{T}_1, \mathcal{T}_2, t_{\epsilon_1}\}$. Thus, from sign-preserving theorem of continuous function, there exists an interval $[t_{k_j}, t_{k_j} + \delta)$ with $t_{k_j} \geq \mathcal{T}_3, \delta > 0$ such that $\dot{\Psi}(\bar{\Theta}(t), t) < -\frac{c_0}{4c}\alpha\epsilon^2 < 0$ holds for $\forall t \in [t_{k_j}, t_{k_j} + \delta)$, we select $\epsilon_1 = \frac{c_0}{4c}\alpha\epsilon^2\delta$, then we have

$$\Psi(\bar{\Theta}(t_{k_j}), t_{k_j}) - \Psi(\bar{\Theta}(t_{k_j} + \delta), t_{k_j} + \delta) > \epsilon_1. \tag{35}$$

which contradicts (26). This completes the proof of the claim.
The detailed proof of claim 1 is omitted here due to space limitation.
Now we show the parameter convergence. From (16), we have

$$\bar{\Theta}^T\bar{\Theta} \leq \frac{\max\{c_0, c_1\}}{c}\bar{x}^T\bar{x} + \frac{\max\{c_0, c_1\}}{\min\{c_0, c_1\}}\bar{\Theta}^T\bar{\Theta}. \tag{36}$$

$$\|\bar{\Theta}\|^2 \leq \frac{\max\{c_0, c_1\}}{c}\|\bar{x}(\mathcal{T}_\epsilon)\|^2 + \frac{\max\{c_0, c_1\}}{\min\{c_0, c_1\}}\|\bar{\Theta}(\mathcal{T}_\epsilon)\|^2. \tag{37}$$

According to (22), for any $\epsilon > 0$, there exists a time instant $t_\epsilon > 0$ such that

$$\|\bar{x}(t)\| \leq \sqrt{\frac{c}{2\max\{c_0, c_1\}}}\epsilon, \; t > t_\epsilon. \tag{38}$$

By virtual the Claim, there exists a time instant $\mathcal{T}_\epsilon > t_\epsilon$ such that

$$\|\bar{\Theta}(t)\| < \sqrt{\frac{\min\{c_0, c_1\}}{2\max\{c_0, c_1\}}}\epsilon. \tag{39}$$

From the initial condition $\bar{x}(\mathcal{T}_\epsilon)$ and $\bar{\Theta}(\mathcal{T}_\epsilon)$, according to (37), (38) and (39),

$$\|\bar{\Theta}(t)\| \leq \sqrt{\frac{max\{c_0,c_1\}}{c}\|\bar{x}(\mathcal{T}_\epsilon)\|^2 + \frac{max\{c_0,c_1\}}{min\{c_0,c_1\}}\|\bar{\Theta}(\mathcal{T}_\epsilon)\|^2} \leq \epsilon, \ \forall t > \mathcal{T}_\epsilon, \quad (40)$$

which implies the (23). Therefore the equilibrium $(\bar{x}, \bar{\Theta}) = 0$ is attractive. Since (21) holds uniformly with respect to initial time instant, so also do (22) and (23). It follows that $(\bar{x}, \bar{\Theta}) = 0$ is a globally uniformly asymptotically stable equilibrium. So $\lim\limits_{t\to\infty} \|\bar{x}(t)\| = 0, \lim\limits_{t\to\infty} \|\bar{\Theta}(t)\| = 0, \lim\limits_{t\to\infty} \|x_i - x_0\| = 0, \forall x_i(0) \in R, \lim\limits_{t\to\infty} \|\hat{\theta}_{0i} - \theta_0\| = 0,$ $\lim\limits_{t\to\infty} \|\hat{\theta}_i - \theta_i\| = 0, \forall \hat{\theta}_{0i}(0) \in R^m, \hat{\theta}_i(0) \in R^m, i = 1, 2, \cdots, N.$

4   Simulations

In this section, we give a numerical simulation of multi-agent systems with jointly connected topologies to validate our theoretical results. Consider a multi-agent system consisting of a leader and five agents. We suppose that the leader's unknown velocity dynamics is parameterized as

$$v_0(t) = [\sin^2(t), \cos^2(t)]\theta_0, \quad (41)$$

where θ_0 is selected as $[\frac{\sqrt{2}}{2}, \frac{\sqrt{3}}{2}]^T$. The unknown nonlinear dynamics of followers are parameterized as

$$f_i(x_i, t) = [x_i \sin^2(t), x_i \cos^2(t)]\theta_i, i = 1, 2, \cdots, 5, \quad (42)$$

where $\theta_i = [1, 1]^T, i = 1, 2, \cdots, 5$, is the true parameters to be estimated.
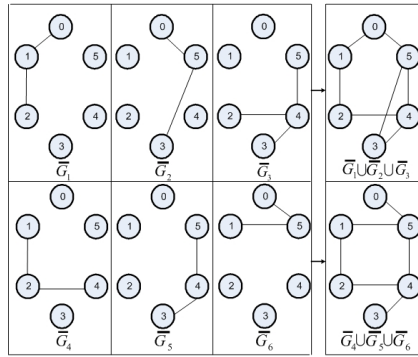
Fig. 1. Jointly connected graphs
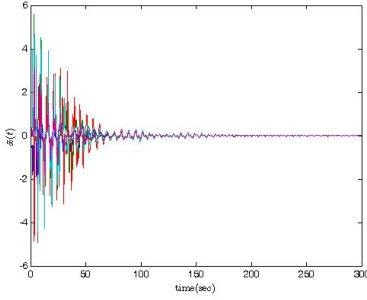
Fig. 2. Consensus is reached asymptotically
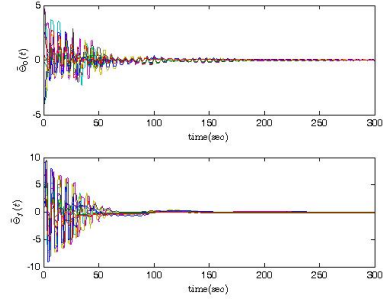
Fig. 3. The parameters converge to zero asymptotically

We suppose the possible interconnected graphs are $\{\bar{G}_1, \bar{G}_2, \bar{G}_3, \bar{G}_4, \bar{G}_5, \bar{G}_6\}$, which are shown in Fig.1, Note that $\bar{G}_1 \cup \bar{G}_2 \cup \bar{G}_3$ and $\bar{G}_4 \cup \bar{G}_5 \cup \bar{G}_6$ are jointly connected through some time intervals. The simulation results are shown in Fig.2 and Fig.3. From Fig.2 and Fig.3, we can see that the control law and adaptive laws we proposed are feasible and effective.

5   Conclusions

In this paper, we consider the leader-following adaptive consensus problem for multi-agent system with unknown nonlinear dynamics and jointly connected topologies. A novel adaptive consensus algorithm is proposed by using linear parameterizations of unknown nonlinear dynamics of all agents. By stability analysis and parameter convergence analysis of the proposed algorithm, adaptive consensus can be realized based on neighboring graphs using algebraic graph theory and Lyapunov theory. The connectedness of interconnected graphs ensures consensus achievement. Parameter convergence analysis shows that the PE condition plays a key role.

References

1. Fax, A., Murray, R.M.: Information flow and cooperative control of vehicle formations. IEEE Transaction on Automatic Control 49(9), 1465–1476 (2004)
2. Ren, W.: Consensus strategies for cooperative control of vehicle formations. Control Theory & Applications. IET 1(2), 505–512 (2007)
3. Tanner, H.G., Jadbabaie, A., Pappas, G.J.: Stable flocking of mobile agents, part I: Fixed topology. In: Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, Hawaii USA, vol. 2, pp. 2010–2015 (2003)

4. Ren, W., Beard, R.W.: Consensus seeking in multi-agent systems under dynamically changing interaction topologies. IEEE Transaction on Automatic Control 50(5), 655–661 (2005)
5. Salehi, A.T., Jadbabaie, A.: A necessary and sufficient condition for consensus over random networks. IEEE Transaction on Automatic Control 53(3), 791–795 (2008)
6. Lin, P., Jia, Y.: Multi-agent consensus with diverse time-delays and jointly-connected topologies. Automatica 47, 848–856 (2011)
7. Kashyap, A., Basar, T., Srikant, R.: Quantized consensus. Automatica 43, 1192–1203 (2007)
8. Xiao, F., Wang, L.: Consensus problems for high-dimensional multi-agent systems. Control Theory & Applications, IET 1(3), 830–837 (2007)
9. Bai, H., Arcak, M., Wen, J.T.: Adaptive design for reference velocity recovery in motion coordination. Systems & Control Letters 57, 602–610 (2008)
10. Bai, H., Arcak, M., Wen, J.T.: Adaptive motion coordination: Using relative velocity feedback to track a reference velocity. Automatica 45, 1020–1025 (2009)
11. Das, A., Lewis, F.L.: Distributed adaptive control for synchronization of unknown nonlinear networked systems. Automatica 46, 2014–2021 (2010)
12. Hou, Z., Cheng, L., Tan, M.: Decentralized robust adaptive control for the multi-agent system consensus problem using neural networks. IEEE Transaction on Systems, Man, and Cybernetics Part B: Cybernetics 39(3), 636–647 (2009)
13. Godsil, C.D., Royle, G.: Algebraic Graph Theory. Springer, New York (2001)
14. Marino, R., Tomei, P.: Nonlinear Control Design: Geometric. Prentice-Hall, Europe (1995)
15. Sastry, S., Bodson, M.: Adaptive Control: Stability, Convergence and Robustness. Prentice-Hall, New Jersey (1989)
16. Ni, W., Cheng, D.: Leader-following consensus of multi-agent systems under fixed and switching topologies. Systems & Control Letters 59, 209–217 (2010)

Web Context Analysis Based on Generic Ontology

Liang Zhu, Wanli Zuo, Fengling He, Jiayu Han, and Jingya Lu

College of Computer Science and Technology, Jilin University, Changchun, 130012,
China Key Laboratory of Symbolic Computation and Knowledge Engineering
of the Ministry of Education of China
`{zhuliang11,hanjy0821,lujy12}@mails.jlu.edu.cn,`
`{zuowl,hefl}@jlu.edu.cn`

Abstract. Now, the popular search engine system does not take into account the context information of search word, so the returned list of web pages contain a large number of irrelevant web pages. One word can reflect its true semantics only in context. This paper proposed the concepts of "contextual word", "web page parsing with contextualinformation", and "context representation" for the first time. Based on general ontology, we use the techniques of the word sense disambiguation to determine the context of the word to realize the web-page parsing in the level of word sense and sentence semantics according to the background of the word in the web-page. First, we transformthe web page into DOM tree, do the web parsing in the tradition method to remove the noise in the web page, extract the main body of the web page, and then use the real time search technology to get the last-modified-time of the web page. Second, we do the Lexical analysis on the body of the web page. Based on general ontology and natural language processing techniques, we mark the word or terms, and get the interpretation corresponding to the context. Third, we use the Named Entity Recognition Technology to get the time and the location information in the web pages,then we organize the information which we obtained into a structure called web context representation that we proposed. Based on the above theoretical basis, the author implements a complete set of web page contextual parse tool—JLUCAS. After a large number analysis of comparative experiments, JLUCAS achieves excellent results in everyaspect. This fully demonstrates thatthe theory and algorithm proposed in this paper can solve the problem of automatic web page contextual parsing, and lay a good basis for ultimately realizing the web contextual search engine.

Keywords: contextual word, contextual analysis, context representation, JLUCAS, contextual search.

1   Introduction

Most of today's search engine systems retrieve information that we need based on keywords and inverted indexes. When the user enters one or a set of keywords, the search engine returns a list of all web pages those contain search terms through searching in the inverted indexes. Because of the multiple meaning of search word, all

web pages containing every meaning of the search word will be returned and included in the result. For instance, for the search word "apple", the returned results will include all web pages those contain apple. Some of these pages are relevant to fruit and some others are relevant to electronic products, such as mobile phones. For a fruit farmer, he may want to get the pages related to fruit, and on the contrary, for apple fans, they may want to get the pages related to electronic products, such as mobile phones. As the retrieval system of search engine does not know the user's interest and intent, the only thing it will do is to give the user the whole web pages containing the word apple. Then, the user selects the pages those he want to get in the results. By this way, not only increase the burden on users, but also a waste of network bandwidth.

The above problem reflects the limitations of current search engines: lack of understanding of users' interests and intentions and do not know the respective field of web pages. To address the problem, contextual search came into being [1]. The purpose of contextual search is to understand and master the user's interest and intention, and then return only pages related to the query intention. For example, in the above example of the apple, for one farmer, his interest is generally fruit, the system returns only fruit-relevant pages; for one apple fan, his interest is generally mobile phones or other electronic products, the system will only return the pages related to electronic products. It can achieve the goal of personalized search.

Contextual search was first proposed by Reiner Kraft [1]. In the academic field, the theoretical research on contextual search is under way. We believe that the contextual search is an important change in traditional search area. To support contextual search effectively, it will break the traditional technology from the external search mode to the internal index structure, otherwise, we can only take the following path, "search first, and then filter". By this way, not only inefficient, but also does not give the user real contextual search experience.

To implement the contextual search engine, we need: (1) Find user interest and search intent, and then express as a special data structure called "User Context". (2) Define "Web Context", implement web pages contextual analysis. (3) Build the Contextual Inverted Index based on "Web Context". (4) Search information in the Contextual Inverted Index based on "User Context" and search words, the user get the web pages those are more in line with retrieval intention. (5) Contextual Sort to determine the order of the results, according to the result of the contextual match. Due to the length limitations, the article only concern the definition of the web context and web contextual analysis those we propose. It is an innovation of this paper.

In the area of web content extraction, after more than ten years development, many effective ways have been proposed by our predecessors. CAI Deng uses visual features to extract the information from the web page and VIPS algorithm have been proposed by him firstly. The algorithm is based on web visual content structure information and it also combines with the DOM tree [2, 3]. LI Xiao-Dong proposes an algorithm that extracts the path information of information we need in the DOM hierarchy as coordinates of information extraction, designs a new inductive learning algorithm to semi-automatically generate rules and then it generates JAVA classes based on these rules [4-7].

In the area of word sense disambiguation, there are also a number of methods: Guided machine learning methods have been applied to word sense disambiguation, the algorithm give a guided machine learning WSD method based on the vector space model [8-10]. Because the guided machine learning method requires a large number of human labeling, for this problem, LU Song proposes a method which avoids the manual word sense tagging without guided learning [11]. WU Yun-Fang proposes an algorithm that combines with the integrated approach, three kinds of integration methods have been applied in the word sense disambiguation application firstly. Multiplication, the mean and maximum values have good performance in the process of disambiguation [12].

Of these methods in depth study and practice of comparison, we fully combine for the context mining.

This paper presents a complete theoretical framework of web context analysis, it can implement the context analysis on the basis of the traditional web parsing. In order to support context-based retrieval, we must determine the right meaning of the polysemy words in the text document, web context analysis is right based on the document context information to determine the right meaning of the polysemy words. But, it does not only include the word sense disambiguation, it also includes the extraction of other context elements, such as time and locations.

Section 2 provides an overview of the formal representation of the web context and the overall framework of automated web context analysis. Subsequently, describe the extraction of web content, Chinese word segment and POS tagging, a word sense disambiguation method based on the common ontology. Finally, we give a reasonable organization of web context and describe the following work briefly.

2   The Overall Framework of Web Context Analysis and the Formal Representation of Web Context

2.1   The Overall Framework of Web Context Analysis

For all levels of processing object and the problem characteristics in web context analysis, we will fully combine the techniques which are used in various parts in the processing, unified establish the model. The model contains Web page context important functional modules like extraction, Chinese word segmentation and POS tagging, word sense disambiguation of polysemy substantives and gain of Page last updated time .Shown in Figure 1, Among them, Web content extraction using rule-based approach to extract the web page body text and images in it, and removed pages noise (like advertising links and copyright statement) before extracting the body of web pages. Access to last update time use real-time search technology, this mainly bases on that time of web pages updating is in line with the Poisson distribution.

Word sense disambiguation is the most important step in the web context analysis, mainly for determining the right meaning of the polysemy words and the following implementation of the web contextual index system. Here, we mainly use word sense disambiguation technology based on generic ontology.
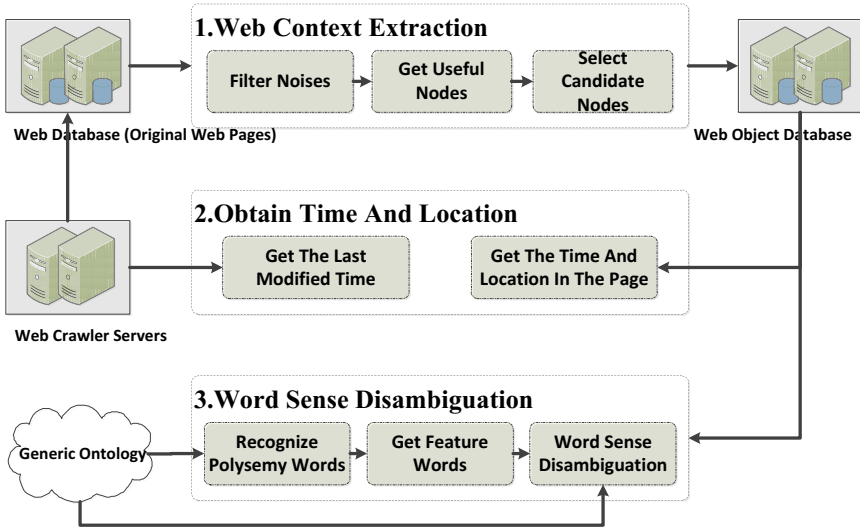
Fig. 1. The Overall Framework Of Web Context Analysis

2.2   The Formal Representation of Web Context

Known by Part Ⅰ Introduction, web context analysis is the precondition and founda-tion of the implementation of the whole web contextual search engine, and follow-up contextual inverted index will be build based on the results of web context analysis system, therefore, the web context analysis directly impacts on the quality of contex-tual indexes, while the contextual index is the core data organization of entire contex-tual search engine system. So, a good context representation is critical.

Web context mainly includes the following elements: a set of pairs between words and context words in the main content of the web page, the location information in the web page, the time information in the web pages, the creation time of the web page and last modified time. This paper presents a formal description of a web context, as follows:

PC=<{<W,C>},PG,PT>

Where W is word or term, C is context word, PG is the location information in the web page, PT is the time information (includes creation time, last modified time, time information in the web page, etc.).

3   A Rule-Based Web Context Extraction Algorithm

3.1   The Overview of Web Content Extraction Method

In this paper, we mainly use rule-based approach to extract web context. We build rules according to the location and characteristics information of the text fragments in web pages, filter, merge, and ultimately get the effective and useful context text. The algorithm we propose is more accurate than the algorithm that simply uses label rules to extract context.

The features of detail pages:

Have more text, and these words would normally not be a hyperlink.

Generally have clear text paragraphs and more punctuation.

Have less links. The main part of theme-oriented page is text. The Page has less links than navigation pages.

Visually, the body text locates on the central location of the web page, rather than on the edge.

Firstly, we construct the DOM tree to remove all noise. After filtering out the noises of page, we extract and splice the effective content. Firstly, use web analytical tool to generate the page's DOM tree; Secondly, filter out all nodes of the DOM tree and record them; Thirdly, merge the continuous text in the nodes those have been recorded, and make it to become the largest content ratio node; Finally, select the largest content ratio node as candidate node by comparing the content ratio and node ratio of useful nodes, and then return. Among them, the content ratio and node ratio is defined as follows:

Definition 1: Content Ratio

$$contentRatio = words\,/\,total\,words$$

Content Ratio (Word Ratio) refers to the proportion of the full text of one node and the total length of the useful text of extracted candidate node. It reflects the degree of importance of the node, the higher the degree of importance of the node, the more important the context text of all text nodes under the node, and that is, the greater the likelihood of the final effective content.

Definition 2: Node Ratio

$$wordsDensity = word\,/\,node$$

Node Ratio (Word Density) refers to the average number of text for all child nodes of the node. It further departures from the node's point of view, quantitative describes the importance of content text of all child nodes of parent node. The greater the word density, the greater the likelihood of that these content texts become into final effective texts.

4   Web Content Extraction Algorithm

Algorithm 1. Web Content Extraction
1: INPUT：strHtml
2: OUTPUT：strTextContent
3: mTree ← parser.Load(strHtml)
4: for all child node N in mTree.DocumentNodedo
5:      filterNode(N)
6: end for
7: nodeInfList ← GetRemainNodes(mTree)
8: nodeList ← MergeNodes(nodeInfList)
9: for all node N in nodeListdo
10:      L.crList ← ComputeContentRatio(N)
11:      L.wdList ← ComputeWordsDensity(N)
12: end for
13: candiateNode ← GetBodyNode(L,nodeList)
14: strTextContent ← prettyPrint(candiateNode)

Description: In Algorithm 1, mTree represents the DOM tree structure corresponding to the HTML source, filterNode filter every text node to judge whether the input nodes meet the requirement of link-text ratio. As another saying, it judge that the text of the node is or not the noise (such as advertisements etc.). If it is the noise, then we remove the node from DOM tree structure. GetRemainNodes is used for getting the set of nodes those have been filtered. PrettyPrint is used for printing all content text of one node. Among them, the algorithm that tests whether a node should be removed is as follows:

Algorithm 2. A Node Should Be Removed ?

1: INPUT：N : HtmlNode
2: OUTPUT：None
3: ifN.HasNoChildNodesthen
4:     return
5: end if
6: links← GetNumLinks(N)
7: word ←GetWordsAndMaxSegLen(N, maxSegLen)
8: ration ← 0
9: if words == 0 then
10:     ration ← 0.25 + 1
11: else
12: ration←links/words
13: end if
14: if ration > 0.25 and maxSegLen< 100 then
15:     for all child node n in N do
16:         removeAll(n)
17:     end for
18: end if

---

5   A Word Sense Disambiguation Algorithm Based on Generic Ontology

There are roughly two kinds of word sense disambiguation algorithms, the one process text and execute the word sense disambiguation by calculate the semantic similarity of Chinese words based on generic ontology (such as HowNet, WordNet, etc.), the other one execute the word sense disambiguation based on the statistic vector space in the context. Both of them have advantages and disadvantages, we execute the word sense disambiguation based on the generic ontology in this paper, and the generic ontology is HowNet.

5.1   Compute the Similarity of Word Concept Based on HowNet System

The task of word sense disambiguation in the web contextual search system proposed by us is to determine the only interpretation of polysemy substantives in the context.

Because the feature words of polysemy substantives we extract are also substantives, the computing of word semantic similarity here is refer to the similarity of substantives.

Compute The Similarity of Sememe

$$Sim(p_1, p_2) = \frac{\alpha}{d + \alpha}$$

Among them, p_1 and p_2 represent two sememe. d is the path length of p_1 and p_2 in the sememe hierarchy structure, it is an unsigned integer. α is an adjustable parameter.

Compute The Similarity Of Concept Of Substantives

One substantive can be described as a characteristic structure in the HowNet system. We mark the different similarity as $Sim_j(S_1, S_2)$.

$$Sim(S_1, S_2) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} Sim_j(S_1, S_2)$$

$\beta_i(1 \le i \le 4)$ is an adjustable parameter.

$\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1, \beta_1 \ge \beta_2 \ge \beta_3 \ge \beta_4$, it reflects that the role $Sim_1(S_1, S_2)$ to $Sim_4(S_1, S_2)$ play to the overall similarity is in descending order.

The four similarity calculation eventually can be attributed to the sememe similarity calculation. The similarity of sememe calculation method can be found in the above section.

Until now, we have obtained the similarity of any two substantives. The following describes the word sense disambiguation algorithm based on similarity.

5.2   A WSD Algorithm Based on Generic Ontology

A polysemy substantive in different context may have different meanings, and the main task of web context analysis is to determine the right meaning of the polysemy substantive in special context environment. Here, we execute Chinese word segment and POS tagging on the original text (the main content of a web page). The Chinese word segment and POS tagging of ICTCLAS have been made in a very high precision and accuracy in this step. We assume that the POS tagging is correct. At this time, if a polysemy substantive have only one meaning in special context with the determined POS, and then we need not execute the word sense disambiguation, and we can get the right meaning according to the POS tagging. Therefore, the polysemy substantives this paper studies refers to words those exist multiple meanings for their particular POS.

The main idea of word sense disambiguation in this paper: Take HowNet as the knowledge source. Firstly, we calculate the correlation between the polysemy substantive and its context words, find the words related with the polysemy substantive, and record them as the feature words. Secondly, we calculate the similarity of the feature words of polysemy substantives and the instance words in HowNet. The

meaning of the instance words which have the greatest similarity value is the right meaning of the polysemy substantive in the particular context environment. Specific processes are shown in Figure 2.
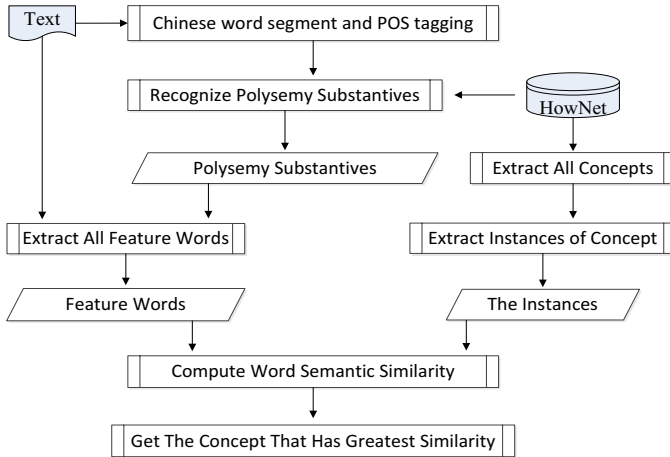
Fig. 2. The Flow Diagram Of WSD

Algorithm 3. Word Sense Disambiguation

1: INPUT：text : string, word : string
2: OUTPUT：meaning
3: FeatureWordList ← GetFeatureWords(text,word)
4: InstanceWordList ← GetInstanceWords(word)
5: Count ← GetMeaningCount(word)
6:meaningList ← GetMeaning(word)
7: MaxSimilarity ← 0, MaxIndex ← 0
8: for (i=0;i<Count;++i) do
9:   similarity ← 0
10:   for all word w1 in FeatureWordListdo
11:     for all word w2 in InstanceWordList[i] do
12:        tempSimilarity ← ComputeSimilarity(w1,w2)
13:        if tempSimilarity> similarity then
14:     similarity ← tempSimilarity
15:       end if
16:     end for
17:    end if
18:    if similarity >MaxSimilaritythen
19:       MaxSimilarity ← 0
20: MaxIndex ← i
21:   end if
22: end for
23: meaning ← meaningList.at(MaxIndex)

Recognize The Polysemy Substantive

The text after text preprocessing (word segment and POS tagging) can basically eliminate the two following ambiguity: the ambiguity in the period of word segment and the ambiguity due to the different POS. If one word has two or more meanings, you can never eliminate ambiguity in the period of preprocessing. So we search the meanings of the word based on its POS in the HowNet knowledge base, if the number of the meanings is two or more, then we can mark the word as the ambiguous word, otherwise we think it is not an ambiguous word.

Get The Feature Words

The so-called phenomenon of ambiguity is that one word has different meanings in different contexts. The polysemy word is called the ambiguous word. Words in the context can play a constraint role to the ambiguous word, but not all of the words of the context can. The words that can restrict the meaning of the ambiguous word are called the feature words. So we need to filter the content and extract feature words from the context.

6   Experiments and Discussion

Through the aforementioned theoretical analysis, we already have the necessary theoretical foundation of the web context analysis and technical support. To verify the effectiveness of the proposed method, the ideal method is to put this system onto real web contextual search engine system and test it, but we do not have the experimental environment. Therefore, we only show the corresponding experimental results and compared values as follows.

6.1   The Experiment Result of Web Content Extraction

The raw data in the web content extraction experiment is from the portal news pages, school news pages, government official pages, personal blogs, forums and other unallocated category pages. Through experimental comparison, the proposed method in this paper can solve the problem of web content extraction effectively, and the accuracy rate has reached 99.99%.

Table 1. Experimental Result

	Total Count	Pages (Right)	Accuracy Rate
News	10000000	9999775	99.99%
School	10000000	9999943	99.99%
Government	10000000	9991432	99.91%
Blogs	10000000	9992315	99.92%
BBS	10000000	9992643	99.93%
Others	20000000	17346056	86.73%

We can see that the web context extraction algorithm that we propose achieves very high accuracy rate for all the pages those contain the main context. For the other pages those do not include the regular text paragraph, the accuracy rate of the algorithm is not very high, but we can also accept the result. These pages do not have a higher influence and constraint for the establishment of follow-up contextual indexes and the implement of the web contextual search system. Web context analysis makes sense only for the pages those contain large amounts of text. Therefore, the experiments show that this algorithm can achieve better execution results.

6.2   The Experiment Result of WSD Algorithm

To verify the effectiveness of the word sense disambiguation algorithm, we show the experimental results in TABLE 2.

It can be seen from Table 2 that the word sense disambiguation strategy used in this paper is able to achieve higher accuracy rate for all kinds of text in test set. This provides a strong guarantee for the accuracy of web context analysis.

Table 2. Experimental Result

	Ploysemy Words Count	Words(Right)	Accuracy Rate
News	1329877	1232845	92.70%
Novels	2123439	2113432	99.53%
Blogs	1432123	1421032	99.23%
Notices	834876	834533	99.96%
Others	1200138	1187131	98.92%

6.3   The Experimental Result of Web Context Analysis

The paper proposed the concepts of web context and web context analysis for the first time. We do lots of experiments, because no one does the same work, we only show our results here. We are sure we achieve a good result.

Here, we randomly select two million web pages for testing and have achieved very good results. The mean accuracy rate can achieve 95.4%. So, the results we achieve can lay a good basis for ultimately realizing contextual indexes and web contexual search engine.

7   Conclusion and the Future Work

Web context analysis is the precondition of foundation of the implement of contextual index and the whole web contextual search engine system. We study the method of web context analysis based on generic ontology in this paper and implement a prototype system to verify the effectiveness of the method we propose. In the process of web context analysis, we mainly execute Chinese word segmentation algorithm for the content of the web page, execute word sense disambiguation algorithm for the

polysemy substantives, and determine the right meaning of the polysemy substantive in the specific context. In addition, the task of web context analysis includes getting the time and location information in the web page, and they prepare for the establishment of the context indexes.

In the process of web context analysis, we extract the main content of the web page firstly. Secondly, we execute Chinese word segmentation algorithm and POS tagging algorithm for the content when we get the accurate content of the web page. Finally, we recognize the polysemy substantives, execute the word sense disambiguation algorithm based on generic ontology, and determine the right meaning of the polysemy substantive in the specific text paragraph. In the period of word sense disambiguation, we use the computation of the semantic similarity of words based on HowNet knowledge base. In the proceeding of extracting the feature words, we use the relevancy to filter the candidate feature words, make them more accurate, and ensure that these feature words have more efficient influence and binding for the ambiguous words. The experiments show that our method achieves better results.

In the following work, we will further improve the functionality of web context analysis and the accuracy of analytical results based on the contents of this paper. We will further research and develop effective organization and representation of web context elements and make them more propitious to the establishment of the web contextual index. The contextual index is the content that we need further research, and we will take the contextual index as the starting point to build the entire web contextual search engine. The content we research in this paper is the starting point of the entire project. There are many tasks to do in future and they are not easy. We will do the research of web search engine in future, especially the web contextual search engine system.

References

1. Kraft, R., Chang, C.C., Maghoul, F., Kumar, R.: Searching with context. In: Proceedings of the 15th International Conference on World Wide Web, pp. 477–486. ACM (2006)
2. Cai, D., Yu, S., Wen, J.R., Ma, W.Y.: VIPS: a visionbased page segmentation algorithm. Microsoft Technical Report, MSR-TR-2003-79 (2003)
3. Shaohua, Y., Hailue, L., Yanbo, H.: Automatic data extraction from template-generated Web pages. Journal of Software 19, 209–223 (2008)
4. Xiaodong, L., Yuqing, G.: DOM-based information extraction for the web sources. Chinese Journal of Computers 25, 526–533 (2002)

5. Gupta, S., Kaiser, G., Neistadt, D., Grimm, P.: DOM-based content extraction of HTML documents. In: Proceedings of the 12th International Conference on World Wide Web, pp. 207–214. ACM (2003)
6. Wang, L., Liu, Z.-T., Wang, Y.-H., Liao, T.: Web Page Main Text Extraction Based on Content Similarity. Computer Engineering 36(6), 102–104 (2010)
7. Qi, W., Tang, S.W., Yang, D.Q., Wang, T.J.: DOM-Based Automatic Extraction of Topical Information from Web Pages. Journal of Computer Research and Development, 10 (2004)
8. Han, Z., Li, W., Mo, Q.: Research on methods for extracting text information from HTML pages. Application Research of Computers 12,012 (2008)
9. Zhou, J., Zhu, Z., Cao, X.: Research on Content Extraction from Chinese Web Page Based on Statistic and Content-Feature. Journal of Chinese Information Processing 23(5), 80–85 (2009)
10. Lu, S., Bai, S., Huang, X., Zhang, J.: Supervised word sense disambiguation based on Vector Space Model. Journal of Computer Research & Development 38, 662–667 (2001)
11. Lu, S., Bai, S., Huang, X.: An Unsupervised Approach to Word Sense Disambiguation Based on Sense-Words in Vector Space Model. Journal of Software 13(6), 1082–1089 (2002)
12. Wu, Y., Wang, M., Jin, P., Yu, S.: Ensembles of Classifiers for Chinese Word Sense Disambiguation. Journal of Computer Research and Development 45(8), 1354–1361 (2008)

Redundant Nodes Elimination
in Wireless Sensor Networks

Shouzhi Huang[1] and Xuezeng Zhao[2]

[1] Research Institute of Petroleum Exploration & Development, China National Petroleum Corporation, 100083 Beijing, China
[2] School of Mechanical and Electrical Engineering, Harbin Institute of Technology, 150001 Harbin, China
huangszhit@163.com, zhaoxz@hit.edu.cn

Abstract. In the real application situations, the wireless sensor networks (WSNs) usually consist of a large number of sensor nodes with limited energy to monitor the environment. The energy efficiency is one of the most important challenges for WSNs in order to extend the network lifetime. There are several widely-used methods to achieve this goal, such as data fusion at coordination sensor nodes, and turning off redundant nodes, etc. The existing methods of redundant nodes elimination are analyzed, and their disadvantages are discussed. Then a new method of redundant node elimination based on the coverage probability is proposed to maximize the network lifetime for the heterogeneous WSNs. Combining with EECRS protocol, an energy efficient protocol based on redundant nodes elimination is proposed. Using random sensors deployment with different sensing radii, it is shown that the sensing coverage quality after using proposed redundant nodes determination method can be guaranteed to 99.9% with more than 8 neighbor nodes for each sensor node, and proposed method has less on-duty nodes compared with other methods. And that can reduce the energy consumption of network. Compared with perimeter coverage protocol and CCP protocol, proposed protocol can extend the network lifetime.

Keywords: wireless sensor networks, redundant nodes, coverage probability.

1 Introduction

Wireless sensor networks (WSNs) are constituted of a large number of sensor nodes which consist of several low-power sensors, a microcontroller, a radio transceiver, and an energy source [1-3]. WSNs have many features such as low power consumption, low data rate transmission, and high flexibility. Nowadays WSNs are widely applied to many areas including industry, environmental protection, and military, etc., to monitor the environmental information, for instance, temperature, chemicals, electrical signals, and target tracking.

The most important challenges for WSNs are the coverage and the network lifetime. Currently many researchers focus on the methods that can extend the

network lifetime because the sensor nodes have a limited battery life. Common methods to reduce the energy consumption for the networks are data coding at the coordinators [4,5], turning off the redundant sensor nodes [6-9], and balancing the energy consumption for the entire network through suitable routing protocols [10,11].

In many situations, such as harsh field, disaster areas and animal activity areas, the sensors are always deployed randomly. In these cases, many sensor nodes need to be deployed in the monitoring areas to guarantee full coverage of the areas [12]. This will result in lots of redundant nodes. To identify and turn off the redundant nodes can save energy consumption for the entire network to extend the network lifetime.

In this paper, we first analyze the existing methods. We then propose a method that can eliminate the redundant nodes based on coverage probability. In the proposed approach, it is assumed that the sensor nodes based on Boolean model for the sensing range and the different sensing radii. Then an energy efficient protocol based on redundant nodes elimination is proposed combined with EECRS protocol [13]. With our protocol, the network lifetime is indeed extended compared with CCP protocol and perimeter coverage protocol by simulation.

2   Review of Other Works

Redundancy and coverage problem are considered in many literatures before. Xing studied the relationship between coverage and connectivity, and proposed a coverage maintenance scheme, and a coverage configuration protocol (CCP) [6]. In this protocol sensors consult an eligible rule that each sensor finds all the intersection points among the borders of its neighbors sensing radii and determines the eligibility of turning it off if each of those intersection points is covered with the desired sensing degree. However, void area may be occurred when the node, whose sensing radius has intersection points with the border of monitoring area, is turned off as the eligibility to sleep determined by CCP. Computing complexity of CCP is high, $O(n^3)$.

In reference [7], Huang proposed a new conception of perimeter coverage, which switched the whole sensing area coverage to the perimeter coverage for the sensing range circle. This algorithm considers the sensor nodes with different sensing radii and can be extended to 3-D network environment. However, the perimeter of sensing circle coverage along may also cause the coverage void area within the sensing range.

By combining the computational geometry and the graph theoretical methods, specifically the Voronoi diagram and graph search algorithms, Megerian at al established an optimal polynomial time worst and average case algorithm for coverage calculation for homogeneous isotropic sensors in reference [8]. However, this method needs a lot computing at the beginning that may cause large time delay and can only be used in homogeneous network.

In reference [9], Tian and Georganas proposed a node-scheduling scheme, which could reduce system overall energy consumption, therefore increasing system lifetime, by turning off some redundant nodes. The coverage-based off-duty eligibility rule and back off-based node-scheduling scheme guarantees that the original sensing coverage is maintained after turning off redundant nodes. It only considers the neighbors which have a distance equal or less than nodes' sensing radius.

3   Energy Efficient Protocol Based on Redundant Nodes Elimination

To solve the redundancy problem in heterogeneous network, we propose redundant nodes determination method based on coverage probability. And combining with EECRS protocol, an energy efficient protocol based on redundant nodes elimination is proposed.

3.1   Redundant Nodes Determination Model Based on Coverage Probability

We assume that the sensor nodes have a Boolean sensing model, as shown in Equation (1), and different sensing ranges R_s ($R_s/R_{s0} \in [1,2]$, R_{s0} is a setting value). We also assume the communication range of sensor node $R_C \geq R_S$.

$$f_x = \begin{cases} 1, & \text{when } x \text{ is within the sensing area of sensor nodes} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where x is a point in the monitoring area, f_x is the sensing probability.

Definition: Neighbor. The neighbor set of node i is defined as

$$N(i) = \{n \in j \, | \, d_{ij} \leq R_{si} + R_{sj}, j \neq i\} \tag{2}$$

where j is node set in the monitoring region, d_{ij} denotes the Euclidean distance between node i and node j.

Theorem: A sensor node is redundant if its sensing area is fully covered by its neighbors.

Assume that the sensing range of sensor node i denotes as S_i. As shown in Fig.1, the shadow part S_{ij} is the intersection of node i and its neighbor node j. P_1, P_2 are the intersection points. We can see that

$$R_{si} \sin \alpha = R_{sj} \sin \beta \tag{3}$$

$$R_{si} \cos \alpha + R_{sj} \cos \beta = d_{ij} \tag{4}$$

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{5}$$

where (x_i, y_i), (x_j, y_j) are the coordinates of node i, j respectively, R_{si} is the sensing range of sensor node i, R_{sj} is the sensing range of node j.

From Equation (3)-(5), we can get the intersection area S_{ij} as

$$S_{ij} = 2 \times [(\pi R_{si}^2 \times \frac{\alpha}{2\pi} - \frac{1}{2} R_{si}^2 \sin \alpha \cos \alpha) + (\pi R_{sj}^2 \times \frac{\beta}{2\pi} - \frac{1}{2} R_{sj}^2 \sin \beta \cos \beta)] \tag{6}$$

The probability that node i is covered by neighbor node j is determined by Equation (7) below.

$$p_{ij} = \frac{S_{ij}}{\pi R_{si}^2} \tag{7}$$
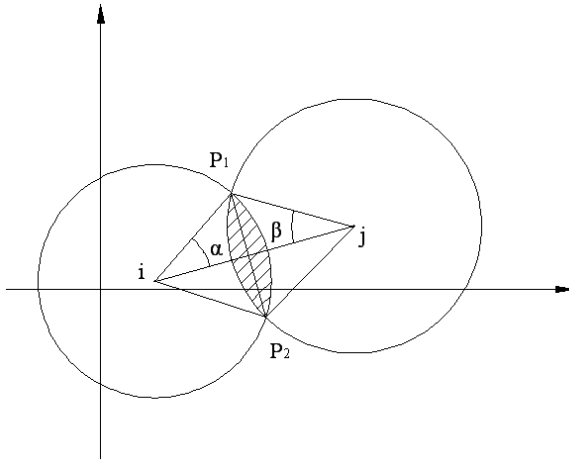
Fig. 1. Intersection area of node i and its neighbor node j
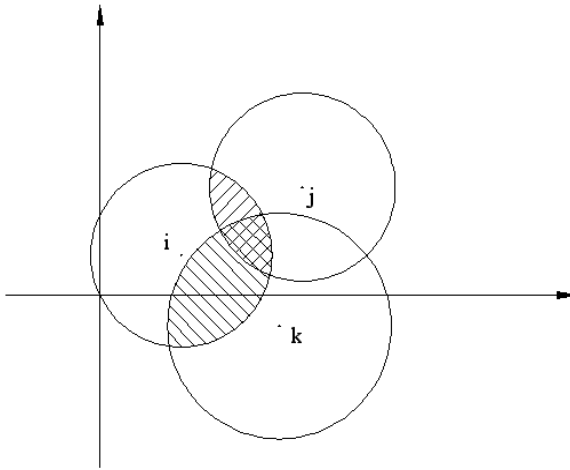
Fig. 2. Intersection area of node i and its neighbor node j and k

If the neighbor node j and k are both partially cover node i, and their coverage areas are overlapped as shown in Fig.2, the probability that node i is covered by neighbor node j and k is

$$p_{ijk} = \frac{S_{ij} + S_{ik} - S_{ijk}}{\pi R_{si}^2} \qquad (8)$$

With n neighbor nodes, the probability that node i is covered is

$$p_i = 1 - \prod_{j=1, k=2, j<k}^{n} (1 - p_{ijk})$$ (9)
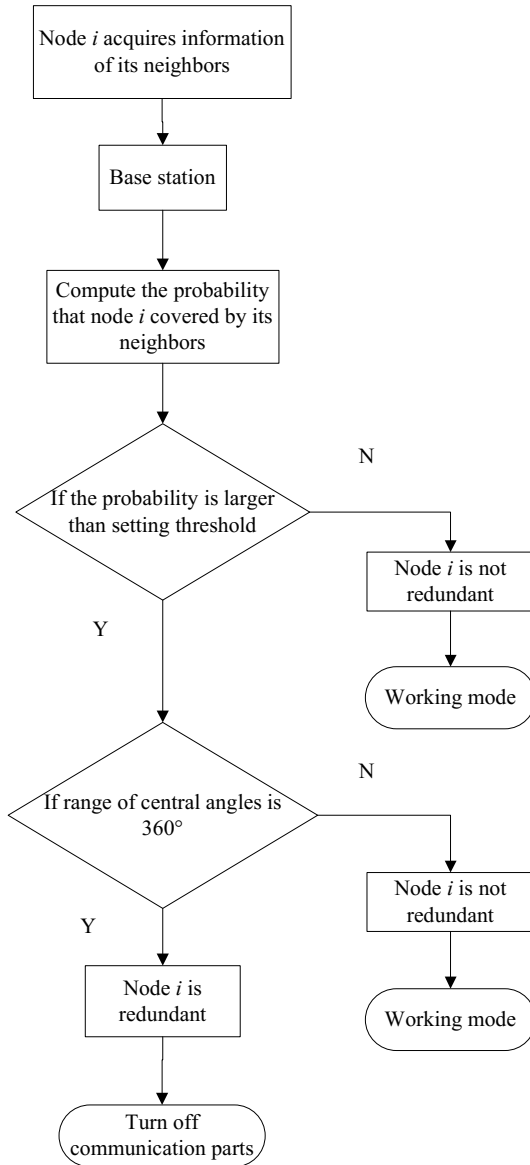
Fig. 3. Flow chart of redundant nodes determination algorithm based on coverage probability

In order to reduce the error, here we introduce central angle as the supplement of our method. Shown as Fig.2, the intersection angle between sensor i and sensor j is $\theta_{ij}=2\alpha$. Let $\theta_{j1}=-\alpha$, $\theta_{j2}=\alpha$, and angle θ_{ij} can be marked as $[\theta_{j1}, \theta_{j2}]$ which is called the range of intersection angle that node i covered by its neighbor node j.

The range of intersection angle that node i covered by all neighbor nodes need to be 360° to make sure the sensing circle of node i is covered by its neighbor nodes, shown as Equation (10).

$$\bigcup_{j=1}^{n}[\theta_{j1},\theta_{j2}]=[-\pi,\pi] \tag{10}$$

Flow chart of redundant nodes determination algorithm based on coverage probability is shown as Fig.3 below.

With the help of the coverage probability of sensor node that covered by its neighbors, we can determine if this node is redundant or not. In other words, the node is redundant if the coverage probability by its neighbors reaches the threshold we set.

Comparing with these existing methods, our proposed method has following advantages: (1) it can be used in heterogeneous network; (2) it has relatively low computational complexity, $O(n^2)$; (3) the determination of coverage redundancy has enough accuracy and the whole network coverage quality can be guaranteed after eliminating the redundant nodes.

3.2   Redundant Nodes Elimination Protocol

Combining with EECRS protocol, an energy efficient protocol based on redundant nodes elimination is proposed. This protocol contains four phases, redundant nodes determination phase, clustering phase, steady work phase and cluster head reselection phase.

In redundant nodes determination phase, sensor nodes collect neighbors' information to determine if they are redundant or not using the method proposed above. The nodes will turn into working mode if they are not redundant. When the nodes are determined as redundancy, they will turn into delay mode with random delay time to avoid coverage blind caused by turning off several nodes at same time. After random delay, the nodes will get into redundancy determination again. If they are still determined as redundancy, they will turn into sleep mode and turn off their communication parts to save energy. The flow chart of redundant nodes determination phase is shown as Fig.4.

After that, sensor nodes are divided into several clusters with method in EECRS protocol. Firstly, base station randomly selects several nodes as the temporary cluster heads, and broadcasts their ID in network. After confirming their situation, the temporary cluster heads will broadcast their ID and waiting for the joining requests. Other nodes select the nearest temporary cluster head according to the signal strengths and send joining requests. When temporary clusters are set up, the temporary cluster heads will send all information of cluster member nodes to base station. The nodes with highest energy in each cluster are selected as the cluster heads to be in charge of

data aggregation and transmission between clusters and base station, and the formal clusters are set up in the same way.

In steady work phase, cluster heads creates TDMA table according to information of member nodes in cluster and broadcast it within the cluster. The member nodes send data to cluster head at their time slots, and turn off their transceiver at other time to save energy. Data in cluster are aggregating at cluster head and transferred to base station in multi-hop via other cluster head.

When the energy values of cluster heads are lower than setting threshold or the cluster heads are determined as redundant nodes, base station will choose new cluster heads based on the energy values of nodes and energy consumption of data transmission within cluster. After new cluster head is selected, it will broadcast its information within the cluster and start new round of work.
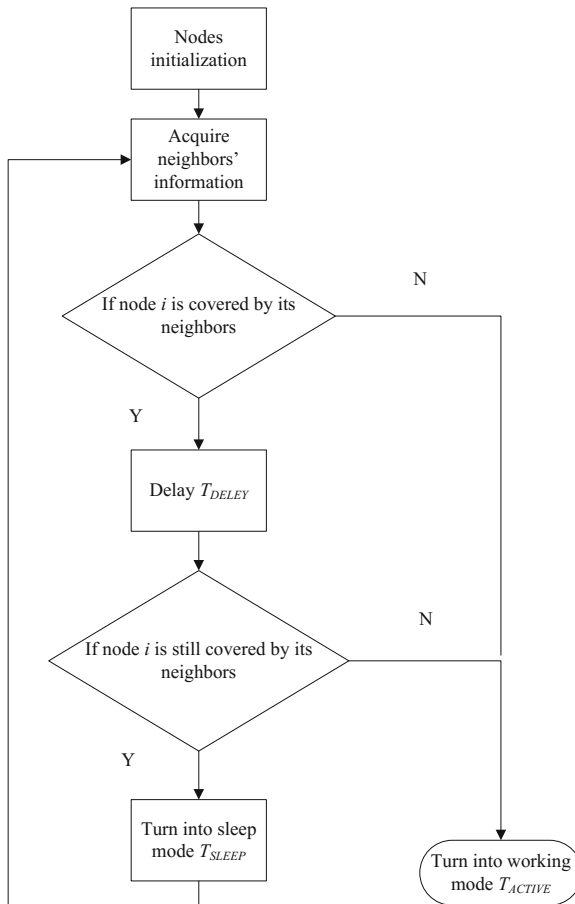
Fig. 4. Flow chart of redundant nodes determination phase

4   Simulation

To analyze the accuracy of the redundant nodes determination method, we simulate the sensor nodes redundancy by using our algorithm. We randomly deploy 100 nodes in a square space, 100m by 100m. Each node has a sensing range from 5m to 10m and a communication range of 10m. Each node knows its neighbors' identification and their locations. To calculate the sensing coverage, we divided the entire space into 1m by 1m units to convert the coverage probability into the number of these units being covered.

As shown in Fig.5, the difference between the degree of computed coverage by using our coverage probability method and real degree of coverage is less than 15% when the number of neighbors is less than 8. When the number of neighbors is more than 8, the degree of coverage is 99.9%. This can be considered that the sensor node is fully covered.
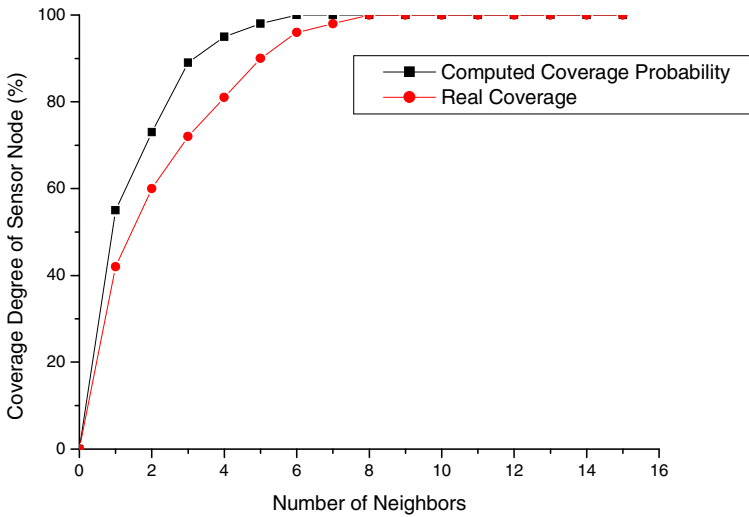
Fig. 5. Error between computed degree of coverage probability and real one

In the simulation, we randomly deploy 100-500 nodes in a square field, 100m by 100m. Both the sensing range of sensor nodes and the communication range are set at 10m. Each node knows neighbors' identifications and their locations. The numbers of on-duty nodes for these three methods, coverage probability, perimeter coverage and CCP, are shown in Fig.6.
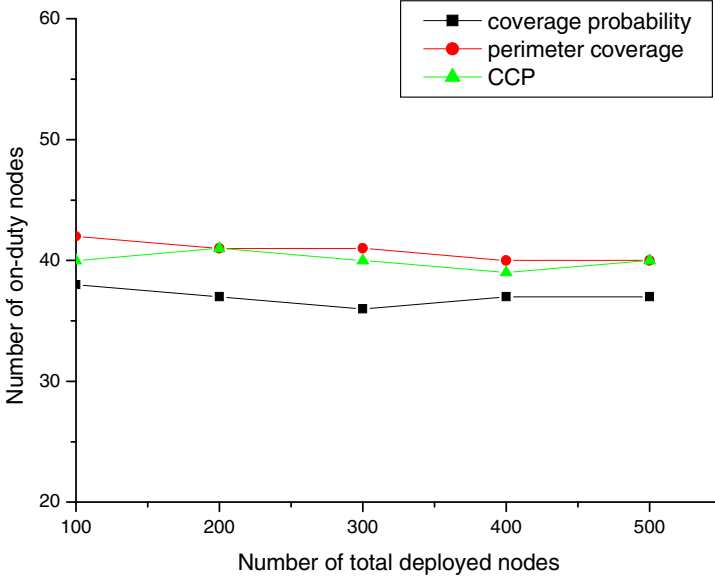
Fig. 6. Number of on-duty nodes of coverage probability, perimeter coverage and CCP with different total number of deployed nodes

Under guarantee of coverage degree, the less on-duty nodes in wireless sensor networks, the less energy that wireless sensor network consume. In Fig.6, it shown that the coverage probability method we proposed had less on-duty nodes than other two methods. So it is more energy efficiency than others.

To testify the energy efficiency of proposed protocol, we compare our protocol with CCP protocol and perimeter coverage protocol through simulation. The simulation environment parameters are as follows: 100 sensor nodes are randomly deployed in the area of 100m by 100m. The sensing radius of nodes sets within 5-10m. The transmission power of nodes is adjustable, and the maximum of that is 40m. Every node knows its position and their neighbors'. T_{ACTIVE}=60s, T_{DELAY}=5s, T_{SLEEP}=60s, the simulation lasts 1200s.

In the simulation, system checks every node's working condition every 10s. The node is considered dead when its energy value is too low to communicate with its neighbors. The numbers of sensor nodes alive of three protocols during simulation are shown in Fig.7.

As shown in Fig.7, the proposed protocol based on redundant nodes elimination has more nodes alive compared with other two protocols, and extends network lifetime. Our protocol extends network lifetime at 8.5% compared with perimeter coverage protocol, and extends 14.3% compared with CCP protocol.
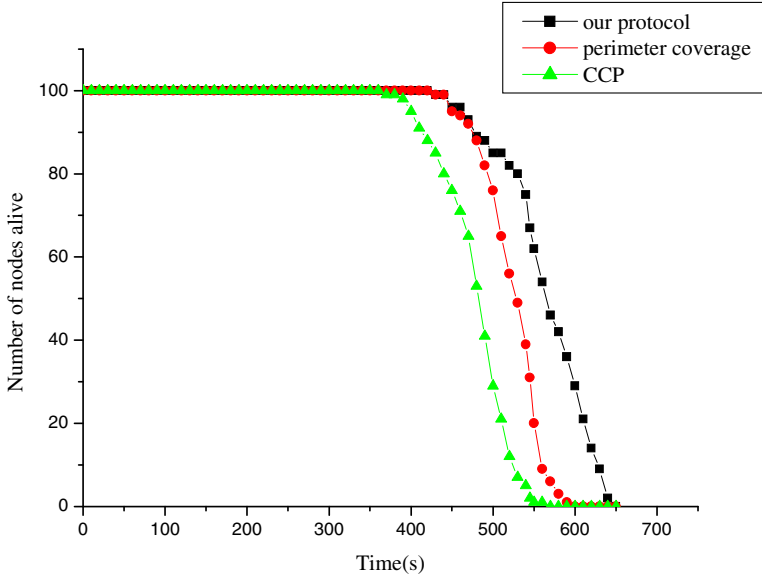
Fig. 7. The number of nodes alive of three protocols

5   Conclusion

In this paper, we have proposed a method of redundant nodes elimination based on coverage probability. In the proposed approach, the sensor nodes have different sensing radii. Combining with EECRS protocol, an energy efficient protocol based on redundant nodes elimination is proposed. With the help of simulation it is shown that the coverage quality of any sensor node is guaranteed to 99.9% when it has more than 8 neighbors. And the number of on-duty nodes is smaller in coverage probability method than other two that can reduce the energy consumption of the whole network. Compared with perimeter coverage protocol and CCP protocol, the proposed protocol is more energy efficient through simulations. In the future, we will keep focusing on how to balance the energy efficiency and coverage quality in WSNs.

References

1. Kulkarni, R.V., Forster, A., Venayagamoorthy, G.K.: Computational Intelligence in Wireless Sensor Networks: a Survey. IEEE Communications Survey & Tutorials 13(1), 68–96 (2011)
2. Zhu, C., Zheng, C.L., Shu, L., et al.: A Survey on Coverage and Connectivity Issues in Wireless Sensor Networks. Journal of Network and Computer Applications 35(2), 619–632 (2012)

3. Yuea, J., Zhang, W.M., Xiao, W.D., et al.: Energy Efficient and Balanced Cluster-Based Data Aggregation Algorithm for Wireless Sensor Networks. Procedia Engineering, 2009–2015 (2012)

4. Zungeru, A.M., Ang, L.M., Seng, K.P.: Classical and Swarm Intelligence Based Routing Protocols for Wireless Sensor Networks: a Survey and Comparison. Journal of Network and Computer Applications 35(5), 1508–1536 (2012)

5. Francesco, M.D., Das, S.K., Anastasi, G.: Data Collection in Wireless Sensor Networks with Mobile Elements: a Survey. ACM Transactions on Sensor Networks 8(1), 1–31 (2011)

6. Xing, G.L., Wang, X.R., Zhang, Y.F.: Integrated Coverage and Connectivity Configuration for Energy Conservation in Sensor Networks. ACM Trans. Sensor Networks 1(1), 36–72 (2005)

7. Huang, C.F., Tseng, Y.C.: The Coverage Problem in a Wireless Sensor Network. Mobile Networks and Applications 10(4), 519–528 (2005)

8. Megerian, S., Koushanfar, F., Potkonjak, M., et al.: Worst and Best-Case Coverage in Sensor Networks. IEEE Transaction on Mobile Computing 4(1), 84–92 (2004)

9. Tian, D., Georganas, N.D.: A Coverage-preserving Node Scheduling Scheme for Large Wireless Sensor Networks. In: WSNA 2002 Proceedings of the 1st ACM International Workshop on Wireless Sensor Networks and Applications, pp. 32–41. ACM Press, Atlanta (2002)

10. Heinzelman, W.B., Chandrakasan, A.P., Balakrishnan, H.: An Application-Specific Protocol Architecture for Wireless Microsensor Networks. IEEE Transactions on Wireless Communications 1(4), 660–670 (2002)

11. Younis, O., Fahmy, S.: HEED: a Hybrid, Energy-efficient, Distributed Clustering Approach for Ad Hoc Sensor Networks. IEEE Transactions on Mobile Computing 3(4), 366–379 (2004)

12. Vaidehi, V., Selvan, U.S., Jayendran, J., et al.: Redundant Node Deactivation by Scheduling in Wireless Sensor Networks. In: 2011 International Conference on Recent Trends in Information Technology (ICRTIT), pp. 613–617. Curran Associates, Chennai (2011)

13. Huang, S.Z., Zhao, X.Z.: An Energy-Efficient Cluster Head and Router Selecting Protocol for Wireless Sensor Networks. Applied Mechanics and Materials 226-228, 1807–1810 (2012)

Fast Algorithms for Solving RFMrLrR Circulant Linear Systems in Communication

Zhao-Lin Jiang and Ru-Li Chen

Department of Mathematics, Linyi University, Linyi 276005, P.R. China
jzh1208@sina.com, chenruli0428@163.com

Abstract. In this paper, fast algorithms for solving RFMrLrR circulant linear systems are presented by the fast algorithm for computing polynomials. The extended algorithms are used to solve the RLMrFrL circulant linear systems. Examples show the effectiveness of the algorithms.

Keywords: RFMrLrR circulant matrix, linear system, Fast algorithm.

1 Introduction

The $f(x)$-circulant matrix has a wide application, especially on the generalized cyclic codes [1, 2]. The properties and structures of the $x^n + rx - r$-circulant matrices, which are called RFMrLrR circulant matrices, are better than those of the general $f(x)$-circulant matrices, so there are good algorithms for solving the RFMrLrR circulant linear system. The advantage of the presented algorithm is that it can solve $AX = b$ whether the coefficient matrix is singular or nonsingular.

Definition 1. A row first-minus-rlast r-right(RFMrLrR) circulant matrix with the first row $(a_0, a_1, ..., a_{n-1})$ is meant a square matrix over the complex field $\mathbf{C}$ of the form $A = \text{RFM} \, r \, \text{LRcirc}_r \text{fr}(a_0, a_1, \ldots, a_{n-1})$

$$= \begin{pmatrix} a_0 & a_1 & \cdots & a_{n-2} & a_{n-1} \\ ra_{n-1} & a_0 - ra_{n-1} & \ddots & a_{n-3} & a_{n-2} \\ ra_{n-2} & ra_{n-1} - ra_{n-2} & \ddots & \ddots & a_{n-3} \\ \vdots & \cdots & \ddots & \ddots & \vdots \\ ra_1 & ra_2 - ra_1 & \cdots & ra_{n-1} - ra_{n-2} & a_0 - ra_{n-1} \end{pmatrix} \quad (1)$$

In this paper, let r be a complex number and satisfy $r \neq 0$ and $r \neq -\frac{n^n}{(n-1)^{n-1}}$. It is easily verified that the polynominal $g(x) = x^n + rx - r$ has no repeated roots in the complex field.

We define $\Theta_{(r,-r)}$ as the basic RFMrLrR circulant matrix over $\mathbf{C}$, that is,

$$\Theta_{(r,-r)} = \text{RFM } r \text{ LRcirc}_r \text{fr}(0,1,0,\ldots,0) \qquad (2)$$

It is easily verified that the polynomial $g(x) = x^n + rx - r$ is both the minimal polynomial and the characteristic polynomial of the matrix $\Theta_{(r,-r)}$, i.e., $\Theta_{(r,-r)}$ is nonsingular nonderogatory. In addition, $\Theta_{(r,-r)}^n = rI_n - r\Theta_{(r,-r)}$.

In view of the structure of the powers of the basic RFMrLrR circulant matric $\Theta_{(r,-r)}$ over $\mathbf{C}$. it is clear that

$$A = \text{RFM } r \text{ LRcirc}_r \text{fr}(a_0, a_1, \ldots, a_{n-1}) = \sum_{i=0}^{n-1} a_i \Theta^i{}_{(r,-r)} \qquad (3)$$

Thus, A is a RFMrLrR circulant matrix over C if and only if $A = f(\Theta_{(r,-r)})$ for some polynomial $f(x)$ over $\mathbf{C}$. The polynomial $f(x) = \sum_{i=0}^{n-1} a_i x^i$ will be called the representer of the RFMrLrR circulant matrix A over $\mathbf{C}$.

In addition, the product of two RFMrLrR circulant matrices is a RFMrLrR circulant matrix and the inverse of a nonsingular RFMrLrR circulant matrix is also a RFMrLrR circulant matrix. Furthermore, a RFMrLrR circulant matrices commute under multiplication [3-8].

Definition 2. A row last-minus-$r$first r-left (RLMrFrL) circulant matrix with the first row $(a_1, a_2, \ldots, a_{n-1})$ is meant a square matrix over the complex field $\mathbf{C}$ of the form

$$A = \text{RLM } r \text{ FLcirc}_r \text{fr}(a_1, a_2, \ldots, a_{n-1})$$

$$= \begin{pmatrix} a_0 & a_1 & \cdots & a_{n-2} & a_{n-1} \\ a_1 & a_2 & \ddots & a_{n-1} - ra_0 & ra_0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ a_{n-2} & a_{n-1} - ra_0 & \cdots & ra_{n-4} - ra_{n-3} & ra_{n-3} \\ a_{n-1} - ra_0 & ra_0 - ra_1 & \cdots & ra_{n-3} - ra_{n-2} & ra_{n-2} \end{pmatrix} \qquad (4)$$

Lemma 1. Let $A = \text{RFM } r \text{ LRcirc}_r \text{fr}(a_0, a_1, \ldots, a_{n-1})$ be a RFMrLrR circulant matrix over $\mathbf{C}$ and $B = \text{RLM } r \text{ FLcirc}_r \text{fr}(a_{n-1}, a_{n-2}, \ldots, a_0)$ be a RLMrFrL circulant matrix over $\mathbf{C}$. Then $B\hat{I}_n = A$ or $B = A\hat{I}_n$, where $\hat{I}_n$ is the backward identity matrix of the form

$$\hat{I}_n = \mathrm{LCirc}(0,\ldots,0,1) \tag{5}$$

Lemma 2 ([3]). Let $\mathbf{C}[x]$ be the polynomial ring over a field $\mathbf{C}$, and let $f(x), g(x) \in \mathbf{C}[x]$. Suppose that the polynomial matrix $\begin{pmatrix} f(x) & 1 & 0 \\ g(x) & 0 & 1 \end{pmatrix}$ is changed into the polynomial matrix $\begin{pmatrix} d(x) & u(x) & v(x) \\ 0 & s(x) & t(x) \end{pmatrix}$ by a series of elementary row operations, then $(f(x), g(x)) = d(x)$ and $f(x)u(x) + g(x)v(x) = d(x)$.

Lemma 3. Let $A = \mathrm{RFM}\,r\,\mathrm{LRcirc}_r\,\mathrm{fr}(a_0, a_1, \ldots, a_{n-1})$ be a RFM*r*L*r*R circulant matrix over $\mathbf{C}$. Then A is nonsingular if and only if $(f(x), g(x)) = 1$, where $f(x) = \sum_{i=0}^{n-1} a_i x^i$ is the representer *of* A and $g(x) = x^n + rx - r$.

In [4], for an $m \times n$ matrix A, any solution to the equation $AXA = A$ is called a generalized inverse of A.. In addition, X satisfies $X = XAX$ then A and X are said to be semi-inverses $A^{\{1,2\}}$. In this paper we consider only square matrices A. If A has index 1, the generalized inverse X of A is called the group inverse $A^{\#}$ of A. Clearly A and X are group inverse if and only if they are semi-inverses and $AX = XA$.

2   Fast Algorithms for Solving the RFM*r*L*r*R and the RLM*r*F*r*L Circulant Linear System

Consider the RFM*r*L*r*R circulant linear system

$$AX = b \tag{6}$$

where A is given in equation (1), $X = (x_1, x_2, \ldots, x_n)^T$, $b = (b_{n-1}, \ldots, b_1, b_0)^T$.

If A is nonsingular, then equation (6) has a unique solution $X = A^{-1}b$.

If A is singular and equation (6) has a solution, then the general solution of equation (6) is $X = A^{\{1,2\}}b + (I_n - A^{\{1,2\}}A)Y$, where Y is an arbitrary n-dimension column vector.

The key of the problem is how to find $A^{-1}b$, $A^{\{1,2\}}b$, $A^{\{1,2\}}A$, for this purpose, we first prove the following results.

Theorem   1.   Let $A = \text{RFM } r \text{ LRcirc}_r \text{fr}(a_0, a_1, \ldots, a_{n-1})$ be   a   nonsingular RFMrLrR circulant matrix of order n over $\mathbf{C}$ and $b = (b_{n-1}, \ldots, b_1, b_0)^T$. Then there exists a unique   RFMrLrR circulant matrix

$$C = \text{RFM } r \text{ LRcirc}_r \text{fr}(c_0, c_1, \ldots, c_{n-1})$$

of order order n over $\mathbf{C}$ such that the unique solution of Equation (6) is the last column of $\mathbf{C}$,   i.e.

$$X = (c_{n-1}, \ldots, c_1, c_0 - r c_{n-1})^T$$

Proof. Since matrix $A = \text{RFM } r \text{ LRcirc}_r \text{fr}(a_0, a_1, \ldots, a_{n-1})$ is nonsingular, then by Lemma 3, we have $(f(x), g(x)) = 1$, where $f(x) = \sum_{i=0}^{n-1} a_i x^i$ is the representer of A and $g(x) = x^n + rx - r$.

Let $B = \text{RFM } r \text{ LRcirc}_r \text{fr}(b_0 + r b_{n-1}, b_1, \ldots, b_{n-1})$ be the RFMrLrR circulant matrix of order n constructed by $b = (b_{n-1}, \cdots, b_1, b_0)^T$. Then the representer of B is $b(x) = (b_0 + r b_{n-1}) + \sum_{i=1}^{n-1} b_i x^i$. Therefore, we can change the polynomial matrix $\begin{pmatrix} f(x) & 1 & 0 & b(x) \\ g(x) & 0 & 1 & 0 \end{pmatrix}$ into the polynomial matrix $\begin{pmatrix} 1 & u(x) & v(x) & c(x) \\ 0 & s(x) & t(x) & c_1(x) \end{pmatrix}$ by a series of elementary row operations. By Lemma 2, we have

$$\begin{pmatrix} u(x) & v(x) \\ s(x) & t(x) \end{pmatrix} \begin{pmatrix} f(x) \\ g(x) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \ \begin{pmatrix} u(x) & v(x) \\ s(x) & t(x) \end{pmatrix} \begin{pmatrix} b(x) \\ 0 \end{pmatrix} = \begin{pmatrix} c(x) \\ c_1(x) \end{pmatrix}$$

i.e. $f(x)u(x) + g(x)v(x) = 1, \ u(x)b(x) = c(x)$.

Substituting x by $\Theta_{(r,-r)}$ in the above two equations respectively, we have

$$f(\Theta_{(r,-r)})u(\Theta_{(r,-r)}) + g(\Theta_{(r,-r)})v(\Theta_{(r,-r)}) = I_n, \ u(\Theta_{(r,-r)})b(\Theta_{(r,-r)}) = c(\Theta_{(r,-r)}).$$

Since $f(\Theta_{(r,-r)}) = A$, $g(\Theta_{(r,-r)}) = 0$ and $b(\Theta_{(r,-r)}) = B$, then

$$Au(\Theta_{(r,-r)}) = I_n \tag{7}$$

$$u(\Theta_{(r,-r)})B = c(\Theta_{(r,-r)}) \tag{8}$$

By equation (7) we know that $u(\Theta_{(r,-r)})$ is a unique inverse A^{-1} of A. According to equation (8) and the characters of the RFMrLrR circulant matrix, we know that the last column of $\mathbf{C}$ is $X = (c_{n-1}, \ldots, c_1, c_0 - r c_{n-1})^T = A^{-1}b$. Since $AA^{-1}b = b$, then $A^{-1}b = (c_{n-1}, \ldots, c_1, c_0 - r c_{n-1})^T$ is the solution of equation (6).

Since both A^{-1} and B are unique, then $A^{-1}B$ is also unique. So $X = (c_{n-1}, \ldots, c_1, c_0 - rc_{n-1})^T = A^{-1}b$ is unique.

Theorem 2. Let $A = \mathrm{RFM}\, r\, \mathrm{LRcirc}_r\mathrm{fr}(a_0, a_1, \ldots, a_{n-1})$ be a singular RFM*r*L*r*R circulant matrix of order n over $\mathbf{C}$ and $b = (b_{n-1}, \ldots, b_1, b_0)^T$. If the solution of equation (6) exists, then there exist a unique RFM*r*L*r*R circulant matrix $C = \mathrm{RFM}\, r\, \mathrm{LRcirc}_r\mathrm{fr}(c_0, c_1, \ldots, c_{n-1})$ and a unique RFM*r*L*r*R circulant matrix $E = \mathrm{RFM}\, r\, \mathrm{LRcirc}_r\mathrm{fr}(e_0, e_1, \ldots, e_{n-1})$ of order n over $\mathbf{C}$ such that

$X_1 = (c_{n-1}, \ldots, c_1, c_0 - rc_{n-1})^T$ is a unique special solution of equation (6) and $X_2 = X_1 + (I_n - E)Y$ is a general solution of equation (6), where Y is an arbitrary n-dimension column vector.

Proof. Since $A = \mathrm{RFM}\, r\, \mathrm{LRcirc}_r\mathrm{fr}(a_0, a_1, \ldots, a_{n-1})$, then the representer of A is

$$f(x) = \sum_{i=0}^{n-1} a_i x^i$$

and the characteristic polynomial of $\Theta_{(r,-r)}$ is $g(x) = x^n + rx - r$. We can change the polynomial matrix $\begin{pmatrix} f(x) \\ g(x) \end{pmatrix}$ into the polynomial matrix $\begin{pmatrix} d(x) \\ 0 \end{pmatrix}$ by a series of elementary row operations. Since A is singular, by Lemma 2 and Lemma 3, we know that $d(x)$ is a largest common factor, which is not equal to 1, of $f(x)$ and $g(x)$.

Let $f(x) = d(x)f_1(x)$ and $g(x) = d(x)g_1(x)$, then $(f_1(x),\ g_1(x)) = 1$. Since $(d(x),\ g_1(x)) = 1$, we have $(f(x),\ g_1(x)) = (d(x)f_1(x),\ g_1(x)) = 1$. Since $(d(x),\ g_1(x)) = 1$ and $(f(x),\ g_1(x)) = 1$, we have

$(f(x)d(x),\ g_1(x)) = 1$.

Let $B = \mathrm{RFM}\, r\, \mathrm{LRcirc}_r\mathrm{fr}(b_0 + rb_{n-1}, b_1, \ldots, b_{n-1})$ be the RFMrLrR circulant matrix of order n constructed by $b = (b_{n-1}, \ldots, b_1, b_0)^T$. Then the representer of B is $b(x) = (b_0 + rb_{n-1}) + \sum_{i=1}^{n-1} b_i x^i$. Therefore, we can change the polynomial

matrix $\begin{pmatrix} f(x)d(x) & \vdots & 1 & 0 & \vdots & d(x)b(x) & \vdots & d(x)f(x) \\ g_1(x) & \vdots & 0 & 1 & \vdots & 0 & \vdots & 0 \end{pmatrix}$ into the

polynomial matrix $\begin{pmatrix} 1 & \vdots & u(x) & v(x) & \vdots & c(x) & \vdots & e(x) \\ 0 & \vdots & s(x) & t(x) & \vdots & c_1(x) & \vdots & e_1(x) \end{pmatrix}$ by a series of

elementary row operations. Then, by Lemma 2, we have

$$f(x)d(x)u(x) + g_1(x)v(x) = 1 \tag{9}$$

$$d(x)u(x)b(x) = c(x) \tag{10}$$

$$d(x)u(x)f(x) = e(x) \tag{11}$$

$$f(x)d(x)u(x)f(x) + g(x)f_1(x)v(x) = f(x) \tag{12}$$

Substituting x by $\Theta_{(r,-r)}$ in equations (10), (11) and (12), respectively, we have
$d(\Theta_{(r,-r)})u(\Theta_{(r,-r)})b(\Theta_{(r,-r)}) = c(\Theta_{(r,-r)})$, $d(\Theta_{(r,-r)})u(\Theta_{(r,-r)})f(\Theta_{(r,-r)}) = e(\Theta_{(r,-r)})$,
$f(\Theta_{(r,-r)})d(\Theta_{(r,-r)})u(\Theta_{(r,-r)})f(\Theta_{(r,-r)}) + g(\Theta_{(r,-r)})f_1(\Theta_{(r,-r)})v(\Theta_{(r,-r)}) = f(\Theta_{(r,-r)})$.
Since $f(\Theta_{(r,-r)}) = A$, $g(\Theta_{(r,-r)}) = 0$ and $b(\Theta_{(r,-r)}) = B$ then

$$d(\Theta_{(r,-r)})u(\Theta_{(r,-r)})B = c(\Theta_{(r,-r)}) \tag{13}$$

$$d(\Theta_{(r,-r)})u(\Theta_{(r,-r)})A = e(\Theta_{(r,-r)}) \tag{14}$$

$$Ad(\Theta_{(r,-r)})u(\Theta_{(r,-r)})A = A \tag{15}$$

In the same way , we have

$$d(\Theta_{(r,-r)})u(\Theta_{(r,-r)})Ad(\Theta_{(r,-r)})u(\Theta_{(r,-r)}) = d(\Theta_{(r,-r)})u(\Theta_{(r,-r)}) \tag{16}$$

By equation (15) and (16), we know $T = d(\Theta_{(r,-r)})u(\Theta_{(r,-r)})$ is a semi-inverses
$A^{\{1,2\}}$ of A. Let $C = TB = c(\Theta_{(r,-r)}) = \text{RFM } r \text{ LRcirc}_r \text{fr}(c_0, c_1, ..., c_{n-1})$ and
let $E = TA = e(\Theta_{(r,-r)}) = \text{RFM } r \text{ LRcirc}_r \text{fr}(e_0, e_1, ..., e_{n-1})$. According to
equation (13) and the characters of the RFMrLrR circulant matrix, we know that the last
column of C is $(c_{n-1}, \cdots, c_1, c_0 - rc_{n-1})^T = Tb$. Since $AX = b$ has a solution,
then $ATb = b$, i.e., Tb is a solution of equation (6). We have
$A[Tb + (I_n - E)Y] = ATb + A(I_n - E)Y = b + AY - AEY = b + AY - ATAY = b + AY - AY = b$
So $X_2 = X_1 + (I_n - E)Y$ is a general solution of equation (6), where Y is an
arbitrary n-dimension column vector and $X_1 = Tb$.

Since both A and T are RFMrLrR circulant matrices, then $AT = TA$. If there
exists another RFMrLrR circulant matrix T such that $AT_1A = A$, $T_1AT_1 = T_1$,

$T_1A = AT_1$. Let $AT = TA = H$ and $AT_1 = T_1A = F$. Clearly $H^2 = H$ and $F^2 = F$ .Thus we have $AX = b$, , $F = T_1A = T_1ATA = FH$. So $H = F$.

Hence

$$T = TAT = HT = FT = T_1AT = T_1H = T_1F = T_1AT_1 = T_1.$$

So T is unique. Hence $TB = C$, $TA = E$ and Tb are also unique.

By Theorem 1 and Theorem 2, we have the following fast algorithm for solving the RFMrLrR circulant linear system (6) :

Algorithm 1:

Step 1. From the RFMrLrR circulant linear system (6), we get the polynomial

$$f(x) = \sum_{i=0}^{n-1} a_i x^i, \quad g(x) = x^n + rx - r \text{ and } b(x) = (b_0 + rb_{n-1}) + \sum_{i=1}^{n-1} b_i x^i;$$

Step 2. Change the polynomial matrix $\begin{pmatrix} f(x) & b(x) \\ g(x) & 0 \end{pmatrix}$ into the polynomial matrix

$\begin{pmatrix} d(x) & c(x) \\ 0 & c_1(x) \end{pmatrix}$ by a series of elementary row operations;

Step 3. If $d(x) = 1$, then the RFMrLrR circulant linear system (6) has a unique solution. Substituting x by $\Theta_{(r,-r)}$ in polynomial $c(x)$, we obtain the RFMrLrR circulant matrix $C = c(\Theta_{(r,-r)}) = \text{RFM } r \text{ LRcirc}_r \text{fr}(c_0, c_1, ..., c_{n-1})$. So the unique solution of $AX = b$ is $(c_{n-1}, \cdots, c_1, c_0 - rc_{n-1})^T$;

Step 4. If $d(x) \neq 1$, dividing $g(x)$ by $d(x)$, we get the quotient $g_1(x)$ and change the polynomial matrix

$$\begin{pmatrix} f(x)d(x) & d(x)b(x) & d(x)f(x) & d(x)f(x)b(x) \\ g_1(x) & 0 & 0 & 0 \end{pmatrix} \text{ into the polynomial}$$

matrix $\begin{pmatrix} 1 & c(x) & e(x) & r(x) \\ 0 & c_1(x) & e_1(x) & r_1(x) \end{pmatrix}$ by a series of elementary row operations;

Step 5. Substituting x by $\Theta_{(r,-r)}$ in polynomial $r(x)$, we get the RFMrLrR circulant matrix $R = r(\Theta_{(r,-r)}) = AA^{\{1,2\}}B$. If the last column of R isn't b, then $AX = b$ has no solution. Otherwise, the RFMrLrR circulant linear system $AX = b$ has a solution. Substituting x by $\Theta_{(r,-r)}$ in polynomial $c(x)$ and $e(x)$, we have the RFMrLrR circulant matrices $C = c(\Theta_{(r,-r)}) = A^{\{1,2\}}B$ and $E = e(\Theta_{(r,-r)}) = A^{\{1,2\}}A$. So the unique special solution of $AX = b$ is

$$X_1 = (c_{n-1}, \cdots, c_1, c_0 - rc_{n-1})^T \text{ and a general solution of } AX = b \text{ is}$$

$X_2 = X_1 + (I_n - E)Y$, where Y is an arbitrary n-dimension column vector.

By Lemma 1 and Theorem 1, we have the following theorem.

Theorem 3. Let $A = \mathrm{RLM}\, r\, \mathrm{FLcirc}_r \mathrm{fr}(a_{n-1}, a_{n-2}, ..., a_1, a_0)$ be a nonsingular RLMrFrL circulant matrix of order n over $\mathbf{C}$ and $b = (b_{n-1}, \cdots, b_1, b_0)^T$. Then there exists a unique RFMrLrR circulant matrix

$C = \mathrm{RFM}\, r\, \mathrm{LRcirc}_r \mathrm{fr}(c_0, c_1, ..., c_{n-1})$ of order n over $\mathbf{C}$ such that the unique solution of $AX = b$ is $X = (c_0 - rc_{n-1}, c_1, \cdots, c_{n-2}, c_{n-1})^T$.

By Lemma 1 and Theorem 2, we have the following theorem.

Theorem 4. Let $A = \mathrm{RLM}\, r\, \mathrm{FLcirc}_r \mathrm{fr}(a_{n-1}, a_{n-2}, ..., a_1, a_0)$ be a singular RLMrFrL circulant matrix of order n and $b = (b_{n-1}, \cdots, b_1, b_0)^T$. If the solution of $AX = b$ exists, then there exist a unique RFMrLrR circulant matrix $C = \mathrm{RFM}\, r\, \mathrm{LRcirc}_r \mathrm{fr}(c_0, c_1, ..., c_{n-1})$ and a unique RFMrLrR circulant matrix $E = \mathrm{RFM}\, r\, \mathrm{LRcirc}_r \mathrm{fr}(e_0, e_1, ..., e_{n-1})$ of order n such that

$X_1 = (c_0 - rc_{n-1}, c_1, \cdots, c_{n-2}, c_{n-1})^T$ is a unique special solution of $AX = b$ and $X_2 = X_1 + \hat{I}_n (I_n - E)Y$ is a general solution of $AX = b$, where Y is an arbitrary n-dimension column vector and $\hat{I}_n$ is given in equation (5).

By Theorem 3, Theorem 4 and Algorithm 1, we can get the fast algorithm for solving the RLMrFrL circulant linear system $AX = b$:

Algorithm 2:

Step 1. From the RLMrFrL circulant linear system, we get the polynomial

$$f(x) = \sum_{i=0}^{n-1} a_i x^i, \quad g(x) = x^n + rx - r \text{ and } b(x) = (b_0 + rb_{n-1}) + \sum_{i=1}^{n-1} b_i x^i;$$

Step 2. Change the polynomial matrix $\begin{pmatrix} f(x) & b(x) \\ g(x) & 0 \end{pmatrix}$ into the polynomial matrix

$\begin{pmatrix} d(x) & c(x) \\ 0 & c_1(x) \end{pmatrix}$ by a series of elementary row operations;

Step 3. If $d(x) = 1$, then the RLMrFrL circulant linear system has a unique solution. Substituting x by $\Theta_{(r,-r)}$ in polynomial $c(x)$, we obtain the RFMrLrR circulant matrix $C = c(\Theta_{(r,-r)}) = \mathrm{RFM}\, r\, \mathrm{LRcirc}_r \mathrm{fr}(c_0, c_1, ..., c_{n-1})$. So the unique solution of $AX = b$ is $(c_0 - rc_{n-1}, c_1, \cdots, c_{n-2}, c_{n-1})^T$;

Step 4. If $d(x) \neq 1$, dividing $g(x)$ by $d(x)$, we get the quotient $g_1(x)$ and change the polynomial matrix

$$\begin{pmatrix} f(x)d(x) & d(x)b(x) & d(x)f(x) & d(x)f(x)b(x) \\ g_1(x) & 0 & 0 & 0 \end{pmatrix}$$ into the polynomial

matrix $\begin{pmatrix} 1 & c(x) & e(x) & r(x) \\ 0 & c_1(x) & e_1(x) & r_1(x) \end{pmatrix}$ by a series of elementary row operations;

Step 5. Substituting x by $\Theta_{(r,-r)}$ in polynomial $r(x)$, we get the RFMrLrR circulant matrix $R = r(\Theta_{(r,-r)}) = AA^{\{1,2\}}B$. If the last column of R isn't b, then $AX = b$ has no solution. Otherwise, the RLMrFrL circulant linear system $AX = b$ has a solution. Substituting x by $\Theta_{(r,-r)}$ in polynomial $c(x)$ and $e(x)$, we have the RFMrLrR circulant matrices $C = c(\Theta_{(r,-r)}) = A^{\{1,2\}}B$ and $E = e(\Theta_{(r,-r)}) = A^{\{1,2\}}A$. So the unique special solution of $AX = b$ is $X_1 = (c_0 - rc_{n-1}, c_1, \cdots, c_{n-2}, c_{n-1})^T$ and a general solution of $AX = b$ is $X_2 = X_1 + \hat{I}_n(I_n - E)Y$, where Y is an arbitrary n-dimension column vector.

3   Examples

Examples 1. Solve the RFM2L2R circulant linear system
$AX = b,$

where $A = \mathrm{RFM}\,2\,\mathrm{LRcirc}_2\mathrm{fr}(1,1,3,1)$ and $b = (1,-1,0,2)^T$.

From $A = \mathrm{RFM}\,2\,\mathrm{LRcirc}_2\mathrm{fr}(1,1,3,1)$ and $b = (1,-1,0,2)^T$, we get the polynomial $f(x) = 1 + x + 3x^2 + x^3$, $g(x) = -2 + 2x + x^4$ and $b(x) = 4 - x^2 + x^3$. Then

$$A(x) = \begin{pmatrix} f(x) & b(x) \\ g(x) & 0 \end{pmatrix} = \begin{pmatrix} 1 + x + 3x^2 + x^3 & 4 - x^2 + x^3 \\ -2 + 2x + x^4 & 0 \end{pmatrix}.$$

We transform the polynomial matrix $A(x)$ by a series of elementary row operations as follows:

$$A(x) = \begin{pmatrix} 1 + x + 3x^2 + x^3 & 4 - x^2 + x^3 \\ -2 + 2x + x^4 & 0 \end{pmatrix}$$

$$\xrightarrow{(2)-x(1)} \begin{pmatrix} 1 + x + 3x^2 + x^3 & 4 - x^2 + x^3 \\ -2 + x - x^2 - 3x^3 & -4 + x^3 - x^4 \end{pmatrix}$$

$$\xrightarrow{(2)+3(1)} \begin{pmatrix} 1+x+3x^2+x^3 & 4-x^2+x^3 \\ 1+4x+8x^2 & 12-4x-3x^2+4x^3-x^4 \end{pmatrix}$$

$$\xrightarrow{8(1)-x(2)} \begin{pmatrix} 8+7x+20x^2 & 32-12x-4x^2+11x^3-4x^4+x^5 \\ 1+4x+8x^2 & 12-4x-3x^2+4x^3-x^4 \end{pmatrix}$$

$$\xrightarrow{2(1)-5(2)} \begin{pmatrix} 11-6x & 4-4x+7x^2+2x^3-3x^4+2x^5 \\ 1+4x+8x^2 & 12-4x-3x^2+4x^3-x^4 \end{pmatrix}$$

$$\xrightarrow{3(2)+4x(1)} \begin{pmatrix} 11-6x & 4-4x+7x^2+2x^3-3x^4+2x^5 \\ 3+56x & 36+4x-25x^2+40x^3+5x^4-12x^5+8x^6 \end{pmatrix}$$

$$\xrightarrow{28(1)+3(2)} \begin{pmatrix} 317 & 220-100x+121x^2+176x^3-69x^4+20x^5+24x^6 \\ 3+56x & 36+4x-25x^2+40x^3+5x^4-12x^5+8x^6 \end{pmatrix}$$

$$\xrightarrow{\frac{1}{317}(1)} \begin{pmatrix} 1 & \frac{220}{317}-\frac{100}{317}x+\frac{121}{317}x^2+\frac{176}{317}x^3-\frac{69}{317}x^4+\frac{20}{317}x^5+\frac{24}{317}x^6 \\ 3+56x & 36+4x-25x^2+40x^3+5x^4-12x^5+8x^6 \end{pmatrix}.$$

Since $d(x)=1$, then the RFM2L2R circulant linear system $AX=b$ has a unique solution. On the other hand,

$c(x) = \frac{220}{317} - \frac{100}{317}x + \frac{121}{317}x^2 + \frac{176}{317}x^3 - \frac{69}{317}x^4 + \frac{20}{317}x^5 + \frac{24}{317}x^6.$

Substituting x by $\Theta_{(2,-2)}$ in polynomial $c(x)$, we have the RFM2L2R circulant matrix $C = c(\Theta_{(2,-2)}) = RFM\,2LRcir_2\,fr(\frac{82}{317},\frac{78}{317},\frac{129}{317},\frac{128}{317})$. So the unique solution of $AX=b$ is the last column of C, i.e., $X = (\frac{128}{317},\frac{129}{317},\frac{78}{317},-\frac{174}{317})^T$.

Examples 2. Solve the RFM2L2R circulant linear system
$AX = b,$

where $A = RFM\,2LRcirc_2\,fr(3,\dfrac{-3+3\sqrt{3}}{2})$ and $b = (1,\sqrt{3}-1)^T$.

From $A = RFM\,2LRcirc_2\,fr(3,\dfrac{-3+3\sqrt{3}}{2})$ and $b = (1,\sqrt{3}-1)^T$, we get

the polynomial $f(x) = 3 + \dfrac{-3+3\sqrt{3}}{2}x,\quad g(x) = -2+2x+x^2$ and

$b(x) = 1+\sqrt{3}+x.$ Then

$$A(x) = \begin{pmatrix} f(x) & b(x) \\ g(x) & 0 \end{pmatrix} = \begin{pmatrix} 3 + \dfrac{-3+3\sqrt{3}}{2}x & 1+\sqrt{3}+x \\ -2+2x+x^2 & 0 \end{pmatrix}.$$

We transform the polynomial matrix $A(x)$ by a series of elementary row operations as follows:

$$A(x) = \begin{pmatrix} 3 + \dfrac{-3+3\sqrt{3}}{2}x & 1+\sqrt{3}+x \\ -2+2x+x^2 & 0 \end{pmatrix}$$

$$\xrightarrow{\overline{(3\sqrt{3}-3)(2)-2x(1)}} \begin{pmatrix} 3 + \dfrac{-3+3\sqrt{3}}{2}x & 1+\sqrt{3}+x \\ 6-6\sqrt{3}+(6\sqrt{3}-12)x & -2(1+\sqrt{3})x-2x^2 \end{pmatrix}$$

$$\xrightarrow{\overline{(2)-(2-2\sqrt{3})(1)}} \begin{pmatrix} 3 + \dfrac{-3+3\sqrt{3}}{2}x & 1+\sqrt{3}+x \\ 0 & 4-4x-2x^2 \end{pmatrix}$$

$$\xrightarrow{\frac{1+\sqrt{3}}{3}(1)} \begin{pmatrix} \sqrt{3}+1+x & \frac{4+2\sqrt{3}}{3}+\frac{1+\sqrt{3}}{3}x \\ 0 & 4-4x-2x^2 \end{pmatrix}.$$

It is obvious that $d(x)=1+\sqrt{3}+x \neq 1$, it denoted that A is singular. Then $g_1(x) = g(x)/d(x) = 1-\sqrt{3}+x$,

$$d(x)f(x) = 3(1+\sqrt{3})+6x+\frac{3\sqrt{3}-3}{2}x^2,$$

$$d(x)b(x) = (1+\sqrt{3})^2 + 2(1+\sqrt{3})x + x^2 \quad \text{and}$$

$$d(x)f(x)b(x) = 3(1+\sqrt{3})^2 + 9(1+\sqrt{3})x + 9x^2 + \frac{-3+3\sqrt{3}}{2}x^3.$$

Thus we structure matrix $B(x)$ and transform the polynomial matrix $B(x)$ by a series of elementary row operations as follows:

$$B(x) = \begin{pmatrix} f(x)d(x) & d(x)b(x) & d(x)f(x) & d(x)f(x)b(x) \\ g_1(x) & 0 & 0 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 3(1+\sqrt{3})+6x+\dfrac{3\sqrt{3}-3}{2}x^2 & P_1(x) & P_2(x) & P_3(x) \\ 1-\sqrt{3}+x & 0 & 0 & 0 \end{pmatrix}$$

$$\xrightarrow{(1)-\frac{3\sqrt{3}-3}{2}x(2)}\begin{pmatrix}3(1+\sqrt{3})+(12-3\sqrt{3})x & P_1(x) & P_2(x) & P_3(x)\\ 1-\sqrt{3}+x & 0 & 0 & 0\end{pmatrix}$$

$$\xrightarrow{(1)-(12-3\sqrt{3})(2)}\begin{pmatrix}18\sqrt{3}-18 & P_1(x) & P_2(x) & P_3(x)\\ 1-\sqrt{3}+x & 0 & 0 & 0\end{pmatrix}$$

$$\xrightarrow{\frac{1}{18\sqrt{3}-18}(1)}\begin{pmatrix}1 & P_4(x)P_1(x) & P_4(x)P_2(x) & P_4(x)P_3(x)\\ 1-\sqrt{3}+x & 0 & 0 & 0\end{pmatrix}$$

Where

$$P_1(x)=(1+\sqrt{3})^2+2(1+\sqrt{3})x+x^2,$$

$$P_2(x)=3(1+\sqrt{3})+6x+\frac{3\sqrt{3}-3}{2}x^2,$$

$$P_3(x)=3(1+\sqrt{3})^2+9(1+\sqrt{3})x+9x^2+\frac{-3+3\sqrt{3}}{2}x^3,$$

$$P_4(x)=\frac{1}{18\sqrt{3}-18}.$$

Substituting x by $\Theta_{(2,-2)}$ in polynomial

$$r(x)=\frac{(\sqrt{3}+1)^3}{12}+\frac{(\sqrt{3}+1)^2}{4}x+\frac{(\sqrt{3}+1)}{4}x^2-\frac{1}{12}x^3,\quad \text{we get the RFM2L2R}$$

circulant matrix $R=r(\Theta_{(2,-2)})=AA^{\{1,2\}}B=\mathrm{RFM\,2\,LRcirc}_2\mathrm{fr}(\sqrt{3}+1,1).$
Since the last column of R is b, then the RFM2L2R circulant linear system
$AX=b$ has a solution. Substituting x by $\Theta_{(2,-2)}$ in polynomial

$$c(x)=\frac{(1+\sqrt{3})^3}{36}+\frac{(1+\sqrt{3})^2}{18}x+\frac{(1+\sqrt{3})}{36}x^2\quad \text{and}$$

$$e(x)=\frac{(1+\sqrt{3})^3}{12}+\frac{(1+\sqrt{3})^2}{6}x+\frac{1}{12}x^2,\quad \text{we have the RFM2L2R circulant}$$

matrix $C=c(\Theta_{(2,-2)})=A^{\{1,2\}}B=\mathrm{RFM\,2\,LRcirc}_2\mathrm{fr}(\frac{\sqrt{3}+3}{6},\frac{\sqrt{3}}{6})$ and

$E=e(\Theta_{(2,-2)})=AA^{\{1,2\}}=\mathrm{RFM\,2\,LRcirc}_2\mathrm{fr}(\frac{\sqrt{3}+3}{6},\frac{\sqrt{3}}{6}).$ So the unique

special solution of $AX=b$ is C, i.e., $X_1(x)=(\frac{\sqrt{3}+3}{18},\frac{\sqrt{3}}{9})^T$ and a general

solution of $AX=b$ is

$$X_2 = X_1 + (I_n - E)Y$$
$$= \frac{1}{18}(3 + \sqrt{3} + 3(3 - \sqrt{3})k_1 - 3\sqrt{3}k_2, 2\sqrt{3} - 6\sqrt{3}k_1 + 3(3 + \sqrt{3})k_2)^T,$$

where $Y = (k_1, k_2)^T \in C$.

References

1. Williams, F.J.M., Sloane, N.J.A.: The Theory Error-correcting Codes. North-Holland, Amsterdam (1981)
2. David, C.: Regular Representations of Semisimple Algebras, Separable Field Extensions, Group Characters, Generalized Circulants, and Generalized Cyclic Codes. Linear Algebra Appl. 218, 147–183 (1995)
3. He, C.Y.: On the Fast Solution of r-circulant Linear Systems. Syst. Syst. Sci. Math. 21(2), 182–189 (2001) (in Chinese)
4. Cline, R.E., Plemmons, R.J., Worm, G.: Generalized Inverses of Toeplitz Matrix. Linear Algebra Appl. 8, 25–33 (1974)
5. Chen, M.K.: On the Solution of Circulant Linear System. SIAM J. Numer. Anal. 24(3), 668–683 (1987)
6. Jiang, Z.L., Xu, Z.B.: A New Algorithm for Computing the Inverse and Generalized Inverse of the Scaled Factor Circulant Matrix. J. Computa. Math. 26, 112–122 (2008)
7. Jiang, Z.L., Xu, Z.B.: Efficient Algorithm for Finding the Inverse and Group Inverse of FLS r-circulant Matrix. J. Appl. Math. Comput. 18, 45–57 (2005)
8. Jiang, Z.L.: Fast Algorithms for Solving FLS r-circulant Linear Systems. In: SCET 2012 (Xi'an), pp. 141–144 (2012)
9. Jiang, Z., Lu, F.: The Sum and Product of Jacobsthal and Jacobsthal-Lucas Numbers Representation by Matrix Method. Ars Combinatoria, 110 (July 2013)
10. Shen, N., Jiang, Z., Li, J.: On Explicit Determinants of the RFMLR and RLMFL Circulant Matrices Involving Certain Famous Numbers. WSEAS Transactions on Mathematics 12(1), 42–53 (2013)
11. Jiang, Z.: Efficient Algorithms for Finding the Minimal Polynomials and the Inverses of Level-k FLS-circulant Matrices. WSEAS Transactions on Mathematics (to appear, 2013)
12. Li, J., Jiang, Z., Shen, N.: Explicit Determinants of the Fibonacci RFPL Circulant and Lucas RFPL Circulant Matrix. JP Journal of Algebra, Number Theory and Applications 28(2), 167–179 (2013)

Determinant of the Generalized Lucas RSFMLR Circulant Matrices in Communication

Yanpeng Zheng, Sugoog Shon[*], Sanghyup Lee, and Dongpyo Oh

Dept. of Information and Telecommunications Engineering, The University of Suwon, Wau-ri, Bongdam-eup, Hwaseong-si, Gyeonggi-do, 445-743, Korea
{zhengyanpeng0702,godo0507}@gmail.com, sshon@suwon.ac.kr, odp1382@naver.com

Abstract. In this paper, the explicit determinants are presented by using generalized Lucas numbers. The techniques used herein are based on the inverse factorization of polynomial. Firstly, we introduce the definitions of the RSFMLR and RSLMFL circulant matrices in communication, and properties of the related generalized Lucas numbers. Then, we present the main results and the detailed process.

Keywords: Determinant, Generalized Lucas numbers, RSFMLR circulant matrix.

1   Introduction

The generalized Lucas numbers $\{L_n\}$ [1] are defined by a third-order recurrence

$$L_n = L_{n-1} + L_{n-2} + L_{n-3}, \quad n \geq 3, \tag{1}$$

with the initial condition $L_0 = 3, L_1 = 1, L_2 = 3$.

Recurrences (1) as well imply the characteristic equation $x^3 - x^2 - x - 1 = 0$, and its roots are denoted by t_1, t_2, t_3. Then

$$\begin{cases} t_1 + t_2 + t_3 = 1 \\ t_1 t_2 + t_1 t_3 + t_2 t_3 = -1 \\ t_1 t_2 t_3 = 1 \end{cases} \tag{2}$$

Furthermore, the Binet form for the generalized Lucas number is

$$L_n = t_1^n + t_2^n + t_3^n. \tag{3}$$

Circulant matrix family have important applications in various disciplines including image processing, communications, signal processing, encoding, and

---

[*] Corresponding author.

preconditioner. They have been put on firm basis with the work of P. Davis [2] and Z. L. Jiang [13].

There are many interests in properties and generalization of some special matrices with famous numbers. Jaiswal evaluated some determinants of circulant whose elements are the generalized Fibonacci numbers [4]. Lin gave the determinant of the Fibonacci-Lucas quasi-cyclic matrices [5]. Lind presented the determinants of circulant and skew-circulant involving Fibonacci numbers in [6]. Shen [7] discussed the determinant of circulant matrix involving Fibonacci and Lucas numbers. Akbulak [8] gave the norms of Toeplitz involving Fibonacci and Lucas numbers. The authors [9, 10] discussed some properties of Fibonacci and Lucas matrices. Stanimirović gave generalized Fibonacci and Lucas matrix in [11].

Definition 1. A row skew first-minus-last right (RSFMLR) circulant matrix with the first row $(a_0, a_1, \ldots, a_{n-1})$, denoted by RSFMLRcircfr$(a_0, a_1, \ldots, a_{n-1})$, is meant a square matrix of the form

$$A = \begin{pmatrix} a_0 & a_1 & \cdots & a_{n-2} & a_{n-1} \\ -a_{n-1} & a_0 - a_{n-1} & \cdots & a_{n-3} & a_{n-2} \\ \vdots & -a_{n-1} - a_{n-2} & \ddots & \ddots & \vdots \\ -a_2 & \vdots & \ddots & \ddots & a_1 \\ -a_1 & -a_2 - a_1 & \cdots & -a_{n-1} - a_{n-2} & a_0 - a_{n-1} \end{pmatrix}_{n \times n}.$$

It can be seen that the matrix with an arbitrary first row and the following   rule for obtaining any other row from the previous one: Get the $i+1^{st}$ row through the first element of the i th row minus the last element of the i th row, and -1 times the last element of the i th row , and then shifting the elements of the i th row (cyclically) one position to the right.

Note that the RSFMLR circulant matrix is a $x^n + x + 1$ circulant matrix [3], and that is neither the extension of skew circulant matrix [12] nor its special case and they are two different kinds of special matrices. Moreover, it is a RSFP r LR circulant matrix [14] and is also a ULS r -circulant matrix [15]with $r = -1$.

If the first row of a RSFMLR circulant matrix be $(L_1, L_2, \ldots, L_n)$, then the matrix is called generalized Lucas RSFMLR circulant matrix.

We define $\Theta_{(-1,1)}$ as the basic RSFMLR circulant matrix, that is

$$\Theta_{(-1,1)} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \cdots & \cdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -1 & -1 & 0 & \cdots & 0 & 0 \end{pmatrix}_{n \times n}$$

It is easily verified that $g(x) = x^n + x + 1$ is both the minimal polynomial and the characteristic polynomial of the matrix $\Theta_{(-1,1)}$. In addition, $\Theta_{(-1,1)}$ is nonderogatory and satisfies $\Theta^j_{(-1,-1)}$ =RSFMLRcircfr $(0,\ldots,0,1,0,\ldots,0)$ and $\Theta^n_{(-1,-1)} = -I_n - \Theta_{(-1,-1)}$.

Then a matrix A can be written in the form

$$A = f(\Theta_{(-1,-1)}) = \sum_{i=0}^{n-1} a_i \Theta^i_{(-1,-1)}$$

if and only if A is a RSFMLR circulant matrix, where the polynomial $f(x) = \sum_{i=0}^{n-1} a_i x^i$ is called the representer of the RSFMLR circulant matrix A.

Definition 2. A row skew last-minus-first left (RSLMFL) circulant matrix with the first row $(a_0, a_1, \ldots, a_{n-1})$, denoted by RSLMFLcircfr $(a_0, a_1, \ldots, a_{n-1})$, is meant a square matrix of the form

$$C = \begin{pmatrix} a_0 & a_1 & \cdots & a_{n-2} & a_{n-1} \\ a_1 & a_2 & \cdot^{\cdot} & a_{n-1}-a_0 & -a_0 \\ \vdots & \cdot^{\cdot} & \cdot^{\cdot} & -a_0-a_1 & \vdots \\ a_{n-2} & \cdot^{\cdot} & \cdot^{\cdot} & \vdots & -a_{n-3} \\ a_{n-1}-a_0 & -a_0-a_1 & \cdots & -a_{n-3}-a_{n-2} & -a_{n-2} \end{pmatrix}_{n\times n}$$

If the first row of a RSLMFL circulant matrix be $(L_1, L_2, \ldots, L_n)$, then the matrix is called generalized Lucas RSLMFL circulant matrix.

Let A =RSFMLRcircfr $(a_n, a_{n-1}, \ldots, a_1)$ and B =RSLMFLcircfr $(a_1, a_2, \ldots, a_n)$.

By explicit computation, we find

$$A = B \hat{I}_n \qquad (4)$$

where $\hat{I}_n$ is the backward identity matrix of the form

$$\hat{I}_n = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdot^{\cdot} & 1 & 0 \\ \vdots & \cdot^{\cdot} & \cdot^{\cdot} & \cdot^{\cdot} & \vdots \\ 0 & 1 & \cdot^{\cdot} & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{pmatrix} \qquad (5)$$

By Proposition 5.1 in [15], we deduce the following lemma:

Lemma 1. Let A = RSFMLRcircfr $(a_0, a_1, \ldots, a_{n-1})$. Then we have $\lambda_i = f(\tau_i) = \sum_{j=1}^{n-1} a_j \tau_i^j$ and $\det A = \prod_{i=1}^{n} \sum_{j=0}^{n-1} a_j \tau_i^j$,where $\tau_i (i = 1, 2, \ldots, n)$ are the roots of the equation

$$x^n + x + 1 = 0. \tag{6}$$

Lemma 2. Let $\tau_i (i = 1, 2, \ldots, n)$ be the roots of the equation (6). If $a = 0$, then

$$\prod_{i=1}^{n} (a\tau_i^3 + b\tau_i^2 + c\tau_i + d) = \prod_{i=1}^{n} (b\tau_i^2 + c\tau_i + d)$$

$$= d^n + b^{n-1}(b - c + d) + d(\rho_1^{n-1} + \rho_2^{n-1}) + (\rho_1^n + \rho_2^n),$$

where $a, b, c \in R$ and $\rho_1 = \dfrac{-c + \sqrt{c^2 - 4bd}}{2}$; $\rho_2 = \dfrac{-c - \sqrt{c^2 - 4bd}}{2}$.

If $a \neq 0$, then

$$\prod_{i=1}^{n} (a\tau_i^3 + b\tau_i^2 + c\tau_i + d) = \frac{(-a)^n}{2}(X_n^2 - X_{2n} - 2X_{n+1} + 2X_n)$$

$$+ \frac{(-a)^{n-1}}{2} d(X_{n-1}^2 - X_{2(n-1)} + 2X_{n-1}) + (-a)^{n-1}(bX_n - a + b - c + d) + d^n$$

where $X_n = x_1^n + x_2^n + x_3^n$, and x_1, x_2, x_3 are the roots of the equation $a\tau_i^3 + b\tau_i^2 + c\tau_i + d = 0$.

Proof. By the $\tau_i (i = 1, 2, \ldots, n)$ satisfy equation (6), we have

$$x^n + x + 1 = \prod_{i=1}^{n} (x - \tau_i).$$

If $a = 0$, please see [Shen and Jiang, WSEAS Trans. Math., 12(1), pp. 42—53 (2013)] for details of the proof. If $a \neq 0$,then

$$\prod_{i=1}^{n} (a\tau_i^3 + b\tau_i^2 + c\tau_i + d)$$

$$= a^n \prod_{i=1}^{n} (\tau_i^3 + \frac{b}{a}\tau_i^2 + \frac{c}{a}\tau_i + \frac{d}{a})$$

$$= a^n \prod_{i=1}^{n} (\tau_i - x_1)(\tau_i - x_2)(\tau_i - x_3)$$

$$= (-a)^n \prod_{i=1}^{n} (x_1 - \tau_i) \prod_{i=1}^{n} (x_2 - \tau_i) \prod_{i=1}^{n} (x_3 - \tau_i)$$

$$= (-a)^n (x_1^n + x_1 + 1)(x_2^n + x_2 + 1)(x_3^n + x_3 + 1)$$

$$= (-a)^n \{ (x_1x_2x_3)^n + x_1x_2x_3[(x_1x_2)^{n-1} + (x_1x_3)^{n-1} + (x_2x_3)^{n-1}]$$
$$+ [(x_1x_2)^n + (x_1x_3)^n + (x_2x_3)^n] + x_1x_2x_3\left(x_1^{n-1} + x_2^{n-1} + x_3^{n-1}\right)$$
$$+ [x_1^n(x_2 + x_3) + x_2^n(x_1 + x_3) + x_3^n(x_1 + x_2)] + (x_1^n + x_2^n + x_3^n)$$
$$+ x_1x_2x_3 + (x_1x_2 + x_1x_3 + x_2x_3) + (x_1 + x_2 + x_3) + 1 \}$$

Let $X_n = x_1^n + x_2^n + x_3^n$, we obtain

$$(x_1x_2)^n + (x_1x_3)^n + (x_2x_3)^n = \frac{X_n^2 - X_{2n}}{2}$$

from $(x_1^n + x_2^n + x_3^n)^2 = x_1^{2n} + x_2^{2n} + x_3^{2n} + 2[(x_1x_2)^n + (x_1x_3)^n + (x_2x_3)^n]$.

Taking the relation of roots and coeffients

$$\begin{cases} x_1 + x_2 + x_3 = -\dfrac{b}{a} \\[2mm] x_1x_2 + x_1x_3 + x_2x_3 = \dfrac{c}{a} \\[2mm] x_1x_2x_3 = -\dfrac{d}{a} \end{cases}$$

into account, we deduce that

$$\prod_{i=1}^{n}(a\tau_i^3 + b\tau_i^2 + c\tau_i + d) = \frac{(-a)^n}{2}(X_n^2 - X_{2n} - 2X_{n+1} + 2X_n)$$
$$+ \frac{(-a)^{n-1}}{2}d(X_{n-1}^2 - X_{2(n-1)} + 2X_{n-1}) + (-a)^{n-1}(bX_n - a + b - c + d) + d^n$$

2   Determinant of the Generalized Lucas RSFMLR and RSLMFL Circulant Matrix

Theorem 1. Let $A = \text{RSFMLRcircfr}(L_1, L_2, \ldots, L_n)$ be a generalized Lucas RSFMLR circulant matrix, If n is odd, then

$$\det A = \frac{-L_n^n(X_n^2 - X_{2n} - 2X_{n+1} + 2X_n) + 2L_n^{n-1}(2L_nX_n + L_{n-1}X_n + 3X_n + 2)}{Y_n^2 + Y_{n-1}^2 - Y_{2n} - Y_{2(n-1)} - 2Y_{n+1} + 2Y_{n-1} + 6}$$
$$+ \frac{L_n^{n-1}(1 + L_{n+1})(X_{n-1}^2 - X_{2(n-1)} + 2X_{n-1}) + 2(1 + L_{n+1})^n}{Y_n^2 + Y_{n-1}^2 - Y_{2n} - Y_{2(n-1)} - 2Y_{n+1} + 2Y_{n-1} + 6},$$

and if n is even, then

$$\det A = \frac{L_n^n(X_n^2 - X_{2n} - 2X_{n+1} + 2X_n) - 2L_n^{n-1}(2L_nX_n + L_{n-1}X_n + 3X_n + 2)}{Y_n^2 + Y_{n-1}^2 - Y_{2n} - Y_{2(n-1)} - 2Y_{n+1} + 2Y_{n-1} + 6}$$

$$-\frac{L_n^{n-1}(1 + L_{n+1})(X_{n-1}^2 - X_{2(n-1)} + 2X_{n-1}) - 2(1 + L_{n+1})^n}{Y_n^2 + Y_{n-1}^2 - Y_{2n} - Y_{2(n-1)} - 2Y_{n+1} + 2Y_{n-1} + 6},$$

where $X_n = \alpha_1^n + \alpha_2^n + \alpha_3^n, \alpha_1, \alpha_2, \alpha_3$ are the roots of the equation $L_n x^3 + (2L_n + L_{n-1} + 3)x^2 + (L_{n+2} + 2)x + L_{n+1} + 1 = 0$ and $Y_n = y_1^n + y_2^n + y_3^n$ y_1, y_2, y_3 are the roots of the equation $y^3 + y^2 + y - 1 = 0$.

Proof. The matrix A has the form

$$A = \begin{pmatrix} L_1 & L_2 & \cdots & L_n \\ -L_n & L_1 - L_n & \cdots & L_{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ -L_3 & -L_4 - L_3 & \cdots & L_2 \\ -L_2 & -L_3 - L_2 & \cdots & L_1 - L_n \end{pmatrix}.$$

According to Lemma 1, the Binet form (3) and (2), we have

$$\det A = \prod_{i=1}^n (L_1 + L_2\tau_i + \cdots + L_n\tau_i^{n-1})$$

$$= \prod_{i=1}^n \sum_{k=1}^n \sum_{j=1}^3 t_j^k \tau_i^{k-1} = \prod_{i=1}^n \sum_{j=1}^3 \frac{t_j(1 - t_j^n\tau_i^n)}{1 - t_j\tau_i}$$

$$= \prod_{i=1}^n \frac{L_n\tau_i^3 + (2L_n + L_{n-1} + 3)\tau_i^2 + (2L_n + 2L_{n-1} + L_{n-2} + 2)\tau_i + 1 + L_{n+1}}{-\tau_i^3 - \tau_i^2 - \tau_i + 1}$$

$$= \prod_{i=1}^n \frac{L_n\tau_i^3 + (2L_n + L_{n-1} + 3)\tau_i^2 + (L_{n+2} + 2)\tau_i + 1 + L_{n+1}}{-\tau_i^3 - \tau_i^2 - \tau_i + 1}$$

From Lemma 2 and (3), we obtain

$$\prod_{i=1}^n \left[L_n\tau_i^3 + (2L_n + L_{n-1} + 3)\tau_i^2 + (L_{n+2} + 2)\tau_i + 1 + L_{n+1} \right]$$

$$= \frac{1}{2}(-L_n)^n(X_n^2 - X_{2n} - 2X_{n+1} + 2X_n)$$

$$+ (-L_n)^{n-1}(2L_nX_n + L_{n-1}X_n + 3X_n + 2)$$

$$+ \frac{1}{2}(-L_n)^{n-1}(1 + L_{n+1})(X_{n-1}^2 - X_{2(n-1)} + 2X_{n-1}) + (1 + L_{n+1})^n,$$

where $X_n = \alpha_1^n + \alpha_2^n + \alpha_3^n, \alpha_1, \alpha_2, \alpha_3$ are the roots of the equation $L_n x^3 + (2L_n + L_{n-1} + 3)x^2 + (L_{n+2} + 2)x + L_{n+1} + 1 = 0$. And

$$\prod_{i=1}^{n}(-\tau_i^3 - \tau_i^2 - \tau_i + 1) = \frac{1}{2}(Y_n^2 + Y_{n-1}^2 - Y_{2n} - Y_{2(n-1)} - 2Y_{n+1} + 2Y_{n-1} + 6),$$

where $Y_n = y_1^n + y_2^n + y_3^n$, y_1, y_2, y_3 are the roots of the equation $y^3 + y^2 + y - 1 = 0$.

Hense, if n is odd, then

$$\det A = \frac{-L_n^n(X_n^2 - X_{2n} - 2X_{n+1} + 2X_n) + 2L_n^{n-1}(2L_n X_n + L_{n-1}X_n + 3X_n + 2)}{Y_n^2 + Y_{n-1}^2 - Y_{2n} - Y_{2(n-1)} - 2Y_{n+1} + 2Y_{n-1} + 6}$$

$$+ \frac{L_n^{n-1}(1 + L_{n+1})(X_{n-1}^2 - X_{2(n-1)} + 2X_{n-1}) + 2(1 + L_{n+1})^n}{Y_n^2 + Y_{n-1}^2 - Y_{2n} - Y_{2(n-1)} - 2Y_{n+1} + 2Y_{n-1} + 6},$$

and if n is even, then

$$\det A = \frac{L_n^n(X_n^2 - X_{2n} - 2X_{n+1} + 2X_n) - 2L_n^{n-1}(2L_n X_n + L_{n-1}X_n + 3X_n + 2)}{Y_n^2 + Y_{n-1}^2 - Y_{2n} - Y_{2(n-1)} - 2Y_{n+1} + 2Y_{n-1} + 6}$$

$$- \frac{L_n^{n-1}(1 + L_{n+1})(X_{n-1}^2 - X_{2(n-1)} + 2X_{n-1}) - 2(1 + L_{n+1})^n}{Y_n^2 + Y_{n-1}^2 - Y_{2n} - Y_{2(n-1)} - 2Y_{n+1} + 2Y_{n-1} + 6}.$$

Theorem 2. Let B = RSFMLRcircfr$(L_n, L_{n-1}, \ldots, L_1)$. If n is odd, then

$$\det B = \frac{-Z_n^2 + Z_{2n} + 2Z_{n+1} + 2(L_{n+1} + 2)Z_n + 2L_{n+1} - 2L_{n-1}}{L_n^2 + L_{n-1}^2 - L_{2n} - L_{2(n-1)} + 2L_n - 2L_{n-2} + 6}$$

$$+ \frac{(L_n + 3)(Z_{n-1}^2 - Z_{2(n-1)} + 2Z_{n-1}) + 2(L_n + 3)^n}{L_n^2 + L_{n-1}^2 - L_{2n} - L_{2(n-1)} + 2L_n - 2L_{n-2} + 6},$$

and if n is even, then

$$\det B = \frac{Z_n^2 - Z_{2n} - 2Z_{n+1} - 2(L_{n+1} + 2)Z_n - 2L_{n+1} + 2L_{n-1}}{L_n^2 + L_{n-1}^2 - L_{2n} - L_{2(n-1)} + 2L_n - 2L_{n-2} + 6}$$

$$- \frac{(L_n + 3)(Z_{n-1}^2 - Z_{2(n-1)} + 2Z_{n-1}) - 2(L_n + 3)^n}{L_n^2 + L_{n-1}^2 - L_{2n} - L_{2(n-1)} + 2L_n - 2L_{n-2} + 6}.$$

where $Z_n = \beta_1^n + \beta_2^n + \beta_3^n, \beta_1, \beta_2, \beta_3$ are the roots of the equation $x^3 + (L_{n+1} + 3)x^2 + (L_{n+2} - L_{n+1} + 5)x + L_n + 3 = 0.$

Proof. The matrix B has the form

$$B = \begin{pmatrix} L_n & L_{n-1} & \cdots & L_1 \\ -L_1 & L_n - L_1 & \cdots & L_2 \\ \vdots & \vdots & \ddots & \vdots \\ -L_{n-2} & -L_{n-3} - L_{n-2} & \cdots & L_{n-1} \\ -L_{n-1} & -L_{n-2} - L_{n-1} & \cdots & L_n - L_1 \end{pmatrix}.$$

According to Lemma 1, (3) and (2), we have

$$\det B = \prod_{i=1}^{n}(L_n + L_{n-1}\tau_i + \cdots + L_1\tau_i^{n-1})$$

$$= \prod_{i=1}^{n}\sum_{k=0}^{n-1}\sum_{j=1}^{3} t_j^{n-k}\tau_i^k = \prod_{i=1}^{n}\sum_{j=1}^{3}\frac{t_j^{n+1} - t_j\tau_i^n}{t_j - \tau_i}$$

$$= \prod_{i=1}^{n}\frac{\tau_i^3 + (L_{n+1} + 3)\tau_i^2 + (L_{n+2} - L_{n+1} + 5)\tau_i + L_n + 3}{-\tau_i^3 + \tau_i^2 + \tau_i + 1}.$$

By lemma 2, the Binet form (3), we obtain

$$= \prod_{i=1}^{n}[\tau_i^3 + (L_{n+1} + 3)\tau_i^2 + (L_{n+2} - L_{n+1} + 5)\tau_i + L_n + 3]$$

$$= \frac{1}{2}(-1)^n(Z_n^2 - Z_{2n} - 2Z_{n+1} + 2Z_n) + (-1)^{n-1}(L_{n+1}Z_n + 3Z_n + L_{n+1} - L_{n-1})$$

$$+ \frac{1}{2}(-1)^{n-1}(L_n + 3)(Z_{n-1}^2 - Z_{2(n-1)} + 2Z_{n-1}) + (L_n + 3)^n,$$

where $Z_n = \beta_1^n + \beta_2^n + \beta_3^n, \beta_1, \beta_2, \beta_3$ are the roots of the equation $x^3 + (L_{n+1} +$

$3)x^2 + (L_{n+2} - L_{n+1} + 5)x + L_n + 3 = 0.$ And

$$\prod_{i=1}^{n}(-\tau_i^3 + \tau_i^2 + \tau_i + 1) = \frac{1}{2}(L_n^2 + L_{n-1}^2 - L_{2n} - L_{2(n-1)} + 2L_n - 2L_{n-2} + 6).$$

Thus, if n is odd, then

$$\det B = \frac{-Z_n^2 + Z_{2n} + 2Z_{n+1} + 2(L_{n+1} + 2)Z_n + 2L_{n+1} - 2L_{n-1}}{L_n^2 + L_{n-1}^2 - L_{2n} - L_{2(n-1)} + 2L_n - 2L_{n-2} + 6}$$

$$+ \frac{(L_n + 3)(Z_{n-1}^2 - Z_{2(n-1)} + 2Z_{n-1}) + 2(L_n + 3)^n}{L_n^2 + L_{n-1}^2 - L_{2n} - L_{2(n-1)} + 2L_n - 2L_{n-2} + 6},$$

and if n is even, then

$$\det B = \frac{Z_n^2 - Z_{2n} - 2Z_{n+1} - 2(L_{n+1}+2)Z_n - 2L_{n+1} + 2L_{n-1}}{L_n^2 + L_{n-1}^2 - L_{2n} - L_{2(n-1)} + 2L_n - 2L_{n-2} + 6}$$

$$-\frac{(L_n+3)(Z_{n-1}^2 - Z_{2(n-1)} + 2Z_{n-1}) - 2(L_n+3)^n}{L_n^2 + L_{n-1}^2 - L_{2n} - L_{2(n-1)} + 2L_n - 2L_{n-2} + 6}.$$

Theorem 3. Let $C = \mathrm{RSLMFLcircfr}(L_1, L_2, \ldots, L_n)$ be the a generalized Lucas RSFMLR circulant matrix, If $n \equiv 0 \pmod 4$, then

$$\det C = \frac{Z_n^2 - Z_{2n} - 2Z_{n+1} - 2(L_{n+1}+2)Z_n - 2L_{n+1} + 2L_{n-1}}{L_n^2 + L_{n-1}^2 - L_{2n} - L_{2(n-1)} + 2L_n - 2L_{n-2} + 6}$$

$$-\frac{(L_n+3)(Z_{n-1}^2 - Z_{2(n-1)} + 2Z_{n-1}) - 2(L_n+3)^n}{L_n^2 + L_{n-1}^2 - L_{2n} - L_{2(n-1)} + 2L_n - 2L_{n-2} + 6},$$

and if $n \equiv 1 \pmod 4$, then

$$\det C = \frac{-Z_n^2 + Z_{2n} + 2Z_{n+1} + 2(L_{n+1}+2)Z_n + 2L_{n+1} - 2L_{n-1}}{L_n^2 + L_{n-1}^2 - L_{2n} - L_{2(n-1)} + 2L_n - 2L_{n-2} + 6}$$

$$+\frac{(L_n+3)(Z_{n-1}^2 - Z_{2(n-1)} + 2Z_{n-1}) + 2(L_n+3)^n}{L_n^2 + L_{n-1}^2 - L_{2n} - L_{2(n-1)} + 2L_n - 2L_{n-2} + 6},$$

and if $n \equiv 2 \pmod 4$, then

$$\det C = \frac{Z_n^2 - Z_{2n} - 2Z_{n+1} - 2(L_{n+1}+2)Z_n - 2L_{n+1} + 2L_{n-1}}{-L_n^2 - L_{n-1}^2 + L_{2n} + L_{2(n-1)} - 2L_n + 2L_{n-2} - 6}$$

$$+\frac{(L_n+3)(Z_{n-1}^2 - Z_{2(n-1)} + 2Z_{n-1}) - 2(L_n+3)^n}{L_n^2 + L_{n-1}^2 - L_{2n} - L_{2(n-1)} + 2L_n - 2L_{n-2} + 6},$$

and if $n \equiv 3 \pmod 4$, then

$$\det C = \frac{Z_n^2 - Z_{2n} - 2Z_{n+1} - 2(L_{n+1}+2)Z_n - 2L_{n+1} + 2L_{n-1}}{L_n^2 + L_{n-1}^2 - L_{2n} - L_{2(n-1)} + 2L_n - 2L_{n-2} + 6}$$

$$-\frac{(L_n+3)(Z_{n-1}^2 - Z_{2(n-1)} + 2Z_{n-1}) + 2(L_n+3)^n}{L_n^2 + L_{n-1}^2 - L_{2n} - L_{2(n-1)} + 2L_n - 2L_{n-2} + 6},$$

where $Z_n = \beta_1^n + \beta_2^n + \beta_3^n$, $\beta_1, \beta_2, \beta_3$ are the roots of the equation $x^3 + (L_{n+1} + 3)x^2 + (L_{n+2} - L_{n+1} + 5)x + L_n + 3 = 0$.

Proof. The theorem can be proved by using Theorem 2 and the equation (4).

References

1. Michele, E.: Derived Sequences, the Tribonacci Recurrence and Cubic Forms. Fibonacci Quart. 39(2), 107–115 (2001)
2. Davis, P.: Circulant Matrices. Wiley, New York (1979)
3. David, C.: Regular Representations of Semisimple Algebras, Separable Field Extensions, Group Characters, Generalized Circulants, and Generalized Cyclic Codes. Linear Algebra and its Appl. 218, 147–183 (1995)
4. Jaiswal, D.: On Determinants Involving Generalized Fibonacci Numbers. Fibonacci Quart. 7, 319–330 (1969)
5. Lin, D.: Fibonacci-Lucas Quasi-cyclic Matrices. Fibonacci Quart. 40, 280–286 (2002)
6. Lind, D.: A Fibonacci Circulant. Fibonacci Quart. 8, 449–455 (1970)
7. Shen, S.Q., Cen, J.M., Hao, Y.: On the Determinants and Inverses of Circulant Matrices with Fibonacci and Lucas Numbers. Appl. Math. Comput. 217, 9790–9797 (2011)
8. Akbulak, M., Bozkurt, D.: On the Norms of Toeplitz Matrices Involving Fibonacci and Lucas Numbers. Hacet. J. Math. Stat. 37, 89–95 (2008)
9. Lee, G.Y., Kim, J.S., Lee, S.G.: Factorizations and Eigenvalues of Fibonacci and Symmetric Fibonacci Matrices. Fibonacci Quart. 40, 203–211 (2002)
10. Miladinović, M., Stanimirović, P.: Singular Case of Generalized Fibonacci and Lucas Matrices. J. Korean Math. Soc. 48, 33–48 (2011)
11. Stanimirović, P., Nikolov, J., Stanimirović, I.: A generalization of Fibonacci and Lucas matrices. Discrete Appl. Math. 156, 2606–2619 (2008)
12. Jiang, X.Y., Gao, Y., Jiang, Z.L.: Explicit Determinants and Inverses of Skew Circulant and Skew Left Circulant Matrices Involving the k-Fibonacci Numbers in Communications-I. Far East Journal of Mathematical Sciences 76(1), 123–137 (2013)
13. Jiang, Z.L., Zhou, Z.X.: Circulant Matrices. Chengdu Technology University Publishing Company, Chengdu (1999)
14. Jiang, X.Y., Jiang, Z.L.: Efficient Algorithm for Finding Inverse and Group Inverse of the RSFPrLR Circulant Matrix in Codes. JP Journal of Algebra, Number Theory and Applications 29(1), 51–70 (2013)
15. Tian, Z.P.: Fast Algorithms for Solving the Inverse Problem of $Ax=b$ in the Class of the ULS r-circulant (retrocirculant) Matrices. Int. J. Algebra 5, 9403–9411 (2011)

Wireless Broadband Networks
in Complex Mine Environment

Bin Liu

School of Computer and Information, Hefei University of Technology,
Hefei, Anhui, China

Abstract. This paper analyzes the WiFi advantages and disadvantages of various types wireless communication, elaborated WiFi communication devices and programs in mine construction; introduced wireless system overall technology roadmap and system application results.The results show that the realization of the WIFi function of the handheld device can satisfy your requirements, hardware and WIFi function works stable and reliable.

Keywords: WiFi, coal mine, wireless network.

1   Introduction

With the computer technology and the rapid development and popularization of the Internet, networks are increasingly being applied in people's life and work [1-4]. Wifi technologies emerge is its convenient networking, RF signal advantages to people's lives has brought great convenience to quickly in the wireless communications field has developed rapidly, and in the coal mine wireless communication technology applications to get people more and more and more attention [5-9]. Wireless Local Network (WLAN) refers to a radio channel to replace the traditional wired transmission medium composed of local area network, a computer network and wireless communication technology of the combination, the application of computer technology to wireless communication devices connected together to form, and can communicate with each resource sharing network system [10,11].

Wireless remote data monitoring system is designed for wireless data remote monitoring system for underground complex geological conditions and the working environment .that uses wireless switching mine complex environment transmission technology.Construction wireless remote data monitoring system platform to solve other equipment downhole remote data transmission problems, to make up for the lack of cable networks, digital mine [12-15]. Meanwhile, the wireless remote data monitoring systems, the ability to raise the overall level of safety in production of coal mining enterprises, especially in the current mining line data monitoring workplace dilemma can not be achieved to improve the enterprise's comprehensive strength and competitiveness, while achieving full coverage of downhole data monitoring objectives.

On the supervision and management of coal production operations staff to improve mine road safety, maintenance, and corporate image has a great role in promoting. Remote data monitoring system can reduce the risk of an image, the protection of the lives and safety of mine workers is important [16]. It is the introduction of advanced information technology, network technology, play all advantages, promote Mine Road leapfrog development of information technology, to improve our level of automation in production Mine Road, the prospect of unmanned mine safety goal has important strategic significance.

2   Wireless Networking in Mine

By studying the establishment of full coverage Inoue and underground broadband wireless network communication system, as the coal industry Ethernet ring network extensions and additions [17-20]. Through peer-proof and intrinsically safe wireless network switch can be realized on shearer, boring machines, winches, mobile substations and other mobile devices, frequent data and video monitoring. In the face installation of network video camera, wireless network switch explosion or install a wired base stations and wireless video camera and other equipment, data and image transmission to the control room, reducing light cables are often drag caused by the failure to raise awareness of mining and other mobile devices monitoring reliability.

2.1   The Frequency Band of WLAN

IEEE802.11 standard is the American Institute of Electrical and Electronics Engineers in 1997 Notice wireless LAN standard, is a pioneer in the development of wireless network technology [21-24]. Not only can make a variety of wireless products from different manufacturers can be interconnected, but also helps the core device performs single-chip solution for the implementation of cost reduction and wireless LAN. There are requirements in a variety of mobile environments, IEEE802.11 standard release, the wireless LAN is widely accepted, has become the common standard wireless LAN. The existing IEEE802.11 protocol group mainly IEEE802.11a, 802.11b, 802.11g, and 802.11i, etc., which have more than 10 agreements with the physical layer can be used to solve optimization, security definitions, MAC layer enhancements and interoperability and other specific issues. 802.11,802.11 b, 802.11a and 802.11g are several key physical layer standard protocol. Which is the first generation 802.11 wireless LAN standard; 802.11b, 802.11a and 802.11g is 802.11 upgrade agreement. Table 1 lists a few standard protocols and application performance comparison.

Based on the above analysis and the confined space of wireless data transmission, we use 802.11a protocol, 5G band wireless transmission. The most commonly used on the market today with 802.11a protocol chip is Marvell's 88W8686 chip production; we are using this chip for wireless data transmission design.

Table 1. Various parameters compared for WLAN wireless LAN standard technical

Name of Agreement	802.11	802.11b	802.11a	802.11g
Published	1997	1999	1999	2003
Frequency of use	2.4GHz	2.4GHz	5GHz	2.4GHz
Highest frequency	2Mbps	11Mbps	54MHz	54MHz
Compatibility	-	Compatible 802.11 when 1/2Mbps	Not compatible with 802.11b	Compatible with 802.11b, 802.11a
Obligation	data	Images, data	Voice, images, data	Voice, images, data
Physical layer	-	CCK, DDS	OFDM	OFDM, CCK, PBCC
Advantage	-	Low cost, stable, mature	Interference, high transmission rate	High transmission rate
Shortcoming	-	Rate is low, large interference	High cost	Large interference
Usage	Seldom	Widely used	Not universal	Not universal

2.2   WiFi Technology

WiFi rate up to 11 Mb / s, operating in the 5 GHz band, speeds up to 54 Mb / s. Its network is convenient, easy to maintain, using a common frequency band is (ISM band), without the need for communication for the record, do not pay channel lease charges, ease of wireless network coverage Mine Road use [25-27]. Advanced WiFi technology, in line with trends, you can use the underground industrial ring network, without a separate network construction, multi-channel; system power can be automatically adjusted according to the environment, after the completion of the system to meet future bandwidth wireless video, data and other downhole variety of applications, scalability strong. Disadvantage is that information security is not high, but the bandwidth is high, suitable for transmitting video information to meet system requirements.

Downhole system supports IEEE802.11a wireless channel, to ensure real-time delivery of voice, data, video and other data streams, is currently able to realize high-speed wireless coverage mining multimedia wireless transmission systems. In the coal mine, WiFi broadband wireless communication system as the existing industrial Ethernet ring network, supplement and perfect, you can expand the coverage of broadband networks, and to ensure safety in production to improve the level of automation of production to achieve the objectives unmanned mine safety has important strategic significance, can promote information technology by leaps and bounds.

3   Structure of Emergency Rescue Communication System in Coal Mine

After the disaster in coal mine underground, underground power supply cut off completely, the environment is very complex. Therefore, the rescue communication system should be 3 parts including the ground command center, Inoue command center and the Inoue wireless rescue communication system from the structure. Rescue system

block diagram is shown in Figure 1,the mine wireless rescue communication system by underground command center, personal terminal and a relay station. Distinguish from the communication technology; it is the rescue communication system of wired and wireless communication technology combined with the.Because of the special structure of underground roadway of coal mine, wireless signal transmission in a limited space, the transmission distance than the ground near. Therefore, in the design of personal terminal, increase the power amplifier module to increase the transmission distance; and between the personal terminal and underground command center, according to the practical needs, increase the relay station, see Figure 1.

The wireless remote monitoring system is mainly coal mine WiFi wideband wireless communication system and digital video surveillance. Wireless broadband wireless communication system as the supplement and perfect the existing industrial Ethernet ring network, can expand the coverage of broadband network. By the front-end video acquisition unit technology face wireless remote video monitoring system (4 video cameras), wireless transmission link and monitor display is composed of 3 parts. The system architecture is shown in figure 2.

The front end of mine flameproof network camera captured video data after the digital compression; enter the flameproof and intrinsically safe wireless network switches, through wireless transmission link sent to the remote monitoring center. Video data to the monitoring center after the decoding flow can carry out real-time view through the monitoring terminal. The monitoring center uses a client / server architecture, the server receives the video after video decoding, the client connect into network access server, monitoring image.At the same time set in the work field flameproof monitor, synchronization of video surveillance.

Through the system structure drawing is not difficult to see, to achieve the function key is to front-end video acquisition unit (4 video camera), the 3 part of the design of wireless transmission and monitoring display.
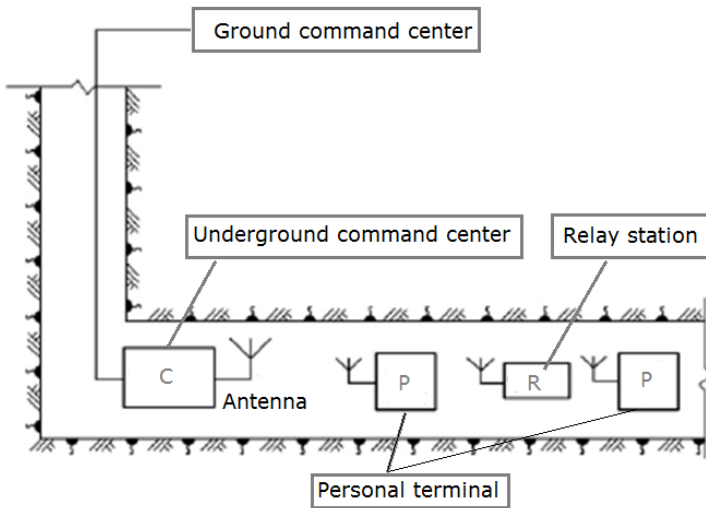
Fig. 1. Underground communication system
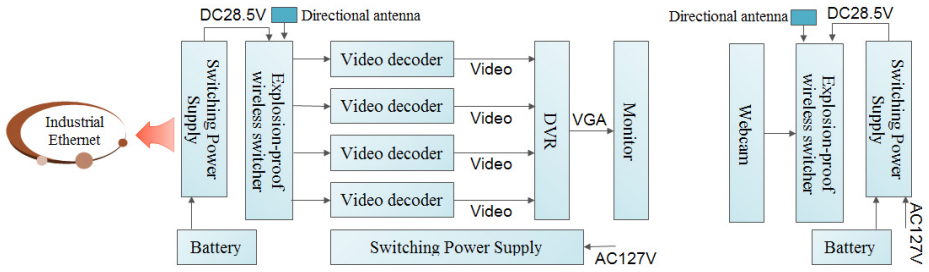
Fig. 2. System structure diagram

4   The Design of the Network Switcher

Ethernet switch with explosion-proof and intrinsically safe explosion-proof form, suitable for AC 127/660V two voltage levels, the device comes with backup power, external power failure, can still work 2 h. Structure is mainly composed of the following four parts: ① flameproof enclosure; ② Industrial Ethernet Switch Module; ③ intrinsically safe interface circuit; ④ power.

Explosion-proof industrial Ethernet switches in addition to the electrical part, the main structure consists of a flameproof enclosure attached to the shell, and some parts, such as gaskets, transparent parts, cables (wires) into devices and junction boxes.

5   The Overall Technical Route

The overall technical route of wireless system as shown in Figure 3.
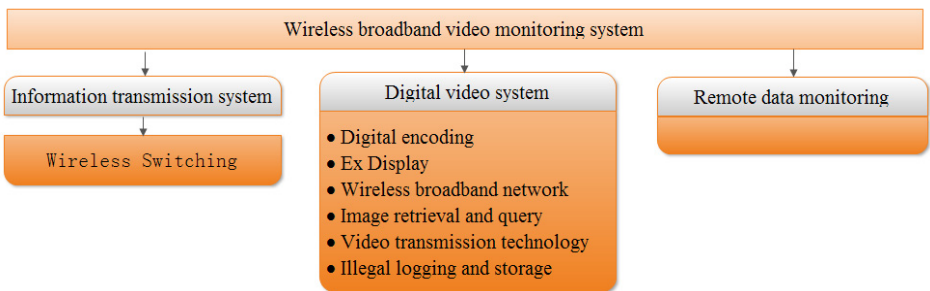
Fig. 3. Diagram of the overall technical scheme

The overall technical line is consists of 4 aspects, first is the wireless broadband scheme of video monitoring system overall construction; second is the development of broadband wireless video monitoring system of the main equipment and systems, including wireless broadband transmission device, digital video surveillance and video

remote monitoring; third is the emphases of technology; fourth is the wireless broadband video monitoring system technology integration and testing scheme.

Research and development of device through the above key technology, through the system integration and testing, we realize transportation monitoring system.

6   System Features

Mining is a high -risk industry, China's current coal-producing one million tons mortality rate is much higher than the international average. In recent years, frequent occurrence of major mining company's production safety accidents to the country and the people a huge loss of life and property. As the state and the government for the growing mining industry attaches great importance to production safety, corporate security and reliable communication system has also been listed as the most important task of all. In order to protect the mine production safety, to prevent mine accidents, we need through the introduction of video surveillance systems, monitor the whole process of mine production, in order to maximize production and operation to ensure mine safety. Currently, video surveillance systems have become an essential technology to prevent the mining industry facilities. Mine site, the special circumstances, complex terrain, seasonal temperature changes, the laying of cable networks more difficult, laying fiber was not possible, so we have to use wireless transmission way to solve the problem of mine security monitoring. Based on wireless MESH routers for the mining industry developed superior performance, quality, reliable, full-featured wireless broadband communication system solution mining, mining unit to meet production safety monitoring, mobile voice communications, personnel and equipment positioning, high-speed data transmission, office automation business information and other communication needs . Mine wireless network scenarios, integrated video surveillance, voice communications, personnel and equipment positioning and many other applications. Composed of a plurality of wireless MESH wireless transmission backbone network routers to replace traditional wired networks. Vehicle configuration camera and vehicle wireless terminals, video information and location information through the car back to the real-time wireless terminal control center, monitoring center based on feedback information to dispatch vehicles. Staff at the mine can be equipped with hand-held terminal, you can achieve voice communication and positioning functions. In the event of an accident, the system can quickly locate staff position. Wireless communication and information systems, proprietary technology based on wireless broadband transmission, with high bandwidth, support for multi-service, multipoint communication, engineering, installation convenience and other advantages, the mining industry can be perfectly applied to production safety requirements.

Using wireless bridge build backbone links, center command center set up at the mine, in the mine area select several high points to establish a relay point, each monitoring point nearest unobstructed access to the relay point, the relay point center using the bridge, point to point wireless link connectivity, eventually forming two complete wireless network surveillance video transmission of information. Wireless

monitoring system set up in the mine, for adequate protection of user investment, and can be repeated at different mines using wireless devices. Will maximize the mining complex problems such as the terrain, monitoring distribution point spread, wired network monitoring system wiring construction difficulty, high cost and other issues, and can provide high-quality video transmission to meet the requirements of network and, ultimately, mine safety and security provide the necessary support . According to the actual situation and needs of mine, mine production safety wireless video surveillance and information management work together to solve scheduling system. This system has the following advantages:

6.1   Advanced and Applicability

This system uses pure digital IP transmission, the latest technology combined with OFDM microwave MPEG4/H.264 encoded spread spectrum modulation techniques, reflects the current computer control technology and computer network technology, the latest level of development, adapt to the times development.

6.2   Economy and Convenience

Fully consider the actual needs of users and information technology trends, according to the user site environment, design selection function and fit the scene, using wired + wireless integrated solutions. Wireless network transmission system has easy to set up, short construction period, low cost, without geographical constraints affecting the merits installation quick and easy.

6.3   Reliability and Security

The system should be designed to have a high reliability, system failure or interruption caused by the accident, to ensure data accuracy, completeness and consistency, and with rapid recovery of function, and the system has complete set of system management strategies that can guarantee the safe operation of the system.

6.4   Open

Every aspect of the system, strict application of internationally accepted standards, so that the entire structure of the system showing a higher openness, compatibility, thus facilitating system upgrades and management ; while also taking into account the surrounding environment, information and communication status and technology trends, with fire, perimeter security, condenser system to achieve linkage with RJ-45 network communication port, which enables remote control.

6.5   Scalability

The system design and equipment selection are taken into account in the future development and use of technology needs, with renewals, extensions and upgrades

possible. And according to the actual requirements of the project in the future expansion of the system functionality, while leaving in the design of the program redundancy to meet future development requirements.

6.6   Fee Rationality

The final program design and equipment selection at reasonable cost to pay strict attention to the premise, not only to fully consider its performance, but also to meet the requirements of reasonable cost.

7   Conclusion

The system uses a WiFi wireless Ethernet switch, with wireless broadband communications functions, that can access video, WiFi phones, monitoring data and other information to comply with IEEE802.11 specification. By proof LCD monitor, video camera flameproof network composed of a full range of face integrated video systems. Available in explosion-proof LCD monitor display video images, real-time display face movement of persons and equipment operation. Violation of various phenomena can automatically record, ready for future reference. System has a self-diagnostic function, easy maintenance, low operating costs.

As the use of wireless communication networks, make wiring easier, and greatly reduce the amount of cable. To facilitate system expansion, they can always add nodes. Controllers and network reliability, less maintenance. Safe and reliable performance of the technology, to achieve a wireless broadband video transmission; able to reduce light cable laying, but also reduce the damage caused by light cable drag and reduce labor intensity of workers; downhole environment and the protection of the health of workers have a good effect .

References

1. Lei, L., Yang, W.: ZigBee-based Mine Safety Monitoring System. Chengdu University (Natural Science) 03 (2011)
2. Jun, S., Xiaoping, Z., Fang, Y.-J.: ZigBee-based wireless sensor networks in Substation. Electrical Applications 09 (2008)
3. Sue lift side, Gao, C., Rong, X.: Mine gas monitoring systems analysis and research. Industrial Safety and Environmental Protection 10 (2009)
4. Li, Q., Fu, C.M.: Mining technology based on Zigbee wireless smart portable anemometer. Industrial Instrumentation & Automation 03 (2011)
5. Hu-yan, Xiu, W., Tao, Y., Wang, Y.-T.: Campus-based indoor positioning LBS Research. Computer Engineering 08 (2010)
6. Jun, S.: AT91sam7s64 and based on the CC2420 Zigbee node design and implementation. Programmable Controller & Factory Automation 04 (2008)
7. Xiaodong, S., Rollins, Gang, R., Ge, T., Shengjia: Intelligent portable gas monitoring instrument design and research. Science Communication 18 (2010)

8. Stone, W., Huang, S., Liu, X.: Wince-based underground portable wireless mobile terminal design. Science Communication 03 (2011)

9. Chen, S., Zhao, T.B., Gao, J.D., Young, W.: ZigBee-based wireless mechanized mining face roof pressure monitoring system. Coal Mining 02 (2011)

10. Hua, C., Meng, Z.: GPRS-based wireless sensor networks and mine safety monitoring system. Coal 08 (2008)

11. Xiao, J., Lvai, Q., Chen, J., Zhu, M.: Wireless sensor network technology in the key issues. Sensor World 07 (2004)

12. Ma, Y., Li, J., Feng, L.: RF chip based on ZigBee technology CC2430. Microcontrollers & Embedded Systems 03 (2006)

13. Feng, Z., Jiang, Z., Zhiguo, Z.: Wireless sensor network MAC protocol. Guangdong Communication Technology 06 (2005)

14. Xiumei, Liu, N.: Utilize 2.4GHz RF chip CC2420 ZigBee Wireless Communication Design. International Electronic Elements 03 (2005)

15. Shao, Q., Wu, W., Li, X.: Adhoc Networks mine based wireless data acquisition system under design and implementation. Journal of Shenyang Institute of Aeronautical Engineering 03 (2005)

16. Lina, L., Tao, D.: Actuators; sensor network operating system TinyOS Key Technology Analysis. Harbin University of Commerce (Natural Sciences Edition) 06 (2005)

17. Yang, W., Feng, X., Cheng, X., Sun, J.: Whole new generation mine wireless information systems theory and key technology. Coal Society 04 (2004)

18. Ren, F., Huang, H., Lin, C.: Wireless sensor networks. Journal of Software 07 (2003)

19. Zhang, J., Jianwen, Wu, C.: ZigBee Technology in Coal Mine Safety Monitoring. China Measurement Technology 04 (2008)

20. Sun, X., Li, J., Zhang, G., Tian, H.: AdHoc routing algorithm commonly used in the analysis. Modern Electronics Technique 13 (2003)

21. Ying, H.: Wireless sensor network based landslide monitoring and Warning system design and research. Chongqing University (2011)

22. Guo, H.: Based on GPRS, Xbee-Pro wireless communication module build wireless sensor network applications. Wuhan University of Technology (2011)

23. Zhao, Z.: ZigBee-based wireless sensor nodes in the coal mine safety monitoring of the pre-study. Jilin University (2008)

24. Wu, C.: ZigBee-based coal mine gas monitoring system and implementation. North University (2008)

25. Paragraph Fann: Wireless sensor network routing protocol. Beijing Forestry University (2009)

26. Great: Dynamic Abnormal Methane Emission disaster response mechanism and control technology research. Coal Science Research Institute (2009)

27. Yong: Wireless sensor network technology in remote ECG acquisition Application Research. Chongqing University (2010)

Ultra Wideband OFDM Channel Estimation Algorithm Based on Instantaneous Power Filtering

Wan Xing [1] and Bi Guangguo [2]

[1] JiangSu Broadcasting Cable Information Network Corporation Limited Nanjing, Nanjing, 210000, China
[2] State Key Laboratory of Mobile Communication Southeast University Nanjing, 210000, China

Abstract. The Ultra Wide-Band (UWB) is one of the important development trends of the new-generation wireless communication technology. The channel estimation technology,is the key part of the Orthogonal Frequency-Division Multiplexing (OFDM). After the strengths and weaknesses of the Least Mean Square (LMS) algorithm in the system implementation were compared, a UWB OFDM channel estimation algorithm based on instantaneous power filtering and applicable to system implementation was put forward.

Keywords: ultra wide-band, orthogonal frequency division multiplexing, instantaneous power filtering, channel estimation.

1 Introduction

LS-based channel estimation is simple to implement but has limited performance. The LMS-based channel estimation algorithm has the theoretically superior performance but poor feasibility, and has limitation in the practical application consequently [1-3]. Well improved in both performance and feasibility, the channel estimation algorithm based on instantaneous power filtering in this study is very suitable for hardware implementation.

2 An Overview of UWB Wireless Communication

The UWB [4-6] communication first appeared in 1960s. Any wireless system can be called as the UWB system if satisfying any one of the conditions listed below: a. the relative bandwidth is 20% greater than the center frequency; b. the absolute bandwidth is greater than 500 MHz. Here the relative bandwidth is defined in [2, 8] .The IEEE 802.15.3a team defined in detail four UWB models [3]. In this study, it was assumed that the adopted channel estimation sequence accorded with the ECMA368/369 UWB Standard [4, 9-11], as shown in Fig. 1. In the following algorithm simulation, tsix OFDM signals were sent each time.

0+0j	1+1j	• • •	-1-1j	0+0j	0+0j	0+0j	0+0j	0+0j	-1+1j	• • •	1-1j
1	2	3~61	62	63	64	65	66	67	68	70~127	128

Fig. 1. Format of sending OFDM channel estimation sequence

3 Conventional UWB OFDM-Based Channel Estimation Algorithm

3.1 A.MMSE-Based Channel Estimation Algorithm

The frequency MMSE channel estimation method solves the interpolation problem of null subcarriers. If the singular value decomposition (SVD) algorithm [12-13] is further employed, the noise components can be filtered out effectively in the time domain, with the energy at the effective channels being preserved. The MMSE algorithm can be expressed as follows:

$$\mathbf{H}_{MMSE} = \mathbf{R}_{HH} \left(\mathbf{R}_{HH} + \sigma_n^2 \cdot \beta / E|x_k|^2 \mathbf{I} \right)^{-1} \tilde{\mathbf{H}}_{LS} \tag{1}$$

In the ECMA368/369 Standard, QPSK and DCM modulation are employed at the transmitting terminal, and the normalization processing of transmitted signals are performed. Therefore, the results of the SVD channel estimation in ECMA368/369 Standard can be modified as:

$$\begin{aligned}
\mathbf{H} &= \mathbf{R}_{HH} \left(\mathbf{R}_{HH} + \sigma_n^2 \cdot \mathbf{I} \right)^{-1} \mathbf{H}_{LS} \\
&= \mathbf{F} \mathbf{R}_{hh} \left(\mathbf{R}_{hh} + \sigma_n^2 \cdot \mathbf{I} \right)^{-1} \mathbf{F}^H \tilde{\mathbf{H}}_{LS}
\end{aligned} \tag{2}$$

It can be seen that the above algorithm makes use of the statistical characteristics of the noise and the channel response, and has superior performance theoretically. In practice, however, it is difficult to obtain the statistical characteristics of channels and noise. Besides, the algorithm is not suitable for non-stationary dynamic channels and requires the inversion of high-order matrices with its high requirement on the hardware as well.

3.2 UWB OFDM Channel Estimation Algorithm Based on Instantaneous Power Filtering

Based on the MMSE channel estimation and in consideration of the rather low time-domain correlation of UWB channels, the following approximation can be obtained:

$$\mathbf{R}_{hh} \approx \mathbf{diag} \left(E[|h_1|^2], E[|h_2|^2], \cdots, E[|h_N|^2], 0, \cdots, 0 \right)$$

Therefore, the MMSE channel estimation can be modified as:

$$\mathbf{H}_{TR} = \mathbf{F}\mathbf{diag}\Big[E[|h_1|^2]\big/\big(E[|h_1|^2]+\sigma_n^2\big),\cdots$$

$$,E[|h_N|^2]\big/\big(E[|h_N|^2]+\sigma_n^2\big),0,\cdots,0\Big]\mathbf{F}^H\mathbf{H}_{LS.}$$

It can be seen that the mean power at each path of the channel and the mean power of the noise need to be acquired for the MMSE-based channel estimation which is essentially a channel estimation algorithm based on mean power filtering. In practice, the only way to acquire the estimated value is using the time average method to substitute for the statistic average method. Having the drawbacks of considerable error, slow convergence speed and large amount of computation, this algorithm is not suitable for cases with large power MSE at each path and can not deal with time-varying channels and noise correspondingly, and the influence of synchronization and multi-paths is great.

Since each implementation of the impact response is different, self-adaptation is impossible with a fixed $E[|h_i|^2]\big/\big(E[|h_i|^2]+\sigma^2\big), i = 1,\cdots, N$ being used to perform filtering on each time-domain point. Therefore, the time-domain channel response and noise acquired by LS estimation at each time are employed for filtering.

$$\mathbf{H}_{Power} = \mathbf{F}\Bigg(\mathbf{diag}\bigg[|\tilde{h}_{1,LS}|^2\big/\big(|\tilde{h}_{1,LS}|^2 + \rho_1\cdot\sigma^2\big),\cdots,$$

$$,|\tilde{h}_{37,LS}|^2\big/\big(|\tilde{h}_{n,LS}|^2 + \rho_n\cdot\sigma^2\big),0,\cdots,0\bigg]\Bigg)\mathbf{F}^H\mathbf{H}_{\tilde{LS}.}$$

$\tilde{h}_{i,LS}$ is the mean of six OFDMs and contains the noise. It can be expressed as follows:

$$\tilde{h}_{i,LS} = h_i + \tilde{n_i}$$

ρ_i is known as the coefficient for the noise power estimation of each path with a purpose of fitting the real noise. The optimal value can be obtained by the following formula.

$$\partial\Bigg(\bigg|\frac{|\tilde{h}_{i,LS}|^2}{|\tilde{h}_{i,LS}|^2 + \rho_i\cdot\sigma^2} - \frac{|h_i|^2}{|h_i|^2 + |n_i|^2}\bigg|^2 \Bigg)\Big/\partial\rho_i = \tilde{0}.$$

Since the estimated noise N_i and n_i of $\tilde{h}_{i,LS}$ is the Gaussian white noise with approximately identical distribution, there is the following equation:

$$\left|\tilde{h}_{i,LS}\right|^2 = \left|h_i\right|^2 + \left|n_i\right|^2 + 2 \cdot \mathrm{Re}(h_i \cdot n_i{}^*)$$

$$= \left|n_i\right|^2 \left(1 + SNR + 2 \cdot \mathrm{Re}(\frac{h_i}{\left|h_i\right|} \cdot \frac{n_i{}^*}{\left|n_i\right|}) \cdot \sqrt{SNR}\right)$$

In practice, $\left|h_i\right|^2 / \left(\left|h_i\right|^2 + \left|n_i\right|^2\right)$ is used to filter each time-domain point. Therefore, the following fitting equation can be obtained:

$$\partial \left(E \left| \frac{\left|n_i\right|^2\left(1+SNR+2\cdot\mathrm{Re}(\frac{h_i}{|h_i|}\cdot\frac{n_i{}^*}{|n_i|})\cdot\sqrt{SNR}\right)}{\left|n_i\right|^2\left(1+SNR+2\cdot\mathrm{Re}(\frac{h_i}{|h_i|}\cdot\frac{n_i{}^*}{|n_i|})\cdot\sqrt{SNR}\right)+\rho\sigma^2} - \frac{\left|h_i\right|^2}{\left|h_i\right|^2+\left|n_i\right|^2}\right|^2 \right) \Big/ \partial\rho = 0$$

It is presumed that the estimated noise power equals to the actual noise power, i.e., $\sigma^2 = \left|n_i\right|^2$. Besides, the following two equations are considered:

$$\frac{\left|h_i\right|^2}{\left|h_i\right|^2+\left|n_i\right|^2} = \frac{\left|n_i\right|^2 SNR}{\left|n_i\right|^2 \cdot (SNR+1)}$$

$$\mathrm{Re}(\frac{h_i}{\left|h_i\right|} \cdot \frac{n_i{}^*}{\left|n_i\right|}) \approx \mathrm{Re}\left(e^{j\theta}\right) = \cos\theta, \ f(\theta) = \frac{1}{\pi} \ \theta \in [0,\pi]$$

The second one is an approximated equation, thus the following equation can be obtained after simplification.

$$\left(2\left((1+q+\rho)^2-4q\right)-(q+1)\cdot\rho\right)-2(q+2)\cdot\rho(q+1+\rho)+\frac{3(q+1)\cdot\rho^2\cdot(q+1+\rho)^2}{(1+q+\rho)^2-4q} = 0$$

Where $q = SNR$. The following table can be obtained through numerical calculation.

Table 1. Optimal at different SNRs

SNR(dB)	0	2	4	6	8	10	12	14
ρ	2.0000	1.0279	0.5213	0.5842	0.6838	0.7727	0.8428	0.8943
SNR(dB)	16	18	20	22	24	26	28	30
ρ	0.9304	0.9548	0.9710	0.9815	0.9882	0.9925	0.9953	0.9970

It is worth noting that the SNR in the above formula is somewhat different from that in the real system, i.e., it refers to the ratio of the power at each path to the noise power, not for all channels. In practice, however, one same noise estimation coefficient ρ is used for filtering estimation in consideration of the feasibility of the system.

4   B.Noise Power Estimation

In the channel estimation calculation based on instantaneous power filtering, the noise power needs to be estimated. For an ECMA368 system, three methods for noise power estimation were developed. In the first, the estimation is performed based on the noise power on null subcarriers. In the standard mode, there are totally six channel estimation sequences. Each sequence involves six null subcarriers. Thus, the power estimation is as follows:

$$\sigma^2 = \sum_{i=1}^{6}\left(\left|Y(i,0)\right|^2 + \sum_{j=63}^{67}\left|Y(i,j)\right|^2\right)\Big/(36*128)$$

The second is based on the noise power on all subcarriers, i.e., the independency of the noise on the same subcarrier of six OFDMs is used for estimation.

$$1/128*1/128*\left[\frac{1}{6}\sum_{i=1}^{6}\sum_{j=1}^{128}\left|\mathbf{Y}(i,j)\right|^2 - \sum_{j=1}^{128}\left|\frac{1}{6}\sum_{i=1}^{6}\mathbf{Y}(i,j)\right|^2\right] = \frac{5}{6}\sigma^2$$

The third is based on time-domain filtering, i.e., the channel estimation sequences after interpolation are converted to the time domain. Since the noise components are distributed at points 38~128, the noise power can be estimated by the following equation.

$$\sigma^2 = \frac{1}{6\times91}\sum_{j=1}^{6}\sum_{i=38}^{128}\left|e_i^T\mathbf{h}_{j,fnval}\right|^2$$

Among the three estimation methods, the first is the fastest, while the second is the most accurate. In order to further improve the precision of the noise power estimation, synchronous sequences can be employed.

5   Analysis on Simulation Results

5.1   A.Simulation of Noise Power Estimation Algorithm

The channel models under two extreme conditions (CM1 and CM4) were chosen for simulation. Through comparison of simulation results with the real noise, the performance curves in Fig. 2 were obtained.
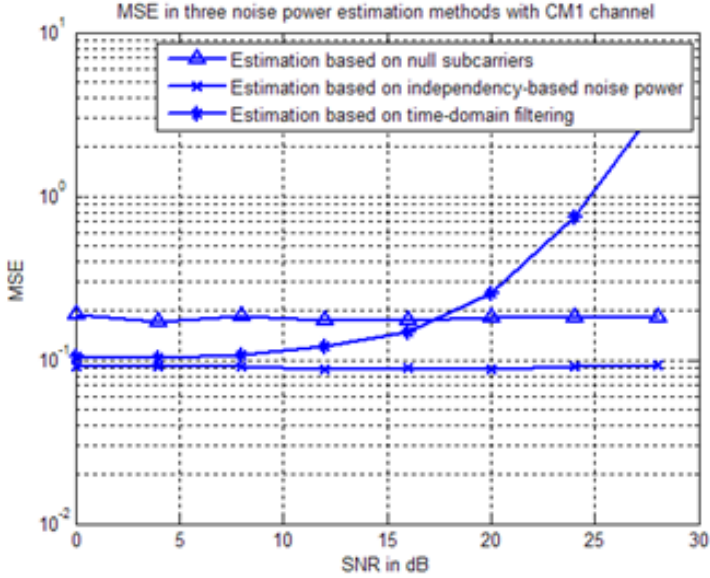
(a)CM1 Channel
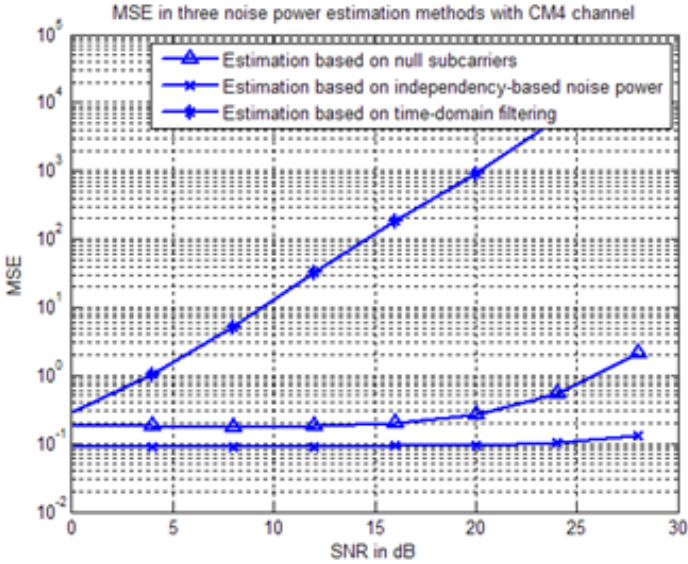
(b) CM4 Channel

Fig. 2. Performance of noise power estimation

From the above curves, it can be seen that the independency-based noise power estimation method achieves the optimal performance, while the estimation method based on null subcarriers exhibits rather stable performance. Both methods are based on the frequency domain, while the error of the noise power estimation method based on time-domain filtering increases gradually with the increase of SNR. The reason is that the frequency-domain interpolation is required in the first place in the method based on time-domain filtering, and the interpolation may cause error diffusion and noise correlation. At high SNR, such error will grow larger and larger compared with minute noise, thus MSE may grow increasingly. However, this method is very suitable for the UWB environment with short channel delay owning to its easy implementation. When the system SNR is high, all kinds of algorithms themselves rather than the noise estimation precision are the decisive factors affecting the system performance, therefore such estimation error can be ignored. With CM4 channels, the method based on time-domain filtering has considerable error due to the large length of channel response. In practice, it can be substituted by the method based on null subcarriers.

5.2   B.Simulation of Channel Estimation Algorithm Based on Instantaneous Power Filtering

With channels CM1 and CM4, $\rho = 1/3, 1/2, 1, 2, 3$ were adopted for estimation respectively. Through comparison with the conventional LS algorithm and the improved LS algorithm based on time-domain interpolation filtering, the curves shown in Fig. 3 were obtained，where T-Noise and N-Noise represents noise estimation based on time-domain filtering and on null subcarriers respectively. The channel time response   usually not exceed 37 sample points oaECMA368/369 standard in channel CM1, therefore instantaneous power filtering is only for the former 37 point.
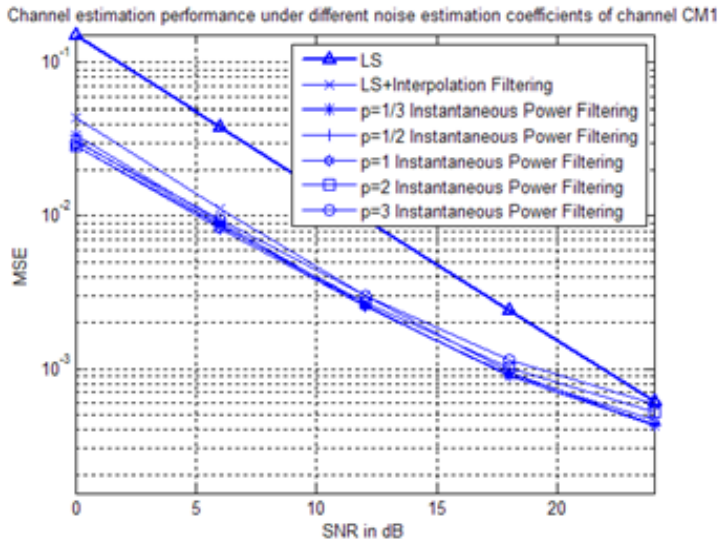
(a)   CM1

Fig. 3. Performance of instantaneous power filtering algorithm for 37 points with channel CM1

Channel estimation performance under different noise estimation coefficients of channel CM4
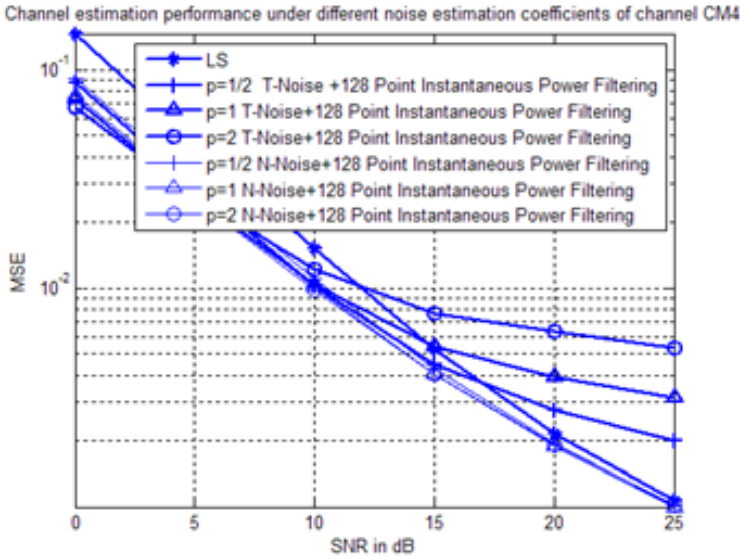
(b) CM4

Fig. 3. (*Continued.*)

In channel CM1, it can be seen that the above curves are very close to the results in Table 2, the reason for which is that the varying trends of the channel power at each path and the signal power at the receiving terminal are progressively However, they are somewhat different from the data in the table because of the different channel power at each path, the error existing in noise power estimation, the signal correlation after interpolation and the averaging operation on the six training sequences at the receiving terminal. It can be seen that the performance of the instantaneous power filtering within the range of 0~16dB is increased by 7dB and 2dB approximately in comparison with the conventional LS algorithm and the one improved through interpolation filtering. It is also shown that the performance is good at low SNR $\rho = 2 \sim 3$ and even better at high SNR $\rho \approx 1$, therefore the value of ρ can be set according to the simulation results.

In view of the large response length of channel CM4, the latter 91 points were not abandoned after the instantaneous power filtering was only performed on the first 37 protective spacing points. In this study, the instantaneous power filtering was performed on all 128 estimation points. Because noise power estimation method based on the time-domain filtering had great error with channel CM4, it could be seen that poor performance was achieved with the time-domain filtering noise power estimation plus the instantaneous power filtering method, while the method based on null subcarriers exhibited good robustness and consequently had rather stable performance with channel CM4.

6   Conclusion

Simulation results showed that the performance of channel estimation algorithm based on instantaneous power filtering was superior to that of the conventional LS algorithm and was robust against the synchronization error, being suitable for the hardware implementation of a baseband system.

References

1. Barrett, T.W.: History of Ultra-wideband (UWB) Radar and Communications: Pioneers and Innovators
2. FCC, Revision of Part 15 of the Commission's Rules Regarding Ultra-Wideband Transmission Systems [EB/OJ] (2002)
3. IEEE P802.15 Working Group for Wireless Personal Area Networks (WPANs), Channel Modeling Sub-committee Report Final, November 18, (2002)
4. ECMA International, Standard ECMA-368-High Rate Ultra Wideband PHY and MAC Standard, Rue du Rhone 114 CH-1204 Geneva, (December 2005)
5. Muquet, B., Wang, Z., Giannakis, G.B., de Courville, M., Duhamel, P.: Cyclic prefixing or zero padding for wireless multicarrier transmissions. IEEE Trans. Commun. 50(12), 2136–2148 (2002)
6. Edfors, O., Sandell, M., van de Beek, J.J., Wilson, S.K., Borjesson, P.O.: OFDM channel estimation by singular value decomposition. IEEE Trans. Communication 46(7), 931–939 (1998)
7. Salberg, A.-B., Swami, A.: Doppler and frequency-offset synchronization in wideband OFDM. IEEE Trans. Wireless Commun. 4(6), 2870–2881 (2005)
8. Kapoglu, I., Li, Y., Swami, A.: Effect of Doppler Spread in OFDM-based UWB systems. IEEE Trans. Wireless Commun. 4(6), 2599–2567 (2005)
9. Stojanovic, M., Catipovic, J., Proakis, J.: Reduced-complexity multichannel processing of underwater acoustic communication signals. J. Acoust. Soc. Am. 98(2), pt. 1, 961–972 (1995)
10. Frassati, et al.: Experimental assessment of OFDM and DSSS modulations for use in littoral waters underwater acoustic communications. In: Proc. IEEE Oceans 2005 Europe Conf. (June 2005)
11. Kim, B.-C., Lu, I.-T.: Parameter study of OFDM Underwater Communications System. In: Proc. IEEE Oceans 2000 Conf. (September 2000)
12. Li, Y., Cimini, L., Sollenberger, N.: Robust channel estimation for OFDM systems with rapid dispersive fading channels. IEEE Trans. Commun. 46(7), 902–915 (1998)
13. Muquet, B., Wang, Z., Giannakis, G.: Cyclic prefix or zero padding for wireless multicarrier transmissions? IEEE Trans. Commun. 50(12), 2136–2148 (2002)

Smooth-Threshold GEE Variable Selection Based on Quadratic Inference Functions with Longitudinal Data

Ruiqin Tian and Liugen Xue

College of Applied Sciences,
Beijing University of Technology,
100 Pingleyuan, Chaoyang District, Beijing, 100124, P.R. China
tianruiqin@emails.bjut.edu.cn, lgxue@bjut.edu.cn

Abstract. A variable selection procedure is proposed using smooth-threshold generalized estimating equations based on quadratic inference functions (SGEE-QIF). The proposed procedure automatically eliminates inactive predictors by setting the corresponding parameters to be zero, and simultaneously estimates the nonzero regression coefficients by solving the SGEE-QIF. The proposed procedure avoids the convex optimization problem and is flexible and easy to implement. We establish the consistency and asymptotic normality of the resulting estimators. Extensive Monte Carlo simulation studies are conducted to examine the finite sample performance of the proposed variable selection procedure.

Keywords: Quadratic inference functions, Variable selection, Longitudinal data, Generalized estimating equations.

1   Introduction

Longitudinal data modeling is a statistical method often used in the analysis of experiments that are designed so that responses on the same experimental units are observed at each repetition [1-3]. Experiments of this type have broad applications, especially in the life and social sciences. Therefore, we consider a longitudinal study with n subjects and n_i observations over time for the i th subject ($i = 1, \cdots, n$)

for a total of $N = \sum_{i=1}^{n} n_i$ observation. Each observation consists of a response

variable Y_{ij} and a covariate vector $X_{ij} \in R^p$ taken from the i th subject at time t_{ij}. We assume that the full data set $\{(X_{ij}, Y_{ij}), \ i = 1, \cdots, n, \ j = 1, \cdots, n_i\}$ is observed and can be modelled as

$$Y_{ij} = X_{ij}^T \beta + \varepsilon_{ij}, \ i = 1, \cdots, n, \ j = 1, \cdots, n_i, \tag{1}$$

where β is a $p \times 1$ vector of unknown regression coefficients, ε_{ij} is random error with $E(\varepsilon_{ij} \mid X_{ij}) = 0$. In addition, we give assumptions on the first two

moments   of   the   observations $\{Y_{ij}\}$,   that   is, $E(Y_{ij}) = X_{ij}^T \beta$ and $Var(Y_{ij}) = v(X_{ij}^T \beta)$, where $v(\cdot)$ is a known variance function.

As is well known, variable selection is an important topic in all regression analysis and most of the variable selection procedures are based on penalized estimation using penalty functions, such as, L_q penalty (Frank and Friedman [4]), Lasso penalty (Tibshirani [10]), SCAD penalty (Fan and Li [3]), the Adaptive Lasso (Zou [17]), and so on. In most of current variable selection literature (Xu and Zhang [12], Xu et al. [13], Xue et al. [14], Zhao and Xue [15]) are based on penalized estimation using penalty functions, which have a singularity at zero. Consequently, these procedures require convex optimization, which incurs a computational burden. To overcome this drawback, Ueki [11] developed a new variable selection procedure called the smooth-threshold estimating equations that can automatically eliminate irrelevant parameters by setting them as zero. Lai et al. [6] explored GEE estimation and smooth-threshold GEE variable selection for single-index models with clustered data.

Generalized estimating equations (GEE) enable one to estimate regression parameters consistently in longitudinal data analysis even when the correlation structure is misspecified. However, when the working correlation is misspecified, the moment estimator of the nuisance parameter α suggested by Liang and Zeger [7] no longer results in the optimal estimator of β . Furthermore, their estimator of α does not exist in some simple case of misspecification (Crowder [1]). For this reason, Qu et al. [8] proposed a method of quadratic inference functions (QIF). It avoids estimating the nuisance correlation structure parameters by assuming that the inverse of working correlation matrix can be approximated by a linear combination of several known basis matrices. The QIF can efficiently take the within-cluster correlation into account and is more efficient than the GEE approach when the working correlation is misspecified. The QIF estimator is also more robust against contamination when there are outlying observations (Qu and Song [9]).

In this paper, we use the smooth-threshold generalized estimating equations based on quadratic inference functions to the linear models with longitudinal data. The proposed procedure automatically eliminates the irrelevant parameters by setting them as zero, and simultaneously estimates the nonzero regression coefficients by solving the SGEE-QIF. Compared to the shrinkage methods and the existing research findings reviewed above, our method offers the following improvements: (i) the proposed procedure avoids the convex optimization problem; (ii) the proposed SGEE-QIF approach is flexible and easy to implement; (iii) the proposed method is easy to deal with the longitudinal correlation structure.

The rest of this paper is organized as follows. In Section 2 we first propose the SGEE-QIF variable selection procedure for the linear model with longitudinal data. Asymptotic properties of the resulting estimators are considered in Section 3. In Section 4 we give the computation of the estimators as well as the choice of the tuning parameters. In Section 5 we carry out simulation studies to assess the finite sample performance of the method.

2   Methodology

Throughout this paper, let β_0 be the fixed true value of β and let $n \to \infty$ while n_i are uniformly bounded. We partition β_0 into active (nonzero) and inactive (zero) coefficients as follows: let $A_0 = \{j : \beta_{0j} \neq 0\}$ and $A_0^c = \{j : \beta_{0j} = 0\}$.

2.1   Quadratic Inference Functions

Denote $Y_i = (Y_{i1}, \cdots, Y_{in_i})^T$, and write X_i and ε_i in a similar fashion. Following the GEE method from Liang and Zeger [7], we can construct the estimating equations for β

$$\sum_{i=1}^{n} X_i^T V_i^{-1}(Y_i - X_i\beta) = 0, \tag{1}$$

where V_i is the covariance matrix of Y_i. Following Liang and Zeger [7], we simplify the covariance of the i th subject V_i by taking $V_i = A_i^{1/2} R(\alpha) A_i^{1/2}$, where $A_i = \mathrm{diag}(Var(Y_{i1}), \cdots, Var(Y_{in_i}))$, $R(\alpha)$ is a common working correlation with a nuisance parameters α. Based on the estimation theory associated with the working correlation structure, the GEE estimator of the regression coefficient proposed by Liang and Zeger [7] is consistent if consistent estimators of the nuisance parameters α can be obtained. For suggested methods for estimating α, see Liang and Zeger [7]. However, even in some simple cases, consistent estimators of α do not always exist (Crowder [1]). To avoid this drawback, Qu et al. [8] introduced the quadratic inference functions by assuming that the inverse of the working correlation matrix $R^{-1}(\alpha)$ can be approximated by a linear combination of a class of basis matrices as

$$\sum_{k=1}^{s} a_k M_k, \tag{2}$$

Where $M_1, \cdots, M_s$ are known matrices, and $a_1, \cdots, a_s$ are unknown constants This is a sufficiently rich class that accommodates, or at least approximates, the correlation structures most commonly used. Details of utilizing a linear combination of some basic matrices to model the inverse of working correlation can be found in Qu et al. [8].

Substituting (2) to (1), we get the following class of estimating functions:

$$\sum_{i=1}^{n} X_i^T A_i^{-1/2} (a_1 M_1 + \cdots + a_s M_s) A_i^{-1/2} (Y_i - X_i \beta) = 0. \tag{3}$$

Unlike the GEE method, we need not find the estimates of parameters $a = (a_1, \cdots, a_s)^T$ by optimizing some function of the information matrix associated with (3). Instead, based on the form of the quasi-score, we define the "extend score" g_n to be

$$g_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} g_i(\beta) = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} X_i^T A_i^{-1/2} M_1 A_i^{-1/2} (Y_i - X_i \beta) \\ \vdots \\ X_i^T A_i^{-1/2} M_s A_i^{-1/2} (Y_i - X_i \beta) \end{pmatrix}. \tag{4}$$

The estimating equations (3) are linear combinations of elements of the extend score vector (4). However, it does not work because the dimension of $g_n(\beta)$ is obviously greater than the number of unknown parameters. Using the idea of generalized method of moments (Hansen [5]), Qu et al. [8] defined the quadratic inference functions,

$$Q_n(\beta) = n g_n^T(\beta) \Omega_n^{-1}(\beta) g_n(\beta), \tag{5}$$

where

$$\Omega_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} g_i(\beta) g_i^T(\beta).$$

Note that Ω_n depends on β. The QIF estimate $\hat{\beta}_n$ is then given by

$$\hat{\beta}_n = \arg\min_{\beta} Q_n(\beta).$$

Then, based on (5), the corresponding estimating equation for β is

$$U(\beta) = n \dot{g}_n^T \Omega_n^{-1} g_n, \tag{6}$$

where $\dot{g}_n$ is the $sp \times p$ matrix $\{\partial g_n / \partial \beta^T\}$.

2.2   Smooth-Threshold Generalized Estimating Equations Based on QIF

Ueki [11] developed a automatic variable selection procedure that can automatically eliminate irrelevant parameters by setting them as zero. The method is easily implemented without solving any convex optimization problems. Motivated by this idea we propose the following smooth-threshold generalized estimating equations based on quadratic inference functions (SGEE-QIF)

$$(I_p - \Delta)U(\beta) + \Delta\beta = 0, \tag{2}$$

where Δ is the diagonal matrix whose diagonal elements are $\delta = (\delta_j)_{j=1,\cdots,p}$, and I_p is the p-dimensional identity matrix. Note that the jth SGEE-QIF with $\delta_j = 1$ reduces to $\beta_j = 0$. Therefore, SGEE-QIF (7) can yield a sparse solution. Unfortunately, we cannot directly obtain the estimator of β by solving (7). This is because the SGEE-QIF involves δ_j, which need to be chosen using some data-driven criteria. For the choice of $\delta = (\delta_j)_{j=1,\cdots,p}$, Ueki [11] suggested that δ_j may be determined by the data, and can be chosen by $\hat{\delta}_j = \min(1, \lambda / |\hat{\beta}_j^{(0)}|^{1+\gamma})$ with an initial estimator $\hat{\beta}_j^{(0)}$. The initial estimator $\hat{\beta}_j^{(0)}$ can be obtained by solving the QIF (5) for the full model. Note that this choice involves two tuning parameters (λ, γ). In Section 4, following the idea of Ueki [11], we use the BIC-type criterion to select the tuning parameters.

Replacing Δ in (7) by $\hat{\Delta}$ with diagonal elements $\hat{\delta} = (\hat{\delta}_j)_{j=1,\cdots,p}$. The SGEE-QIF becomes

$$(I_p - \hat{\Delta})U(\beta) + \hat{\Delta}\beta = 0. \tag{3}$$

The solution of (8) denoted by $\hat{\beta}_{\lambda,\gamma}$ is called the SGEE-QIF estimator.

3   Asymptotic Properties

We define the active set $A = \{j : \hat{\delta}_j \neq 1\}$ which is the set of indices of nonzero parameters, where $\hat{\delta}_j = \min(1, \lambda / |\hat{\beta}_j^{(0)}|^{1+\gamma})$. Next, we study the asymptotic properties of the SGEE-QIF estimators. In our developments we make the following assumptions:

(C1). The parameter space S is compact, and β_0 is an interior point of S.

(C2). $E\varepsilon_i^4 < \infty$, $EX_{ij(r)} < \infty$, $i = 1,\cdots,n$, $j = 1,\cdots,n_i$, $r = 1,\cdots,p$, where $EX_{ij(r)}$ is the rth component of X_{ij}.

$$\text{(C3).} \quad \lim_{n\to\infty} \frac{1}{n} E \begin{pmatrix} X_{i,\mathrm{A}_0}^{T} A_i^{-1/2} M_1 A_i^{-1/2} X_{i,\mathrm{A}_0} \\ \vdots \\ X_{i,\mathrm{A}_0}^{T} A_i^{-1/2} M_s A_i^{-1/2} X_{i,\mathrm{A}_0} \end{pmatrix} \equiv J_0 .$$

(C4). The weighting matrix $\Omega_n(\beta)$ converges almost surely to a constant matrix Ω_0, where Ω_0 is invertible. Furthermore, the first and second partial derivatives of Ω_n in β are all $O_p(1)$.

(C5). $Q_n(\beta)$ is a twice differentiable function of β. Furthermore the third derivatives of $Q_n(\beta)$ are $O_p(n)$.

We assume, under the regularity conditions, the initial estimator using the full model is consistent and asymptotically normally distributed by solving the QIF. The following theorems give the consistency of the SGEE-QIF estimators, variable selection consistency and asymptotic normality.

Theorem 1. ($\sqrt{n}$-consistency) For any positive λ and γ such that $n^{1/2}\gamma \to 0$ and $n^{(1+\gamma)/2}\lambda \to \infty$ as $n \to \infty$, and the conditions (C1)-(C5) hold, there exists a sequence $\hat{\beta}_{\lambda,\gamma}$ of the solutions of (8) such that $\left\| \hat{\beta}_{\lambda,\gamma} - \beta_0 \right\| = O_p(n^{-1/2})$.

Theorem 2. Suppose that the conditions of Theorem 1 hold, as $n \to \infty$, we have

(i)(Variable selection consistency), i.e. $P(\mathrm{A} = \mathrm{A}_0) \to 1$;

(ii) (Asymptotic normality), i.e. $\sqrt{n}(\hat{\beta}_{\lambda,\gamma,\mathrm{A}} - \beta_{\mathrm{A}_0}) \xrightarrow{L} N(0, \Phi)$,

where $\Phi = (J_0^T \Omega_0^{-1}(\beta_{\mathrm{A}_0}) J_0)^{-1}$, and " $\xrightarrow{L}$ " represents the convergence in distribution.

Remark 1: Theorems 1 and 2 imply that the proposed SGEE-QIF procedure is consistent in variable selection, it can identify the zero coefficients with probability tending to 1. By choosing appropriate tuning parameters, the SGEE-QIF estimators have the oracle property; that is, the asymptotic variances for the SGEE-QIF estimators are the same as what we would have if we knew in advance the correct submodel.

Remark 2: Proofs of the part (i) in Theorem 2 is similar to that of Ueki [11]. By this, Theorem 1 and the part (ii) in Theorem 2 is essentially along the same line as the proof of Theorems 4.1-4.2 in Dziak [2]. This will lead to the conclusions. To save space, the proofs are omitted.

4   Some Issues in Practice

In the practical situation, we have to choose the basis matrices for the inverse of the working correlation structure, and choose the proper tuning parameter (λ, γ). In this section, we discuss these practical issues.

4.1   Choosing the Basis Matrices

The choice of basis matrices M_k in (2) is not difficult, especially for those special correlation structures which are frequently used. If we assume $R(\alpha)$ is the first-order autoregressive correlation matrix. The exact inversion $R^{-1}(\alpha)$ can be written as a linear combination of three basis matrices, they are M_1, M_2 and M_3, where M_1 is the identity matrix, M_2 has 1 on two main off-diagonals and 0 elsewhere, and M_3 has 1 on the corners (1, 1) and (m, m), and 0 elsewhere. Suppose $R(\alpha)$ is an exchangeable working correlation matrix, it has 1 on the diagonal, and α everywhere off the diagonal. Then $R^{-1}(\alpha)$ can be written as a linear combination of two basis matrices, M_1 is the identity matrix and M_2 is a matrix with 0 on the diagonal and 1 off the diagonal. More details about choosing the basis matrices can be seen in Zhou and Qu [16].

4.2   Selection of Tuning Parameters

To implement the procedures described above, we need to choose the tuning parameters (λ, γ). Following Ueki [11], we use BIC-type criterion to choose these two parameters. That is, we choose (λ, γ) as the minimizer of

$$BIC_{\lambda,\gamma} = Q_n(\hat{\beta}_{\lambda,\gamma}) + df_{\lambda,\gamma} \log(n),$$

where $\hat{\beta}_{\lambda,\gamma}$ is the SGEE-QIF estimator for given (λ, γ), $df_{\lambda,\gamma}$ is simply the number of nonzero coefficient $\hat{\beta}_{\lambda,\gamma}$. The selected (λ, γ) minimizes the $BIC_{\lambda,\gamma}$.

4.3   Iterative Algorithm

Next, we propose the iterative algorithm to implement the procedures as follows:

Step 1, Calculate the initial estimates $\hat{\beta}^{(0)}$ of β by solving the initial QIF (5) estimator. Let $k = 0$.

Step 2, By using the current estimate $\hat{\beta}^{(k)}$, we choose the tuning parameters (λ, γ) by the BIC criterion.

Step 3, Update the estimator of β as follows:

$$\hat{\beta}_A^{(k+1)} = \hat{\beta}_A^{(k)} - \{ n\dot{g}_{n,A}^T(\hat{\beta}_A^{(k)})\Omega_n^{-1}(\hat{\beta}_A^{(k)})\dot{g}_{n,A}(\hat{\beta}_A^{(k)}) + \hat{G}_A \}^{-1}$$

$$\times \{ \dot{g}_{n,A}^T(\hat{\beta}_A^{(k)})\Omega_n^{-1}(\hat{\beta}_A^{(k)})g_{n,A}(\hat{\beta}_A^{(k)}) + \hat{G}_A\hat{\beta}_A^{(k)} \},$$

$$\hat{\beta}_{A^c,\hat{\Delta}} = 0,$$

where $g_{n,A}(\beta_A) = \dfrac{1}{n}\sum\limits_{i=1}^n \begin{pmatrix} X_{i,A}^T A_i^{-1/2} M_1 A_i^{-1/2}(Y_i - X_{i,A}\beta_A) \\ \vdots \\ X_{i,A}^T A_i^{-1/2} M_s A_i^{-1/2}(Y_i - X_{i,A}\beta_A) \end{pmatrix},$

and $\hat{G}_A = (I_{|A|} - \hat{\Delta}_A)^{-1}\hat{\Delta}_A$.

Step 4, Iterate Steps 2-3 until convergence, and denote the final estimators of β as the SGEE-QIF estimator.

5   Simulation Study

In this section we conduct a simulation study to assess the finite sample performance of the proposed procedures. In the simulation study, the performance of estimator $\hat{\beta}$ will be assessed by using the average the mean square error (AMSE), defined as $\left\| \hat{\beta} - \beta_0 \right\|^2$ averaged over 1000 times simulated data sets.

We simulate data from model (1.1), where $\beta_0 = (\beta_1, \cdots, \beta_{13})^T$ with $\beta_1 = 2.8$, $\beta_2 = -1.8$ and $\beta_3 = 3.8$. While the remaining coefficients, corresponding to the irrelevant variables, are given by zeros. To perform this simulation, we take the covariates $X_{ij}(j = 1, \cdots, 5)$ from a multivariate normal distribution with mean zero, marginal variance 1 and correlation 0.5. Response variable Y_{ij} is generated according to the model. And error vector

$\varepsilon_i = (\varepsilon_{i1}, \cdots, \varepsilon_{i5})^T \sim N(0, \sigma^2 Corr(\varepsilon_i, \alpha))$, where $\sigma^2 = 1$ and $Corr(\varepsilon_i, \alpha)$ is a known correlation matrix with parameter α used to determine the strength of with-subject dependence. Here we consider ε_{ij} has the first-order autoregressive (AR-1) or compound symmetry (CS) correlation (i.e. exchangeable correlation) structure with $\alpha = 0.7$. In the following simulations, we make 1000 simulation runs and take $n = 50$ and 100.

In the simulation study, for each simulated data set, we compare the estimation accuracy and model selection properties of the SGEE-QIF method, the SCAD-penalized QIF and the Lasso-penalized QIF for two different working correlations. For each of these methods, the average of zero coefficients over the 1000 simulated

data sets is reported in Tables 1-2. Notes that "Correct" in Tables means the average number of zero regression coefficients that are correctly estimated as zero, and "Incorrect" depicts the average number of non-zero regression coefficients that are erroneously set to zero. At the same time, we also examine the effect of using a misspecified correlation structure in the model, which are also reported in Tables 1-2.

From Tables 1-2, we can make the following observations.

(i) For the parametric component, the performances of variable selection procedures become better and better as n increase. For example, the values in the column labeled "Correct" become more and more closer to the true number of zero regression coefficients in the models.

(ii) Compared with the penalized QIF based on Lasso and SCAD, SGEE-QIF performs satisfactory in terms of variable selection.

(iii) It is not surprised that the performances of variable selection procedures based on the correct correlation structure work better than based on the incorrect correlation structure. However, we also note that the performance does not significantly depend on working covariance structure.

Table 1. Variable selections for the parametric components under different methods when the correlation structure is correctly specified

		CS			AR(1)		
n	Method	AMSE	Correct	Incorrect	AMSE	Correct	Incorrect
50	SGEE	0.0067	10.0000	0	0.0088	10.0000	0
	SCAD	0.0071	9.8660	0	0.0090	9.8840	0
	Lasso	0.0077	9.8230	0	0.0095	9.8400	0
100	SGEE	0.0036	10.0000	0	0.0038	10.0000	0
	SCAD	0.0036	9.9960	0	0.0039	9.9990	0
	Lasso	0.0038	9.9950	0	0.0044	9.9970	0

Table 2. Variable selections for the parametric components under different methods when the correlation structure is incorrectly specified. The term "CS.AR(1)" means estimation with the fitted misspecified AR(1) correlation structure, while "AR(1).CS" means estimation with the fitted misspecified CS correlation structure.

		CS.AR(1)			AR(1).CS		
n	Method	AMSE	Correct	Incorrect	AMSE	Correct	Incorrect
50	SGEE	0.0092	10.0000	0	0.0115	10.0000	0
	SCAD	0.0096	9.7680	0	0.0119	9.8630	0
	Lasso	0.0104	9.6530	0	0.0125	9.7800	0
100	SGEE	0.0046	10.0000	0	0.0052	10.0000	0
	SCAD	0.0046	9.9920	0	0.0052	9.9960	0
	Lasso	0.0049	9.9840	0	0.0056	9.9920	0

References

1. Crowder, M.: On the use of a working correlation matrix in using generalised linear models for repeated measures. Biometrika 82, 407–410 (1995)
2. Dziak, J.J.: Penalized quadratic inference functions for variable selection in longitudinal research. Ph.D Thesis, The Pennsylvania State University, (2006), `http://www.stat.psu.edu/~jdziak`
3. Fan, J.Q., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360 (2001)
4. Frank, I.E., Friedman, J.H.: A statistical view of some chemometrics regression tools (with discussion). Technometrics 35, 109–148 (1993)
5. Hansen, L.: Large sample properties of generalized method of moments estimators. Econometrica 50, 1029–1054 (1982)
6. Lai, P., Wang, Q., Lian, H.: Bias-corrected GEE estimation and smooth-threshold GEE variable selection for single-index models with clustered data. J. Multivariate Anal. 105, 422–432 (2012)
7. Liang, K.L., Zeger, S.L.: Longitudinal data analysis using generalised estimating equations. Biometrika 73, 13–22 (1986)
8. Qu, A., Lindsay, B.G., Li, B.: Improving generalized estimating equations using quadratic inference functions. Biometrika 87, 823–836 (2000)
9. Qu, A., Song, P.X.K.: Assessing robustness of generalized estimating equations and quadratic inference functions. Biometrika 91, 447–459 (2004)
10. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. Roy. Statist. Soc., Ser. B 58, 267–288 (1996)
11. Ueki, M.: A note on automatic variable selection using smooth-threshold estimating equations. Biometrika 96(4), 1005–1011 (2009)
12. Xu, D., Zhang, Z.: Regularized REML for estimation in heteroscedastic regression models. In: Li, S., Wang, X., Okazaki, Y., Kawabe, J., Murofushi, T., Guan, L. (eds.) Nonlinear Mathematics for Uncertainty and its Applications. AISC, vol. 100, pp. 495–502. Springer, Heidelberg (2011)
13. Xu, D., Zhang, Z., Wu, L.: Variable selection in high-dimensional double generalized linear models. Stat. Papers (2012), doi:10.1007/s00362-012-0481-y
14. Xue, L., Qu, A., Zhou, J.J.: Consistent model selection for marginal generalized additive model for correlated data. J. Amer. Statist. Assoc. 105, 1518–1530 (2010)
15. Zhao, P.X., Xue, L.G.: Variable selection for semiparametric varying coefficient partially linear errors-in-variables models. J. Multivariate Anal. 101, 1872–1883 (2010)
16. Zhou, J., Qu, A.: Informative estimation and selection of correlation structure for longitudinal data. J. Amer. Statist. Assoc. 107, 701–710 (2012)
17. Zou, H.: The adaptive Lasso and its oracle properties. J. Amer. Statist. Assoc. 101, 1418–1429 (2006)

A Distributed Computing Platform
for Task Stream Processing

Xing Weiyan[1], Huang Wenqing[1], Liu Dong[2], and Deng Youyi[1]

[1] Department of Space Commanding, Academy of Equipment, Beijing 101416, China
[2] Key Laboratory, Academy of Equipment, Beijing 101416, China
{xingweiyan,huangwq,dengyy}@gmail.com, ld5m@163.com

Abstract. In order to process task stream, a distributed computing platform was put forward. The hardware of the platform includes one control terminal, multi computing nodes, one storage array and one fiber optic switch. The software of the platform includes DBMS, control software and tasks. The transaction process of the platform includes n ($n \geq 1$) tasks, and each task belongs to a specific level. There are data dependences among the tasks, namely that the output data of the previous task is the input data of the following task. The transaction starts from the first level task, and then to the second, the third, …, at last to the nth level task. After the nth level task finished and output result data, the transaction comes to the end. The platform is applicable to the science computing that is characterized by a task stream composed of multi tasks, such as remote sensing data processing and complex electromagnetism analysis.

Keywords: Distributed computing, Task stream, computing node, Platform.

1   Introduction

The fast development of computer networks improves the birth of distributed computing, which brings a reform in computer science and becomes a hot research field [1][2]. With the wide application of distributed computing, people begins to pursue high computing performance, good data transparency, and high reliability [3][4][5]. Distributed Computing Platform (DCP) is the integration of multi independent computers, which are connected by network and work together to complete a common task [6]. The properties of DCP include: (a) it is composed of many processors or computers; (b) computing resources can be processors, which are connected in a contiguous space by inner bus or switches and communicate each other by shared memory, or the geographically separated computer machines, which are connected by network and communicate each other by messages; (c) the platform is transparent to users, which means that the exact position of resources are unknown to users while the platform is executing a computing task; (d) program or computing can be parallel processed in multi computing resources.

Originally, DCP mainly concern the parallel computing of only one task. However, in practical applications, some applications are composed of several related tasks [7][8]. For example, remote sensing data processing includes 5 tasks, namely source

data unformating & dividing, coding, uncompressing, radiation correction and image mosaicing, and the input data of the following task is computing result of the previous task, which means there are data dependences among the tasks.

Besides, during the transaction processing, the fault in DCP, like computing node failure, is inevitable [9][10]. Furthermore, the scale of DCP is likely to increase or reduce in a long time running if a new node is added into DCP or a failure node is wiped off from DCP.

In order to process a task stream using a distributed computing way, the paper presents a DCP framework, in which the data dependences among the tasks are taken into account. The platform can also handle the events, such as node entering and failure, so as to support long-time incessant running.

The second section summarizes DCP. And the third section describes its software framework. The main contents of the paper are provided in the fourth section in detail. And the conclusions are made in the fifth section.

2   The Summarization of the Platform

The hardware of DCP includes one control terminal, multi computing nodes, one storage array and one fiber optic switch (see Fig. 1).
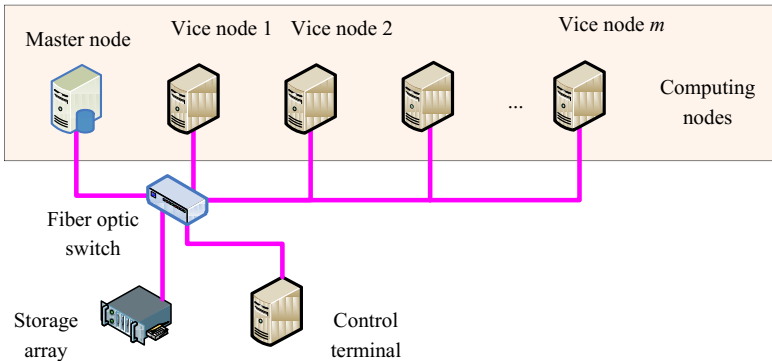
Fig. 1. The Hardware Structure of the DCP

The control terminal is used to setup the work mode and startup the working processes of the platform. The computing nodes are used to execute the computing. The storage array provides a reliable large capacity of storage environment, which is used to store computing results and database files. The fiber optic switch, which connects the control terminal, the computing nodes and the storage array by optical fiber network, is used to provide a high speed communication access for DCP.

DCP uses one computing node as master node, the other nodes as vice nodes. Master node, in which Distributed Computing Platform Software (DCPS) and Database Management System (DBMS) are installed, is in charge of scheduling tasks, executing computing and managing database files. In order to improve the reliability

of DCP, high reliable server is often chosen as the master node. Vice nodes, in which only DCPS is installed, are mainly in charge of executing computing with the control of master node.

The computing process of DCP, named transaction in the paper, includes n ($n \geq 1$) tasks, each of which belongs to one specific level. There are data dependences among the tasks, namely that the output data of the previous task is the input data of the following task. The transaction of DCP starts from the first level task, and then to the second, the third, …, at last to the nth level task. The source data of the first task can be the data files stored in storage array in advance or the real-time data stream from data output device. After the nth level task finished and output result data, the end of the transaction of DCP can be declared. Therefore, the transaction is actually a task stream made of n tasks. Fig. 2 illustrates a transaction example composed of four tasks.
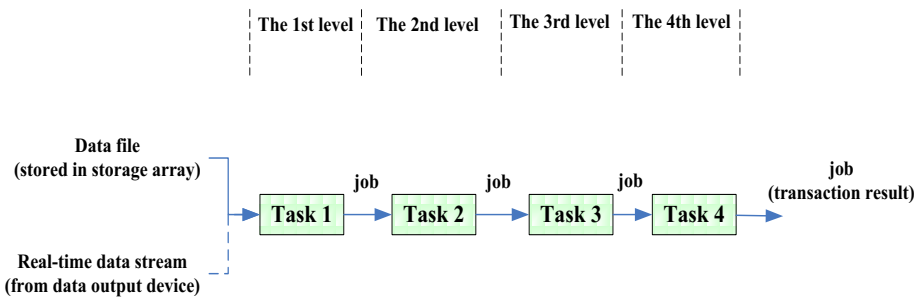
Fig. 2. A Transaction Example

After transaction starts, data flow among the tasks in DCP. If a vice node fails during the process, DCPS in master node will schedule the unfinished job in the failed node to another computing node. Besides, a new computing node can join in DCP to participate in the transaction at any time. Both the mentioned circumstances do not affect the natural computing processes of DCP.

3   The Software of the Platform

3.1   The Framework of Software

DCP supports many desktop operating systems, including Linux, UNIX and Windows. The software in DCP includes DBMS, DCPS and tasks.

Database files, which store configuration information, control information and task information, are saved in storage array.

DBMS is only installed in master node and in charge of maintaining, accessing and modifying database files.

DCPS includes platform management model (DCPS_PMM) and node management model (DCPS_NMM). DCPS_PMN is only installed in master node. And DCPS_NMM is installed in all of the computing nodes, including both master node and vice nodes.

Task is an executable application that DCP users designed and programmed. The main function of task is to process data. The work flow of a representative task is: receiving initial parameters from DCPS_NMM, processing source data, storing result data to storage array, and sending position information of result data to DCPS_NMM.

Transaction means the whole computing process of DCP.

The computing result of task, named as *job* in DCP, is stored in storage array, often in the form of data blocks. Job is the source data of a next level task or the transaction result of DCP if it is produced by the last level task.

3.2   The Structure of DCPS

1) DCPS_PMM

DCPS_PMM is with responsibility for scheduling task in DCP.

DCPS_PMM includes several sub-modules, which is shown in Fig. 3.
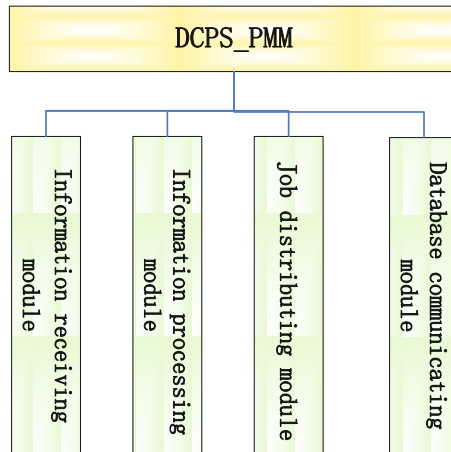
Fig. 3. The sub-modules of DCPS_PMM

The detail functions of the sub-modules are described below:

Information receiving module

After DCPS_PMM starts, information receiving module keeps waiting for the commands and the information from control terminal and other computing nodes.

Information processing module

The module processes the received information according to their types. Its functions include parsing commands, examining the returned result of commands, communicating with DCPS_NMM and issuing commands.

Job distributing module

The module distributes jobs to the DCPS_NMM in corresponding node according to the node load of the time.

Database communicating module

All the necessary processes, such as job distributing and task state updating, should be recorded in database using database communicating module.

2) DCPS_NMM

The main functions of DCPS_NMM are to monitor the local task operations and communicate with DCPS_PMM when necessary.

DCPS_NMM includes several sub-modules, , which is shown in Fig. 4.
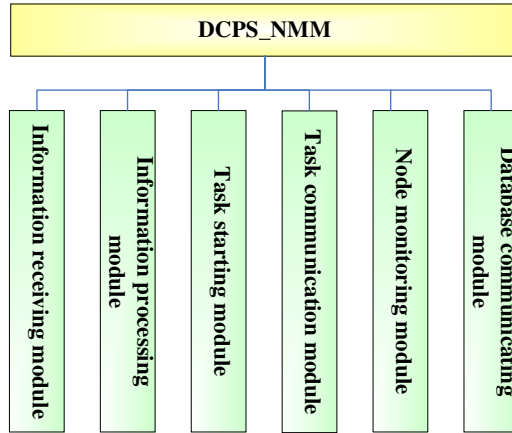
Fig. 4. The sub-modules of DCPS_NMM

The detail functions of the sub-modules are described below:

Information receiving module

After DCPS_NMM starts, it will keep waiting for the commands from DCPS_PMM and the result information from tasks.

Information processing module

The module processes the received information according to their types. Its functions include parsing commands, examining the returned result of commands and send task states to DCPS_PMM when a task starts or finishes.

Task starting module

The module only sees to starting the tasks in computing nodes according to the commands and configuration parameters from DCPS_PMM.

Task communication module

The module sends and receives the job information to and from tasks.

Node monitoring module

The module records a heartbeat signal of the local computing node periodically in database.

Database communicating module

The module is in charge of communicating with DBMS so as to update the state of jobs and tasks.

4   The Work Processes of DCP

4.1   Transaction Configuring

Before we use DCP, we should setup the transaction and input some necessary information to database. The information includes task name, task level, task parameter, task program position and task result position. All the transaction information will be stored in database, and each task corresponds to one record. Database can be used to store more than one transaction so as to support different application fields.

4.2   DCP Initializing

Before executing transaction, DCP should be initialized. The initializing work includes the following steps:

  1)   Master node startups DBMS.
  2)   Each computing node startups DCPS. If DCPS detects the existing of DBMS in the local node, both DCPS_PMM and DCPS_NMM will be activated. If DCPS does not detect the existing of DBMS, only DCPS_NMM will be activated.
  3)   The node monitoring module of DCPS_NMM writes a heartbeat signal periodically to database so as to indicate a good node state. If DBMS detects that the state of a certain node is not updated after a heartbeat period (normally 2 second), it will clean the record in database corresponding to the node so as to indicate the failure of the node.

4.3   Transaction Initializing

The initialization of transaction includes the following steps:

  1)   Control terminal sends starting transaction command, which only includes the transaction ID, to DCPS_PMM.
  2)   After the information receiving module of DCPS_PMM received the starting transaction command, the information processing module will create and write the task flow to database.
  3)   The database communication module in DCPS_PMM gets node ID addresses from database and sends configuration command to the corresponding computing nodes. The command includes the task address, name, initial parameters and result data position.
  4)   After the information receiving module in DCPS_NMM of each computing node received the configuration command, the task starting module startup the tasks according to the contents in the command.
  5)   After the above mentioned processes, all the tasks except the first level task in DCP will be in waiting state since all the non-first level tasks do not have input source data.

4.4   Task Scheduling Algorithm

The task scheduling algorithm of DCP is shown in Fig. 5:
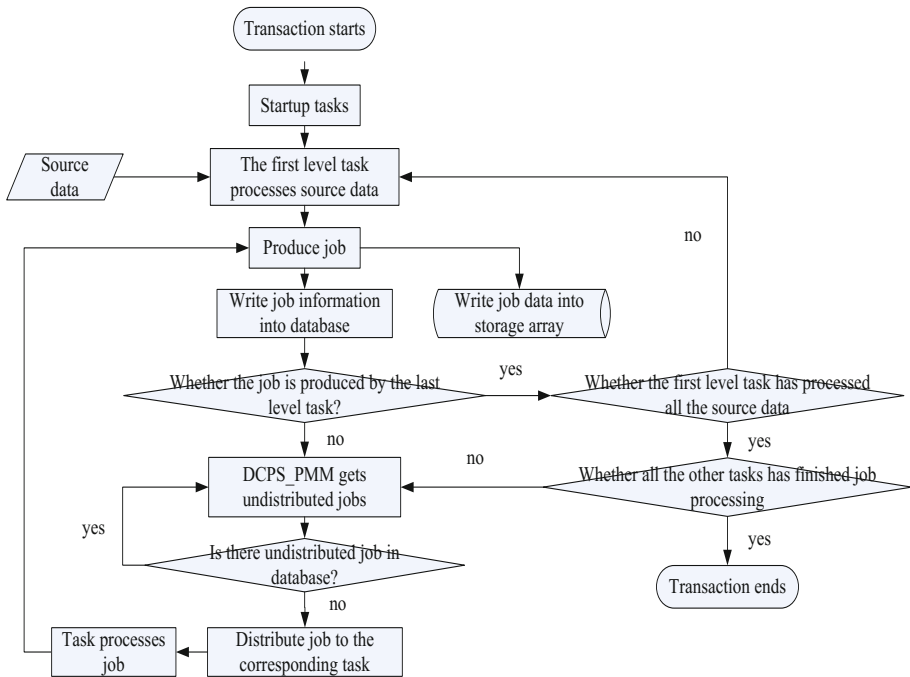
Fig. 5. The Workflow of Task Scheduling Algorithm

1) After the first level task started, it processes the source data acquired through fiber optic network or from storage array on its own. Once the task finished data processing, it stores the result data (namely job) to storage array according to the position parameter in configuration command and sends a job produced command (including task name, job file name, job position and job create time) to DCPS_NMM. The information processing module in DCPS_NMM will transmit the command to DCPS_NMM. If the first level task produced more than one job, the step will be iterated.

2) After the information receiving module in DCPS_PMM received the job produced command, the information processing module will record the information through the database communicating module.

3) The job distributing module of DCPS_PMM queries database to get an undistributed job. If an undistributed job record exists, the module will get the next level task name from database. If the transaction has only one level task, the job distributing module of DCPS_PMM cannot query the name of the next level task, then go to step 6. Otherwise, the module chooses an idle computing node from database records, and sends task executing command to the DCPS_NMM of the idle node; and at same time the module writes the IP address of the idle node, distributing time and a distributed symbol to database.

4) The step 3 is iterated until there is no undistributed job in database.

5) After the information receiving module of DCPS_NMM in computing node received the task executing command, the information processing module will transmit the command to the corresponding task according to the task name in the command. And then the task starts the computing, produces and stores the result data to storage array. After computing is completed, the task sends job completed command (including row number, task name, job file name, job position and the current time) to DCPS_NMM. Then DCPS_NMM will transmit the command to DCPS_PMM in master node.

6) The information processing module of DCPS_PMM updates database according to the row number in job completed command.

7) If the information processing module of DCPS_PMM finds that the task in job completed command is not the last level task, go to step 2. Otherwise, do nothing.

8) If all the source data of the first level task and the nth level task has been processed, transaction ends.

4.5   Self-organization of DCP

In the processes of transaction, DCP can deal with the following two circumstances:

(1) Vice node failure
Vice node failure include two circumstances, which are described in Fig.6.

1) Circumstance 1: vice node fails before executing tasks
If vice node fails before executing tasks, database record corresponding to the failure node cannot be updated normally. After the heartbeat period time, the record in database corresponding to the node will be deleted. Since the failure node did nothing about tasks, the failure has no effect to transaction.

2) Circumstance 2: vice node fails while a task is being executed
If vice node fails while a task is being executed: a) the node state cannot be updated in database; b) the task that is being processed cannot be completed. In order to deal with the circumstance, DBMS deletes the record in database corresponding to the node, sets the record of which belongs to the failure node to be *undistributed*. After that, when DCPS_PMM begins to distribute tasks, it will find the unfinished job in database and redistributes the task.

(2) New node entering
A new node can join into DCP at any time. The self-organization process is described in Fig. 7.

The new node completes the initialization as a vice node at first. And then, it will keep waiting for task executing command from DCPS_PMM in master node.
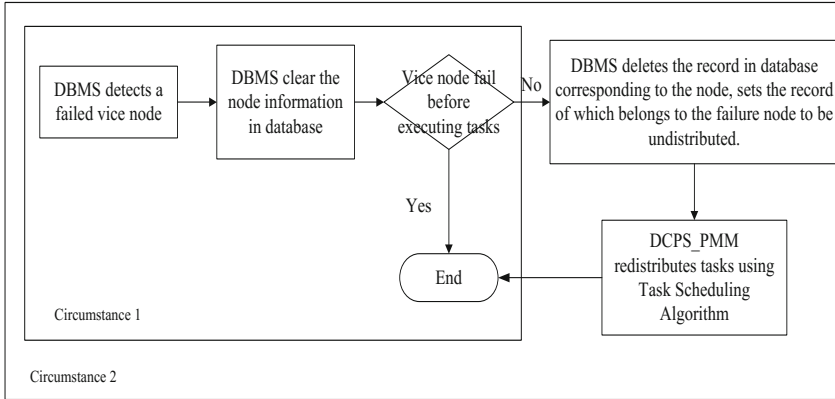
Fig. 6. The self-organization of DCP if a vice node fail
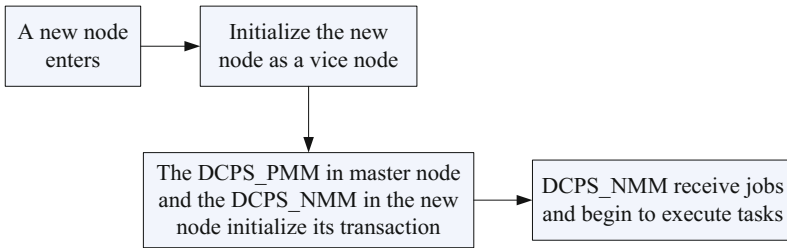
Fig. 7. The self-organization of DCP if a new node enters

5   Conclusions

The paper puts forward a task stream oriented distributed computing platform. The platform provides a reliable, scalable, self-organizable and universal solution for science computing, which is characterized by a task stream of multi tasks, such as remote sensing data processing and complex electromagnetism analysis.

References

1. Zhang, X., Li, H., Liu, Y.: The Research of Optimal Algorithm for Task Scheduling Underground Wireless Network Based on Distributed Computing. In: 2010 International Conference on Manufacturing Automation (ICMA), Hong Kong, pp. 151–155 (2010)
2. Jakovits, P., Srirama, S.N., Kromonov, I.: Stratus: A Distributed Computing Framework for Scientific Simulations on the Cloud. In: IEEE 14th International Conference on High Performance Computing and Communication & 2012 IEEE 9th International Conference on Embedded Software and Systems (HPCC-ICESS), Liverpool, pp. 1053–1059 (2012)

3. Xu, G., Lu, F., Yu, H., Xu, Z.: A Distributed Parallel Computing Environment for Bioinformatics Problems. In: Sixth International Conference on Grid and Cooperative Computing, Los Alamitos, CA, pp. 593–599 (2007)
4. Hawick, K.A., James, H.A., Maciunas, K.J., et al.: Geostationary-Satellite Imagery Applications on Distributed, High-Performance Computing. In: High Performance Computing on the Information Superhighway, Seoul, pp. 50–55 (1997)
5. Hifi, M., Saadi, T., Haddadou, N.: High Performance Peer-to-Peer Distributed Computing with Application to Constrained Two-Dimensional Guillotine Cutting Problem. In: 19th Euromicro International Conference on Parallel, Distributed and Network-Based Processing, Ayia Napa, pp. 552–559 (2011)
6. Liu, H., Sorensen, S.-A., Nazir, A.: On-Line Automatic Resource Selection in Distributed Computing. In: IEEE International Conference on Cluster Computing and Workshops, New Orleans, LA, pp. 1–9 (2009)
7. Souza Ramos, D., Watershed, T.L.A.: Watershed: A High Performance Distributed Stream Processing System. In: 23rd International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD), Vitoria, Espirito Santo, pp. 191–198 (2011)
8. Chen, L., Agrawal, G.: Self-Adaptation in a Middleware for Processing Data Streams. In: 2004 International Conference on Autonomic Computing, pp. 292–293 (2004)
9. Slawinska, M., Slawinski, J., Sunderam, V.: Unibus: Aspects of Heterogeneity and Fault Tolerance in Cloud Computing. In: 2010 IEEE International Symposium on Parallel & Distributed Processing, Workshops and PhD Forum (IPDPSW), Atlanta, GA, pp. 1–10 (2010)
10. Obaidat, M.S., Bedi, H., Bhandari, A., Bosco, M.S.D.: Design and Implementation of a Fault Tolerant Multiple Master Cloud Computing System. In: 2011 International Conference on Internet of Things and 4th International Conference on Cyber, Physical and Social Computing (iThings/CPSCom), Dalian, pp. 82–88 (2011)

Chaos in Traffic Flow Based on LA Model

Wang Jingbo[1], Yang Zhigang[1], and Wang Jingtao[2]

[1] Hebei United University, Tangshan Hebei, China
[2] Tangshan Branch Tiandi Science&Technol CO., LTD, China
`jingbow@163.com`

Abstract. Transportation System which is mainly consisted of moving people and vehicles is open, nonlinear, real-time, and multi-parameter. Traffic simulation and chaos movement states are studied based on car-following model. Firstly, LA model was presented, that is the car-following model with a cubic additional term. It is nonlinear. Then Runge-Kutta method was used to calculate of flow. Finally, the simulated traffic flow was generated with MATLAB based on the LA model. The flow was analysised by a combination of many chaos distinguishing methods. Simulation results are shown that there are chaos characteristics in the traffic flow which was derived by LA model.

Keywords: Traffic Flow, LA model, Chaos, Poincare Section.

1   Introduction

In recent years, the different states of traffic flow became one of the hot spot to study, What the most representative theory is the three-phase traffic flow theory which was proposed by Kerner through a lot of the actual traffic data[1].He divided the traffic flow state into free stream ,synchronous flow and wide moving jam. The stop and go phenomenon is a kind of composite traffic state model in three-phase traffic flow, included the synchronized flow and wide moving jam. And this phenomenon contains abundant nonlinear behavior [2-6].The studying aim of the traffic flow is not only to reveal the rule of the traffic flow's motion, but also is to lay the foundation for the chaotic control by taking advantage of the chaotic behavior in traffic flow [7-9].

2   Methods

2.1   Model

Car-following model is one of the traffic flow model which has the most application and research. It used differential equation to describe traffic running state. The relationship of the vehicle team is analyzed and studied from the reaction of the drivers. Car-following model includes linear car-following and nonlinear car-following. In order to time series of generated the traffic flow were as close as the actual traffic flow, LA model is presented. It is the car-following model with a cubic additional term added cubic additional item. .LA model can exactly explain traffic flow nonlinear characteristics, and its expression is shown in equ (1).

$$\ddot{x}_i(t) = \lambda \frac{\dot{x}_{i-1}(t) - \dot{x}_i(t)}{x_{i-1}(t) - x_i(t)} + \beta(x_{i-1}(t) - x_i(t) - D)^3 \tag{1}$$

Among them, λ is the sensitiveness, D is the ideal Space headway, β is the correction coefficient relative to the Space headway.

In this paper, we used LA model which select N cars in Car-following team. When the first car was constant speed, we added the certain interference to the first car and watched the follow car. Because the actual traffic interference can cause vehicles' acceleration or deceleration regularly .The kind of interference approximate sine change, so the interference of sine change is added to the first car in the simulation study. We choose sine change as equ (2) shown. And we discuss the interference to the traffic flow with the different influence.

$$\dot{x}_0 = v + \omega p \sin \omega t \tag{2}$$

We used Matlab software to construct LA simulation model and simulated the situation of the Car-following term composed of 6 cars. To facilitate the calculation and programming, we substituted time and space of LA model as equ (3) shown.

$$\begin{cases} T = \dfrac{\omega t}{2\pi} \\ X_i(t) = \dfrac{1}{D}(x_i(t) - vt) \end{cases} \tag{3}$$

In this way, the parameters v, ω and D are eliminated. the equation is simplified into equ (4).The simplified Xi(t) get rid of the displacement of constant speed and the ratio of ideal space headway .It's dimensionless. It can only reflect the change of displacement and it does not reflect the value of the displacement. The simplified equation can simple calculation, and also can generate the traffic flow with the same characteristics as LA model. It is shown in equ (4).

$$\begin{cases} \dot{X}_0 = 2\pi P \sin 2\pi T \\ \ddot{X}_i(T) = A \dfrac{\dot{X}_{i-1}(T) - \dot{X}_i(T)}{X_{i-1}(T) - X_i(T)} + B(X_{i-1}(T) - X_i(T) - 1)^3 \\ X_i(T) = -1, \quad \dot{X}_i(T) = 0 \end{cases} \tag{4}$$

Among them, P is the amplitude of interference signal and the unit is meter per second. A is the sensitivity coefficient. B is the sensitive coefficient of space headway. During the simulation process, the sampling interval is 1 second and the vehicle acceleration in sampling time is averaging .Each simulation step get the displacement and speed of each car and the space headway and speed difference between two cars before and after. Runge-Kutta method was used to calculate of flow.

2.2   Algorithm

Firstly, we set the interval time is h and assume the acceleration $\ddot{X}_i(t)$ is consistent in each time, which is decided by the speed and displacement in the last moment.

Then, the calculation of speed and $X_i(t)$ is obtained from the basic trapezoidal formula.

Due to equ (5), $\dot{X}_i(t)$ in equ (4) can be obtained through the following operation, and the form is as equ (6).

$$\dot{X}_i(t) = \frac{dX_i(t)}{dt} = v_i(t) \tag{5}$$

$$X_i(t+h) = X_i(t) + \int_t^{t+h} v_i(\tau)d\tau$$
$$= X_i(t) + \frac{h}{2}[v_i(t) + v_i(t+h)] \tag{6}$$

Third step, after finishing the appointed cycle times, the operation result of $\dot{X}i(t)$ and $X_i(t)$ is derived as the initial condition of $\dot{X}_{i+1}(t)$ and $X_{i+1}(t)$.

3   Results

During the simulation experiment, we select the Car-following team with 6 cars and the interval time as 0.001. It sample 1000 times in a sinusoidal periodic T. The added disturbance amplitude P = 0.05, A = 0.15, B = 800.0. The simulation time is 30 T. And we get the $X_i(t)$ in LA model .We draw a graph with the transverse axis t and the longitudinal axis $X_i(t)$. As shown in fig.1. Note that $X_i(t)$ get rid of the displacement of constant speed and the ratio of ideal space headway .It's dimensionless and it only reflect the change of displacement. It does not reflect the displacement value.

The fig.1 shows that the displacement among 4th, 5th, and 6th car causes oscillation strengthened. And the cycle becomes irregular. It is according with the characteristics of chaos. Chaos is a complex motion. Through the intuitive methods it is very difficult to distinguish. Only to analyze it certain processing, can effectively estimate. There are some Judging methods of chaos: the reconstruction of phase space, Correlation dimension, Poincare section, Lyapunov index, power spectrums and so on. Using nonlinear sequence analysis tool TSTool [10] embedded MATLAB, we get the following data.
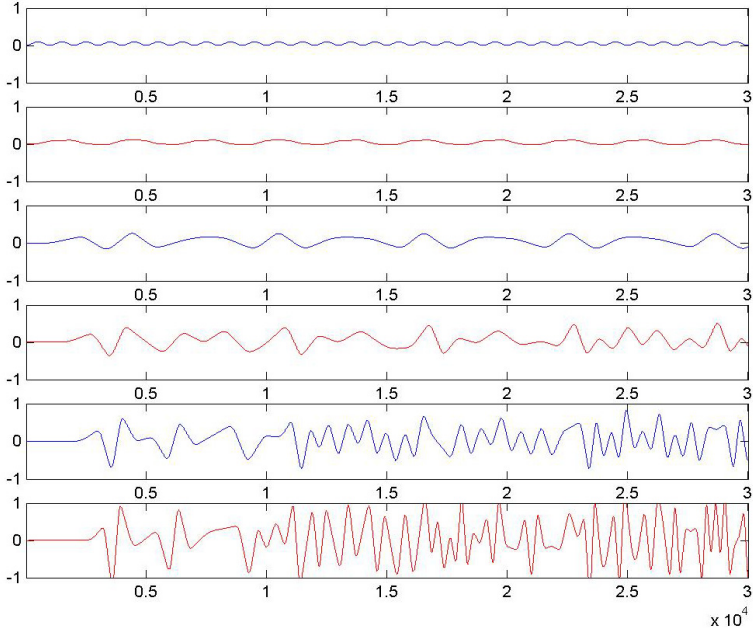
Fig. 1. This shows a figure of the displacement different with the sine change interference

3.1   The Reconstruction of Phase Space

The reconstruction of phase space is a key parameter in phase portrait, bifurcation diagram and Lyapunov exponent. The selection of embedding dimension and time delay is the key in phase space reconstruction. Using TSTool, we get Embedding dimension and Time delay.
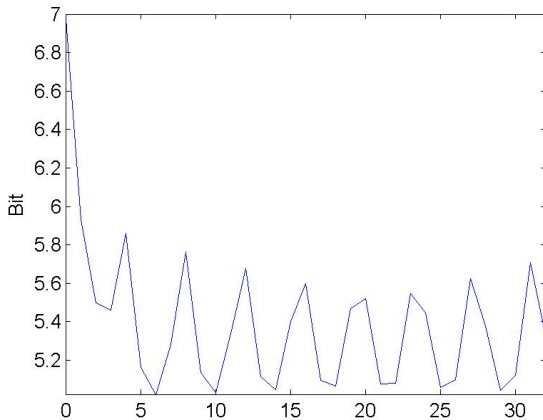
Fig. 2. This shows a figure of the selection of embedding dimension of 1st car

The example of 1st car was used to illustrate the selection of embedding dimension and time delay. As indicated in fig 2. The minimum embedding dimension is the kink in fig 2 and we select m is 2. Time delay is the position of the first minimum in fig 3 and we select T is 4.
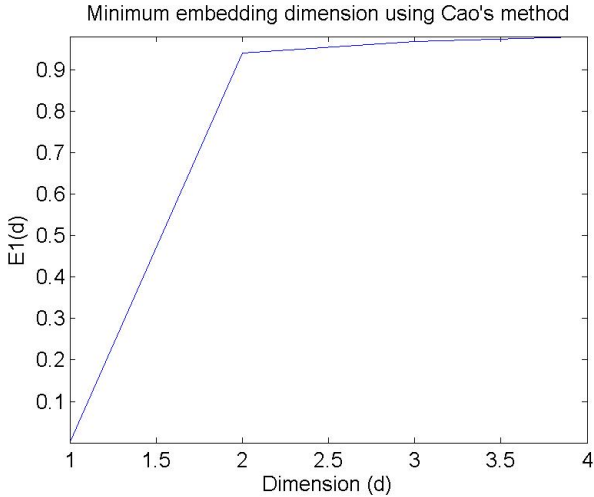
Fig. 3. This shows a figure of the selection of time delay of 1st car

The idea of this method is to calculate 2nd -6th car. Table 1 list the selection of embedding dimension and time delay from 1st car to 6th car.

Table 1. The selection of embedding dimension and time delay

Number	Time delay	Embedding dimension
1st car	4	2
2nd car	725	3
3rd car	530	3
4th car	370	3
5th car	210	3
6th car	170	4

3.2   Correlation Dimension

Correlation dimension is taken as the measurement parameters of chaotic behavior and it also has been applicated widely .Correlation dimension are the advantageous tool to judge of one dimensional time series. Table 2 list the correlation dimension D when the scale r = 0.05.

Table 2. The correlation dimension D in 6 cars with r=0.05

Number	Font size and style
1st car	0.9069
2nd car	2.1627
3rd car	2.5585
4th car	3.6471
5th car	4.3319
6th car	5.6388

3.3   Poincare Section

Poincare section shows the basic shape of chaos attractor which is the intuitive characteristics. But it does not reflect the degree of chaos .In order to display the Poincare section of traffic flow in this paper intuitively, we make the two dimensional phase diagram which was form of displacement difference and speed difference with LA model firstly. As shown in fig. 4. After placing the section in the two dimensional phase diagram, we get the Poincare section as shown in fig. 5.
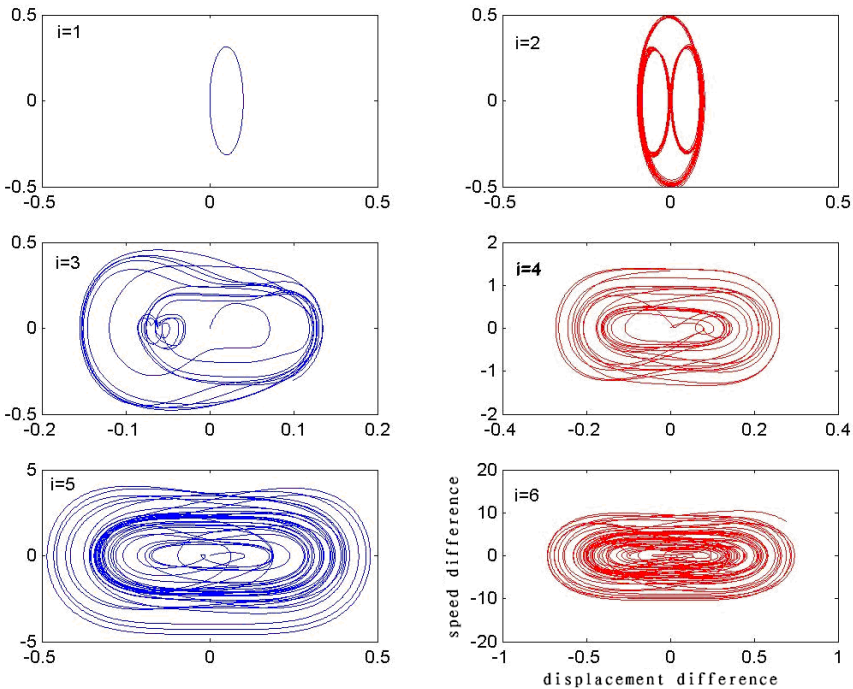
Fig. 4. This shows a figure consisting of the two dimensional phase diagram in 6 cars
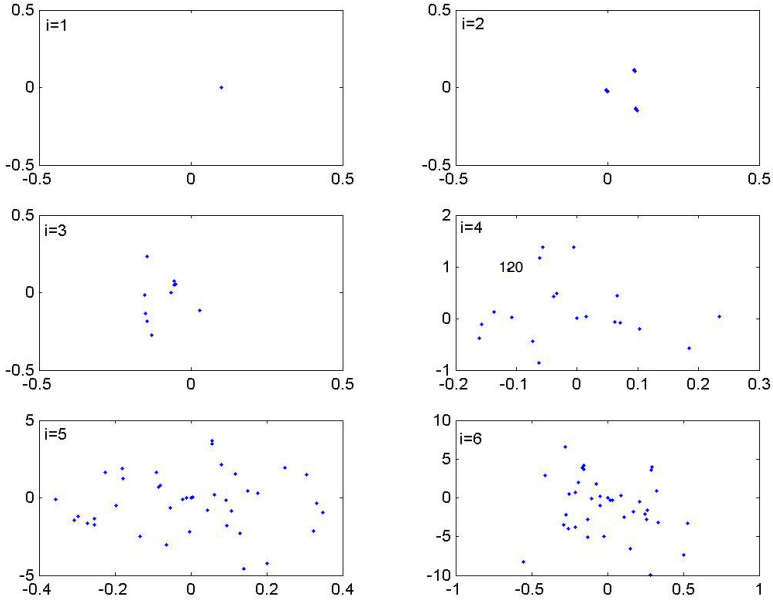
Fig. 5. This shows a figure consisting of the Poincare section in 6 cars

3.4   Lyapunov Index

Lyapunov index is the main parameters to represent the statistical characteristic value in Chaos. It is a metric in the phase space that is the average convergence or the average divergent in the near orbit. Whether there is dynamics chaos for the system, we figure out intuitively from that the maximum Lyapunov index is greater than zero or not. Table 3 list the maximum Lyapunov index in 6 cars.

Table 3. The maximum Lyapunov index in 6 cars

Number	Font size and style
1st car	-0.2873
2nd car	-0.1367
3rd car	-0.0039
4th car	2.8287
5th car	3.0316
6th car	4.5879

3.5   Power Spectrum

It is used to study on the motion of the complex nonlinear system by power spectrum. It is obtained by the Fourier transform in phase space. The power spectrum of the

periodic motion is the discrete peaks and the ratio between frequencies is a rational number. The power spectrum of the quasi periodic motion is different cycle and the ratio between frequencies is an irrational number. The Power spectrum of the random motion is continuous spectrum and the amplitude is independent with the frequency. There still may have a peak power spectrum of chaotic motion. But they will be widened, and also may appear the bandwidth of the background noise. It is very favorable that the identification and analysis of the conversion process between the periodic and the chaotic state by the power spectrum.
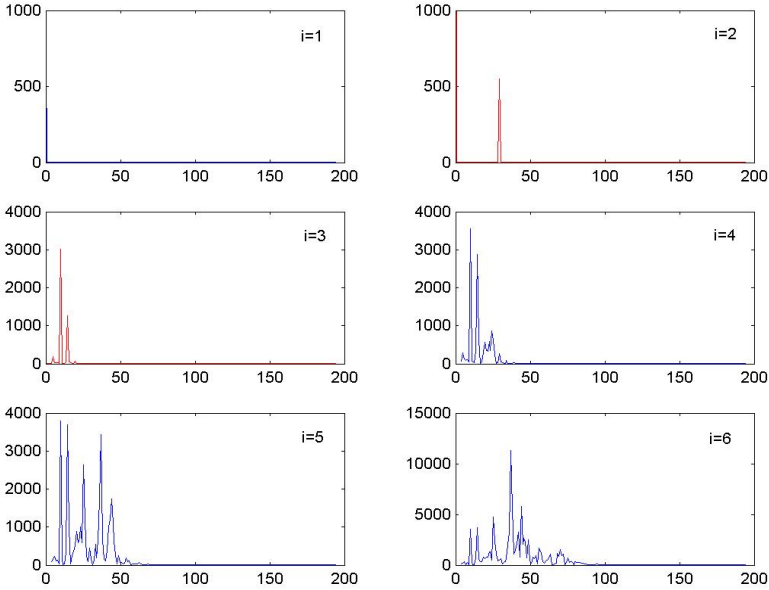
Fig. 6. This shows a figure consisting of the power spectral in 6 cars

Using MATLAB, we get power spectral from 1st car to 6th car, as is shown in fig.6. The power spectrum of the 1st and 2nd cars is the periodic motion. The power spectrum of the 3rd car is the quasi periodic motion. The power spectrum of the 4th to 6th cars is the chaotic motion because there is still a peak power spectrum of chaotic motion. But they were widened, and also appeared the bandwidth of the background noise.

3.6   Changes in the Interference Condition

The interference conditions which appeared in actual cases can be the one-off interference. We select the car-following team composed of 6 cars. The interval time is 0.001, and A = 0.15, B = 800.0, D = 1.0, P = 0.02. The simulation time is 30T. We compared the situation with fig. 1 when we change the interference for one-off sine signal. We observed the change of traffic flow, as shown in fig. 7.
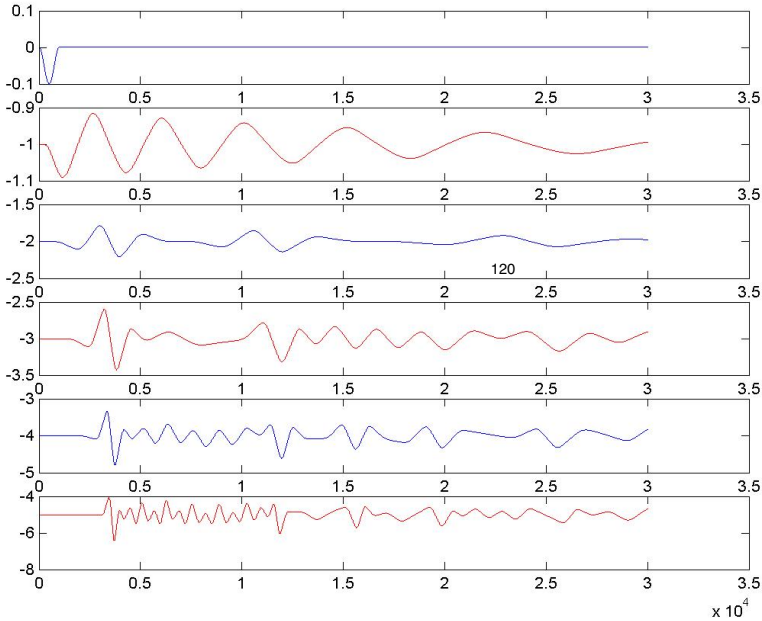
Fig. 7. This shows a figure of the displacement different with the one-off interference

The interference signal used in fig.4 is one-off sine interference signal. There is the interference condition according to the pedestrian. For example, while the car-following team was constant driving in the driveway, a pedestrian crossed the lane quickly .The 1st car suffers the disposable interference. The kind of interference is similar to a half cycle of sine condition. As shown in fig.4, the following car will still produce chaotic behavior, and the amplitude increases gradually. The kind of situation is completely same with Butterfly Effect that Lorenz put forward.

4   Discussion

This paper judged chaotic behavior in LA model. From the simulation results above and its analysis may safely draw the conclusion that the traffic flow generated by LA model exist chaos. And the car in the traffic flow with different number has different motion law. The traffic flow movement presents some regularity, chaos movement and periodic motion is changed for existence. In addition, the interference condition according to the pedestrian is discussed. The following car will still produce chaotic behavior, and the amplitude increases gradually. The kind of situation is completely same with Butterfly Effect.

Reference

1. Kerner, B.S.: Three-phase traffic theory and highway capacity. Physica A 333, 379–440 (2004)
2. Wang, D., He, G.: Summary and prospects of the study on traffic chaos. China Civil Engineering Journal 36(1), 68–73 (2003)
3. Low, D.J., Addison, P.S.: Chaos in a car-following model a desired headway time. In: Proceeding of the 30th ISATA Conference, Florence, Italy, pp. 175–182 (1997)
4. Dong, C., Shao, C., Zhuge, C., Meng, M.: Spatial and Temporal Characteristics for Congested Traffic on Urban Expressway. Journal of Beijing Polytechnic University 38(8), 128–132 (2012)
5. Lu, Y.: Identification of Chaos in Microcosmic Simulation Traffic Flow Based on Improved Surrogate-data Technique. Journal of Wuhan University of Technology (Transportation Science & Engineering) 35(3), 94–97 (2011)
6. Liu, S., Guan, H., Yin, H.: Chaotic behavior in the dynamical evolution of network traffic flow and its control. Acta Physica Sinica 61(9), 64–73 (2012)
7. Liu, L., Li, S., Xie, Y.-l.: Comparative Study on Chaotic Prediction Model in Short-term Traffic Flow. Mathematics in Practice and Theory 41(17), 108–116 (2011)
8. Lu, Y., Zhang, X.: Influence of Feature in Time Series on the System for Real-time Identification of Chaos in Traffic Flow. Journal of Wuhan University of Technology (Transportation Science & Engineering) 35(5), 91–94 (2011)
9. Dong, C.-J., Shao, C.-F., Li, J., Meng, M.: Short-term traffic flow prediction of road network based on chaos theory. Journal of Systems Engineering 26(3), 340–345 (2011)
10. Christian, M., Ulrich, P., Werner, L.: Tools for Nonlinear Data Analysis. In: NOLTA 1998 Crans-Montana, Switzerland, pp. 1253–1255 (1998)

Influence Factors Analysis of Capital Structure of Power and Coal Industry Listed Companies

Rui-shu Liu

Business School China of the University of Nottingham Ningbo China
199 Taikang East Road, Ningbo, China
maria2002528@163.com

Abstract. Capital structure is of great importance to company's financial decision- making. China's power and coal industry are vital to its economy and it really matters to their development, even to a stable development of China's economy whether or not the capital structure of the listed companies and coal industry companies will be scientifically managed. This paper gives a summary of the factors affecting capital structure, finds out every component which is representative of specific financial indexes and proposes hypotheses. Together with the inherent development features and conditions of these industries, we make a quantitative analysis on the factors by using SPSS and the method of principal component analysis. We draw the conclusion that the negative correlations are found between the capital structure of these two industries and profitability, asset liquidity and company growth, but the positive correlations exist between capital structure and company size, debt tax shield, asset structure and internally financing capability.

Keywords: capital structure, listed companies, influence factors.

1 Introduction

Capital structure is the key to financing and management of the company. Whether capital structure is rational or not will have direct impacts on the company's present and future developments and it is a real reflection of the stakeholders' rights and obligations. Capital structure is influential and decisive in company management, and, to a further degree, in the behavior and the value of the company. Thus, capital structure may be referred to as modern companies' vital "gene". A proper capital structure is beneficial to regulate the behavior of the company and improve the value of the company. Giving a broad overview, many problems are found in the capital structure and the financing strategy of listed companies and coal companies Of China, and their structure needs optimizing. While these companies possess typical industry characteristics and capital structure theories have been further advancing in China, no enough capital structure theories are elaborately classified according industries. As every industry has its own distinctive capital structure, a study of capital structure and

its influence factors of the listed companies and coal companies are of great theoretical value and practical significance to promote a sound development of these companies and advance the transformation of state-owned enterprises into the share-holding system.

2 Literature Review

Franco and Miller (1958) state that the market value of a company is determined by its earning power and the risk of its underlying assets, and is independent of the way it chooses to finance its investments or distribute dividends [1]. Justin Pettit (2009) in his Strategic Corporate Finance provides a practical guide to the key issues, their context, and their solutions. He translates principles of applications in valuation and capital structure by managing the left -hand side of the balance sheet, the right -hand side of the balance sheet and the enterprise [2]. Gerald S. (2011) gives a comprehensive discussion of financial theory, empirical work and practice concerning financial policies, strategies and choices. They adopt a practical approach to capital structure by introducing why many theories make sense and how companies solve problems and produce wealth with help of these theories [3]. Wu and Xiao (1999) make an empirical analysis indicating: there exist positive correlations between capital structure and ratio of state-owned shares, the guarantee value of assets and company size while there exist negative correlations between company's capital structure and profitability, company growth, and non-debt tax shield [4]. Wu (2003) believes that various factors including company size would, to a larger extent, affect the financing preference of the company, and may further affect its capital structure, and this shows that capital structure of the listed companies of China differs from that of the mature capital markets [5]. Zhang (2007) concludes that the performance of the private listed companies is better than that of the state-owned companies, and capital structure is the reason behind their different performance [6]. Hu (2008) maintains that capital structure isn't only ratio of total liabilities to total assets, but leverage of liabilities [7]. Zhu (2009) states that financial factors affect capital structure, and capital structure affects the value of the company, but all these can be traced back to optimal capital structure of the listed companies [8]. Ma's (2011) research shows that the performance of a listed company is clearly relevant to its capital structure and management. As a result, optimal capital structure and better management are the methods of improving the performance of the company [9]. Ge (2010), Chen (2010), Yang (2010) and Liu (2011) believe that there is relationship between enterprise's profit ability and capital structure [10-12].

Based on the above analyses, this paper sorts out the three-year (2009-2011) financial affairs of the listed companies and coal companies of China, presents the hypotheses that affect capital structure of these companies. With several analytical methods, the paper finds out the factors affecting capital structure of the listed companies and coal companies and proposes solutions to optimizing the listed companies and coal companies of China.

3   Analyses of the Present Situations of Capital Structure of the Listed Companies and Coal Companies

3.1   The Present Situations of Capital Structure

Table 1 indicates that the assets liabilities of power and coal industry are slightly different from those of all industry and they remain stable yearly but asset liability ratios do not reflect characteristics of the power and coal industry. Table 2 indicates that long-term asset liability ratios of the power and coal industry are far greater than those of all industry.

Table 1. Contrast asset liability ratios (%)

Year	2006	2007	2008	2009	2010
all industry	57.46	57.48	57.71	57.88	57.41
power	57.23	58.16	63.98	64.66	65.83
coal	60.65	61.15	59.02	59.06	58.17

Table 2. Contrast long-term asset liability ratios (%)

Year	2006	2007	2008	2009	2010
all industry	24.08	23.09	24.38	25.66	24.7
power	43.27	43.29	49.47	51.7	51.32
coal	34.31	34.77	31.62	33.56	32.63

Discussing the impacts of internal financing, equity financing and debt financing on the value of the company while considering return of net assets and earnings per share, we can draw the next conclusion: In China, the preferential financing methods of the listed companies and coal companies are internal financing, equity financing and debt financing in order. However, some empirical studies indicate that their financing preference is, in order, equity financing, short-term debt financing, long-term debt financing and internal financing. By comparison, we can see that the orders concerning equity financing and debt financing tend to be similar. We realize that the listed companies and coal companies do not make optimal choices of financing for a long-term development of China's social economy.

3.2   Hypotheses

With analyses of company size, profitability, asset structure, growth, non-debt tax shield, asset liquidity, solvency, and internal financing, this paper puts forward the following hypotheses: Size is positively correlated with capital structure. Profitability is negatively correlated with capital structure. Asset structure is positively correlated with capital structure. Growth is positively correlated with capital structure. Non-debt tax shield is negatively correlated with capital structure. Asset liquidity is positively

correlated with capital structure. Solvency is positively correlated with capital structure. Internal financing is positively correlated with capital structure.

4   Empirical Study

4.1   Model

As stated above, we sort out many variables, which affect capital structure of the listed companies. Having conducted a multiple regression analysis of the influence factors with SPPSS19.0, we establish a multiple linear regression equation, demonstrating how these variables will affect capital structure of the listed companies and coal companies of China. The equation is as follows:

$$Y = a + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n + \varepsilon$$

Y=independent variable, i.e. total asset liability. α=constant. B=regression coefficient. X_n=independent variable, i.e. factors affecting capital structure of China's listed companies and coal companies. ε=random variable. Table 3 shows a hypothesized relationship between capital structure and influence factors.

Table 3. Component Definition

Variable		Component	Index	Metric
Dependent variable	Y	capital structure	asset liability	total liabilities/total assets
Independent variable	X_1	profitability	asset profitability	net income/total assets
	X_2	asset liquidity (solvency)	current ratio	current assets/current liabilities
	X_3	asset liquidity (solvency)	quick ratio	quick assets/current liabilities
	X_4	company size	LN (total assets)	LN (total assets at the end of the year)
	X_5	non-debt tax shield	non-debt tax shield	depreciation of fixed assets/total assets
	X_6	asset structure	ratio of fixed asset	fixed assets/total assets
	X_7	company growth	growth on total assets	growth on total assets / average of total assets prior year
	X_8	internal financing	cash flow	net cash flow/ total assets

According to the publications of the industry situations by Sina Finance, by May, 23rd, 2012, there had been 89 listed companies and coal companies in China. This paper follows the next five criteria of choosing the samples:

To assure that the samples are comparable, only the companies listed on A- Stock Market of China are selected but those listed on B -Stock Market of China are excluded. ST companies are removed as these companies have been running at a loss for years on end and the financial information released are not convincing. Operating revenue structure is examined and the companies are removed if revenue gained from power or coal accounts for less than 50% of the prime operating revenue. The companies are removed if some of the financial data is not available. The companies listed on the marked within the latest three years are removed in order to assure that the corporate behavior is relatively mature and the samples are comparable. There are 45 companies satisfying above criteria and 44 are removed. The period of time for choosing these samples is from 2009 to 2011. Thus there are eventually 135 samples to be studied.

4.2   Descriptive Statistics Analysis

According to the statistics shown in Table 4, we could obtain the following views.

Asset liabilities ratios of power and coal industry listed companies **range from** 26.04% to 89.75%, which indicates that the polarization is serious in asset liabilities ratio of power and coal industry and a stable asset liabilities structure has not come in to being. Furthermore, the mean of it is 59.93%, which is basically equal to that of all domestic listed companies

The range of return on total assets fluctuates between 10.73% and 5.33%, which also polarizes seriously. This implies that the profitability of the whole industry remains to be improved. From the above table, we can see that the current ratio varies between 12.22% and 234.76% and the mean of it is 79.83%, which is less than 100%. As for quick ratio, it varies from 12.08% to 226.44% and the mean of it is 68.01%, which is more than 50%, but less than 75%. Consequently, asset liquidity is low in power and coal industry listed companies, so that solvency remains to be improved. Moreover, the mean of internal financing is merely 0.86%, which indicates internal financing capacity of these listed companies is weak.

Non-debt tax shield of power and coal industry listed companies ranges from 7.7% to 89.82% and the mean of it is 29.50%. These suggest that the ratio of the expenses on taxes except debt interests is imbalanced and the whole ration is at a low level.

As the table shows, fixed assets ratio varies between 18.51% and 88.54% and the mean of it is 49.17%, which indicates that there is evident polarization in fixed assets ratio, but the overall ration is moderate.

The above statistics has shown that growth on total asserts ranges from -6.98% to 176.16%, with an average of 17.83%. These indicate the serious polarization of the growth on total assets in power and coal industry listed companies. These also show even negative growth has occurred in some companies in recent several years. However, the whole industry appears to keep growth in a positive state.

Table 4. Descriptive Statistics Analysis Table of Influence Factors of Capital Structure

	N	Minimum	Maximum	Mean	standard deviation	variance
	statistic	statistic	statistic	statistic	statistic	statistic
Asset liabilities ratio	45	.2604	.8795	.5993	.1577	.025
Return on total assets	45	.1073	.9533	.3376	.1522	.023
Current ratio	45	.1222	2.3476	.7983	.5121	.262
Quick ratio	45	.1208	2.2644	.6801	..4630	.214
LN(total assets)	45	20.4133	26.2620	23.4639	1.5771	2.487
Non-debt tax shield	45	.0770	.8982	.2950	.1468	.022
Fixed assets ratio	45	.1851	.8854	.4917	.1726	.030
Growth on total assets	45	-.0698	1.7616	.1783	.2759	.076
Internal financing	45	-.1812	.2889	.0086	.0621	.004

4.3   Analyses of the Factors

The test result shows that KMO=0.579>0.5, and Sig.=0.000, and this indicates the data is suitable for the component analysis. We use the method of Principal Component Analysis and the sampled data are principal factors. According to the statistics, total percentage adds up to 79.232 when three principal factors are selected and this shows that the explanation ratio of the three principal factors is adequately satisfactory.

By Component Matrix in Table 5, we can see three principal factors, simplifying the data studied, but the matrix does not exactly reflect the relationship between individual factors and it needs to be rotated. For a better explanation of the relationship between variables, Component Matrix is rotated, and then in this way, every component in the matrix becomes orthogonal, as shown in Table 6.

According to Rotated Component Matrix, we can define every principal component:

Component A is positively correlated with current ratio and quick ratio; Current ratio and quick ratio belong to solvency, i.e. asset liquidity indexes. Component A is very negatively correlated with ratio of fixed assets, which indicates capital structure.

Component B is very positively correlated with return on total assets and non-debt tax shield, and next positively correlated with growth on total assets. Return on total assets reflects profitability while growth on total assets reflects company growth.

Component C is very positively correlated with LN (total assets) and internal financing. LN (total assets) indicates company's scale.

From the above analyses, we believe that solvency, asset structure, profitability, non-debt tax shield, company size and internal financing all have significant impacts on capital structure of China's listed companies and coal companies.

Table 5. Component matrix

	Component		
	A	B	C
Return on total assets	.506	.820	.032
Current ratio	.877	-.376	-.158
Quick ratio	.870	-.387	-.108
LN(total assets)	-.153	-.401	.685
Non-debt tax shield	.366	.894	.049
Fixed assets ratio	-.678	.472	-.137
Growth on total assets	.480	.380	.495
Internal financing	-.033	-.024	.803

Table 6. Rotated component matrix

	Component		
	A	B	C
Return on total assets	.047	.950	-.158
Current ratio	.956	.069	-.128
Quick ratio	.953	.066	-.077
LN(total assets)	.026	-.280	.758
Non-debt tax shield	-.113	.949	-.149
Fixed assets ratio	-.814	.050	-.190
Growth on total assets	.213	.653	.383
Internal financing	-.056	.123	.792

We can find Component Score Coefficient Matrix when we explain the factors, demonstrating the above three factors are comparable, as shown in Table 7.

Table 7. Component score coefficient matrix

	Component		
	A	B	C
Return on total assets	.193	.358	.023
Current ratio	.333	-.164	-.112
Quick ratio	.331	-.169	-.076
LN(total assets)	-.058	-.175	.484
Non-debt tax shield	.139	.390	.034
Fixed assets ratio	-.258	.206	-.096
Growth on total assets	.182	.166	.349
Internal financing	-.012	-.011	.566

4.4   Regression Analysis

With SPPSS19.0, we conduct a regression analysis, testing the above hypotheses. Enter Method is used in this paper, namely, enter eight influence component variables into the model and the test result is shown in Table 8.

According to the dada in Table 8, we can see that capital structure of the listed companies and coal companies of China is positively correlated with LN (total assets), non-debt tax shield and internal financing while it is negatively correlated with return on total assets, asset liquidity, fixed assets ratio and growth on total assets. A contrast of the empirical analysis and the hypothesis is shown in Table 9.

Table 8. Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t
	B	Std. Error	Beta	
(Constant)	.301	.240		1.252
Return on total assets	-1.646	.337	-1.590	-4.887
Current ratio	-.305	.114	-.991	-2.665
Quick ratio	.183	.125	-.537	1.458
LN(total assets)	.021	.009	.211	2.243
Non-debt tax shield	1.639	.349	1.525	4.695
Fixed assets ratio	-.004	.106	.004	-.034
Growth on total assets	-.022	.057	-.039	-.388
Internal financing	.042	.228	.017	.184

Table 9. Contrast results

Component	Profitability	Asset liquidity	Company size	Non-debt tax shield	Asset structure	Company growth	Internal financing
Hypothesized relationship	-	+	+	-	+	+	+
Empirical relationship	-	-	+	+	+	-	+

5   Conclusion

Conclusion one: Capital structure is negatively correlated with profitability. When a company's profitability is stronger, its internal accumulation will be stronger. When a company's profitability is weak, it has to turn to external debt financing. Companies will have a high debt ratio if they can not perform well, and this type of case is quite common in China. This result supports the point of the pecking order theory that companies with adequate internal financing prefer a low asset liability. Thus profitability is negatively correlated with capital structure.

Conclusion two: capital structure is negatively correlated with asset liquidities. The empirical analysis result is different from that of the hypothesis expectation. The listed companies and coal companies of China have a high mobility, causing a high degree of capital use. The companies have enough capital to expand reproduction and cut down external capital.

Conclusion three: Capital structure is positively correlated with a company size. Companies which are of larger sizes, good guarantees of debts and good credit will reduce the risk of bankruptcy. The listed companies and coal companies of China are mainly of large and medium-sized and have good credit and low management risk. Thus creditors can trust them and they can obtain debt financing with lower costs, improving their liabilities.

Conclusion four: Capital structure is positively correlated with non-debt tax shield. In academic circles, there have been controversies over impacts of non-debt tax shield on a company's capital structure and some scholars' empirical studies show that they are not obviously correlated. According to the provisions of the tax law, this paper suggests that companies can put extracted depreciation charge into costs for a pretax deduction.

Conclusion five: Capital structure is positively correlated with asset structure. According with the hypotheses mentioned above, If the fixed assets are greater, the asset liabilities will be higher Capital Structure theory states that asset structure affects a company's capital structure.

Conclusion six: Capital structure is positively correlated with a company's growth. If the company's growth becomes better, the asset liabilities will be the greater. The conclusion is opposite to the theoretical analysis. When a company is at a developing stage, it needs more capital but it is acceptable for the company to operate on borrowings and it chooses to have a high liability for the development.

Conclusion seven: Capital structure is positively correlated with internal financing. According to Capital Structure theory, companies with strong internal financing tend to take more liabilities for a tax shield effect. As these companies run the low risk of bankruptcy and are likely to get better issuing provisions such as a low degree of risk premium, the managers choose to accept financial risks related to liabilities more than those of the enterprises with weak internal financing. In this case, the companies with strong internal financing can choose higher liabilities. Agency Cost theory, Trade-off theory and Capital Structure Signaling theory maintain that capital structure is positively correlated with internal financing.

References

1. Modigliani-Miller Theorem - M&M, `http://www.investopedia.com`
2. Pettit, J.: Strategic Corporate Finance: Applications in Valuation and Capital Structure. John Wiley & Sons (2007)
3. Kent Baker, H., Martin, G.S.: Capital Structure & Corporate Financing Decisions: Theory, Evidence, and Practice. Wiley (2011)
4. Wu, S.: Theories and Methods of Modern Accounting. China Economy Press (1997)
5. Wu, X.: China's Listed Companies: Capital Structure and Company Management, pp. 35–39. China Renmin University Press, Beijing (2003)
6. Zhang, Z., He, W., Liang, Z.: Capital Structure and Company Performance: Evidence from China's Stated-owned Listed Companies and Private Listed Companies. Business Manegement (12), 141–151 (2007)
7. Hu, Y.: Studies on Capital Structure Concerning China's Enterprises, pp. 66–72. China Finance and Economy Press, Beijing (2008)
8. Zhu, Y.: Theoretical and Empirical Studies on Listed Companies of China, pp. 59–82. China Renmin Press, Beijing (2009)
9. Ma, X.: Study on Capital Structure, Company Management and Performance Concerning China's Listed Companies-Based on An Empirical Study of 300 Samples from Shanghai and Shenzheng Stock Markets. Modern Business Trade Industry (11), 154–155 (2011)
10. Ge, H., Mo, S.: Empirical Studies on Capital Structure Features and Influence Factors of China's Real Estate Companies—Cases from Listed Companies. Huanghai Academic Forum (14), 14–17 (2010)
11. Chen, Y., Shao, X.: Capital Structure of Listed Companies Influence Factors-The Cross-sectional Data Based on China's Listed Companies. Communication of Finance and Accounting (3), 90 (2010)
12. Yang, M, Shi, J., Yu, Y.: The Empirical Study on Influence Factors of Capital Structure. Communication of Finance and Accounting. (5), 50 (2010)

Numerical Solution for a Kind of Nonlinear Telegraph Equations Using Radial Basis Functions

Ling De Su[1,2], Zi Wu Jiang[1], and Tong Song Jiang[1]

[1] Department of Mathematics, Linyi University, Linyi, 276005, P.R.China
[2] College of Mathematics, Shandong Normal University, Jinan, 250014, P.R. China
jiangtongsong@sina.com, lingdesu@yahoo.cn

Abstract. In this paper, we propose a numerical scheme to solve a kind of the nonlinear telegraph equation by using the Kansa's method with Radial Basis Functions (RBFs). From the numerical results of experiments presented in this paper, we can get that the accuracy between the numerical solutions and the analytical solutions are valid. In this paper, we also give the analysis of the parameter c in IMQ radical basis function for the results.

Keywords: Nonlinear telegraph equation, the Kansa's method, Radial Basis Function, Thin Plate Spline, Inverse Multiquadric.

1 Introduction

This paper is devoted to the numerical computation of the nonlinear telegraph equations takes the following form:

$$u_{tt} + a_1 u_t = a_2 u_{xx} + f(u),\qquad(1)$$

where a_1, a_2 are known constant coefficients, $f(u) = \alpha u^3 + \beta u^2 + \gamma u$, α, β and γ are real constant. Eq.(1) referred to as second-order nonlinear telegraph equation. Telegraph equation describes various phenomena in many applied fields, such as a planar random motion of a particle in fluid flow, transmission of electrical impulses in the axons of nerve and muscle cells, propagation of electromagnetic waves in superconducting media and propagation of pressure waves occurring in pulsatile blood flow in arteries[1,2].

In the recent years much attention has been given in the literature to the development, analysis and implementation of stable numerical methods for the solution of second-order telegraph equations [3,4]. The existence of time-bounded solution of nonlinear bounded perturbations of the telegraph equation with Neumann boundary conditions has recently been considered in [5]. The approach is based upon a Galerkin method combined with the use of some Lyapunov functionals. The exact solution of nonlinear telegraph equations are also given in many papers [6,7].

Over the past several decades, many numerical methods have been developed to solve boundary-value problems involving ordinary and partial differential equations.

Finite difference methods are known as the first techniques for solving partial differential equations [8,9]. However the need to use large amount of CPU time in implicit finite difference schemes limit the applicability of these method. Furthermore, these methods provide the solution of the problem on mesh point only [10] and accuracy of these well known techniques is reduced in non-smooth and non-regular domains.

To avoid the mesh generation, in recent years meshless techniques have attracted attention of researchers. Some meshless schemes are the element free Galerkin method, the reproducing kernel particle, the point interpolation and etc. More descriptio see [11] and reference therein.

In last 25 years, the radical basis functions (RBFs) method is known as a powerful tool for scattered data interpolation problem. Because of the collection technique, this method does not need to evaluate any integral. The main advantage of numerical procedures which use RBFs over traditional techniques is meshless property of these methods. RBFs are used actively for solving partial differential equations. The examples see [12,13,14].

In the last decade, the development of the RBFs as a truly meshless method for approximating the solutions of PDEs has drawn the attention of many researchers in science and engineering. Meshless method has became an important numerical computation method, and there are many academic monographs are published [15,16].

In this article, we present a new numerical scheme to solve nonlinear telegraph problem using the meshfree method with Radial Basis Functions (RBFs). The results of numerical experiments are presented, and are compared with analytical solutions to confirm the good accuracy of the presented scheme.

The layout of the article is as follows : In section 2, we introduce the method and apply this method on the nonlinear telegraph problem. The results of numerical experiments are presented in section 3. Section 4 is dedicated to a brief conclusion. Finally, some references are introduced at the end.

2 The Method of Kansa's

2.1 Radial Basis Function Approximation

The approximation of a distribution $u(X)$, using RBF, may be written as a linear combination of N radial functions, usually it takes the following form:

$$u(X) \approx \sum_{j=1}^{N} \lambda_j \varphi(X, X_j) + \psi(X), \quad \text{for } X \in \Omega \subset R^d , \qquad (2)$$

where N is the number of data points, $X = (x_1, x_2, \cdots, x_d)$, d is the dimension of the problem, the $\lambda's$ are coefficients to be determined and φ is the radial basis function. Eq. (2) can be written without the additional polynomial ψ. In that case,

φ must be unconditional positive definite to guarantee the solvability of the resulting system (e.g. Gaussian or inverse multiquadrics). However, ψ is usually required when φ is conditionally positive definite, i.e, when φ has a polynomial growth towards infinity. We will use RBFs, which defined as:

Thin Plate Spline (TPS) : $\varphi(X, X_j) = \varphi(r_j) = r_j^{2m} \ln r_j, m \in N^+$

Inverse Multiquadric (IMQ) :

$$\varphi(X, X_j) = \varphi(r_j) = \frac{1}{\sqrt{r_j^2 + c^2}}, c > 0 \tag{3}$$

where $r_j = \|X - X_j\|$ is the Euclidean norm. Since φ given by (3) is C^∞ continuous, we can use it directly.

The IMQ radial basis function takes the form: $\frac{1}{\sqrt{r_j^2 + c^2}}, c > 0$. The choice of parameter c is very important for the solution accuracy. Unsuitable parameter will produce the singular matrix. Moreover, how many nodes we choose can also affect the accuracy. The theory of radial basis function difference method can see [17,18].

If P_q^d denotes the space of d-variate polynomial of order not exceeding than q, and letting the polynomials $p_1, p_2, \cdots, p_m$ be the basis of P_q^d in R^d, then the polynomial $\psi(X)$ in Eq. (2) is usually written in the following form:

$$\psi(X) = \sum_{i=1}^{m} \varsigma_i p_i(X_j), \tag{4}$$

where $m = \frac{(q-1+d)!}{d!(q-1)!}$.

To get the coefficients $(\lambda_1, \lambda_2, \cdots, \lambda_N)$ and $(\varsigma_1, \varsigma_2, \cdots, \varsigma_m)$, the collocation method is used. However, in addition to the N equations resulting from collocating Eq. (2) at the N points, an extra m equations are required. This is ensured by the m conditions for Eq. (2),

$$\sum_{j=1}^{N} \lambda_j p_i(X_j) = 0, i = 1, 2, \cdots, m. \tag{5}$$

In a similar representation as Eq. (2), for any linear partial differential operator L, Lu can be approximated by

$$Lu(X) \approx \sum_{j=1}^{N} \lambda_j L\varphi(X, X_j) + L\psi(X). \tag{6}$$

2.2   Nonlinear Telegraph Problems

Let us consider the following time-dependent problems:

$$u_{tt} + a_1 u_t = a_2 u_{xx} + f(u), x \in \Omega \cup \partial\Omega = [a,b] \subset R, 0 < t < T, \qquad (7)$$

with initial condition

$$u(x,0) = g_1(x), x \in \Omega,$$
$$u_t(x,0) = g_2(x), x \in \Omega \qquad (8)$$

and Dirichlet boundary condition

$$u(x,t) = h(x,t), x \in \partial\Omega, 0 < t < T, \qquad (9)$$

where a_1, a_2 are known constant coefficients, $f(u) = \alpha u^3 + \beta u^2 + \gamma u$ and α, β, γ are known constant coefficients, the function $u(x,t)$ is unknown.

First, let us discretize Eq.(7) according to the following θ-weighted scheme:

$$\frac{u(x,t+dt) - 2u(x,t) + u(x,t-dt)}{(dt)^2} + a_1 \frac{u(x,t+dt) - u(x,t-dt)}{2 \cdot dt}$$
$$= a_2 \theta \Delta u(x,t+dt) + a_2(1-\theta)\Delta u(x,t) + f(u(x,t+dt)), \qquad (10)$$

where $0 < \theta < 1$, and dt is the time step size and Δ is the Laplace operator, using the notation $u^n = u(x,t^n)$ where $t^n = t^{n-1} + dt$, we get

$$\left(1 + a_1 \frac{dt}{2}\right) u^{n+1} - a_2(dt)^2 \theta \Delta u^{n+1} = 2u^n$$
$$+ a_2(dt)^2(1-\theta)\Delta u^n + \left(a_1 \frac{dt}{2} - 1\right) u^{n-1} + (dt)^2 f^{n+1} \qquad (11)$$

where $f^{n+1} = f(u^{n+1})$.

Assuming that there are a total of $N-2$ interpolation points, $u^n(x)$ can be approximated by

$$u^n(x) = \sum_{j=1}^{N-2} \lambda_j^n \varphi(r_j) + \lambda_{N-1}^n x + \lambda_N^n. \qquad (12)$$

To determine the coefficients $(\lambda_1, \lambda_2, \cdots, \lambda_{N-1}, \lambda_N)$, the collection method is used by applying Eq. (12) at every point x_i, $i = 1, 2, \cdots N-2$. Thus we obtain

$$u^n(x_i) = \sum_{j=1}^{N-2} \lambda_j^n \varphi(r_{ij}) + \lambda_{N-1}^n x_i + \lambda_N, \tag{13}$$

where $r_{ij} = \sqrt{(x_i - x_j)^2}$. The additional conditions due to Eq. (5) can be written as

$$\sum_{j=1}^{N-2} \lambda_j^n = \sum_{j=1}^{N-2} \lambda_j^n x_j = 0. \tag{14}$$

Writing Eq. (13) together with Eq. (14) in a matrix form

$[u]^n = A[\lambda]^n,$

where $[u]^n = [u_1^n, u_2^n, \cdots, u_{N-2}^n, 0, 0]$, $[\lambda]^n = [\lambda_1^n, \lambda_2^n, \cdots, \lambda_N^n]$ and

$A = [a_{ij}, 1 \le i, j \le N]$ is given as follows:

$$A = \begin{bmatrix} \varphi_{11} & \cdots & \varphi_{1(N-2)} & x_1 & 1 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \varphi_{(N-2)1} & \cdots & \varphi_{(N-2)(N-2)} & x_{N-2} & 1 \\ x_1 & \cdots & x_{N-2} & 0 & 0 \\ 1 & \cdots & 1. & 0 & 0 \end{bmatrix}. \tag{15}$$

Assuming that there are $p < N-2$ internal points and $N-2-p$ boundary points, then the $N \times N$ matrix A can be split into:

$A = A_d + A_b + A_e$, where

$A_d = [a_{ij}$ for $(1 \le i \le p, 1 \le j \le N)$ and 0 elsewhere],

$A_b = [a_{ij}$ for $(p+1 \le i \le N-2, 1 \le j \le N)$ and 0 elsewhere], $\tag{16}$

$A_e = [a_{ij}$ for $(N-1 \le i \le N, 1 \le j \le N)$ and 0 elsewhere].

Using the notation LA to designate the matrix of the same dimension as A and containing the elements $\hat{a}_{ij}$ where $\hat{a}_{ij} = La_{ij}$, $1 \le i, j \le N$, then Eq. (11) together with the boundary conditions Eq. (9) can be written, in matrix form as

$$B[\lambda]^{n+1} = C[\lambda]^n + \left(a_1 \frac{dt}{2} - 1\right) u^{n-1} + (dt)^2 \cdot f^{n+1} + [H]^{n+1}, \tag{17}$$

where

$$C = 2A_d + a_2(1-\theta)(dt)^2 \Delta A_d,$$

$$B = \left(1 + a_1 \frac{dt}{2}\right)A_d - a_2\theta(dt)^2 \Delta A_d + A_b + A_e,$$

$$[H]^{n+1} = \begin{bmatrix} 0 & \cdots & 0 & h_{p+1}^{n+1} & \cdots & h_{N-2}^{n+1} & 0 & 0 \end{bmatrix}^T,$$

$$[f]^{n+1} = \begin{bmatrix} f_1^{n+1} & \cdots & f_p^{n+1} & 0 & \cdots & 0 \end{bmatrix}^T.$$

Eq. (17) is obtained by combining Eq. (11), which applies to the domain points, while Eq. (9) applies to the boundary points.

If $n = 0$ the Eq. (17) has the following form:

$$B[\lambda]^1 = C[\lambda]^0 + \left(a_1 \frac{dt}{2} - 1\right)u^{-1} + (dt)^2 \cdot [f]^1 + [H]^1. \tag{18}$$

To approximate u^{-1}, the second initial condition can be used. We discretise the second initial condition as

$$\frac{u^1(x) - u^{-1}(x)}{2 \cdot dt} = g_2(x), x \in \Omega \quad . \tag{19}$$

Writing Eq. (18) together with Eq. (19) we have

$$\left(B + \left(1 - a_1 \frac{dt}{2}\right)A_d\right)[\lambda]^1 = C[\lambda]^0 + (2 - a_1 \cdot dt)dt[G_2] + (dt)^2 \cdot [f]^1 + [H]^1, \tag{20}$$

where $[G_2] = \begin{bmatrix} g_2^1 & \cdots & g_2^p & 0 & \cdots & 0 \end{bmatrix}^T$. Together with the initial condition Eq. (8) and Eq. (17), we can get all the $\lambda's$, thus we can get the numerical solutions.

Since the coefficient matrix is unchanged in time steps, we use the LU factorization to the coefficient matrix only once and use this factorization in our algorithm.

Remark 1. Although Eq. (17) is valid for any value of $\theta \in [0,1]$, we will use $\theta = \frac{1}{2}$ (the famous Crank-Nicolson scheme).

3   Numerical Examples

In this section we present some numerical results to test the efficiency of the new scheme for solving nonlinear telegraph problems.

Example 1. In this example, $a_1 = 2$, $a_2 = 1$, $\alpha = 1$, $\beta = 0$ and $\gamma = -1$ Eq. (1) takes the following form:

$$u_{tt} + 2u_t = u_{xx} + u^3 - u, \quad 0 < x < 1, t > 0,$$

the analytical solution by [6] is

$$u(x,t) = \frac{1}{2} + \frac{1}{2}\tanh\left(\frac{x}{8} + \frac{3t}{8} + 5\right).$$

We extract the boundary function and the initial conditions from the exact solution.

We use the radial basis functions (TPS, IMQ) for the discussed scheme. These results are obtained for $dx = 0.01$ and $dt = 0.001$. The L_∞, L_2 errors and Root-Mean-Square (RMS) of errors are obtained in Table (1) for t=0.5, 1.0, 1.5, 2.0, 3.0. The results obtained in the table show the very good accuracy.

Table 1. Numerical errors using TPS (m=2) and IMQ (c=0.01) represented at different times, where dx =0.01, dt =0.001, θ =1/2

t	L_∞ - error		L_2 - error		RMS	
	TPS	IMQ	TPS	IMQ	TPS	IMQ
0.5	1.014×10^{-5}	3.599×10^{-6}	7.434×10^{-6}	3.077×10^{-6}	7.325×10^{-6}	3.032×10^{-6}
1.0	1.668×10^{-5}	6.375×10^{-7}	1.209×10^{-5}	5.437×10^{-7}	1.192×10^{-5}	5.357×10^{-7}
1.5	1.087×10^{-5}	1.156×10^{-6}	7.769×10^{-6}	9.843×10^{-7}	7.655×10^{-6}	9.698×10^{-7}
2.0	3.633×10^{-6}	6.750×10^{-7}	2.606×10^{-6}	5.754×10^{-7}	2.568×10^{-6}	5.670×10^{-7}
3.0	2.159×10^{-6}	4.154×10^{-7}	1.580×10^{-6}	3.541×10^{-7}	1.557×10^{-6}	3.489×10^{-7}

We also give the space-time graph of L_∞ error for T=3 with IMQ (c=0.01) radical basis function in Fig. 1.
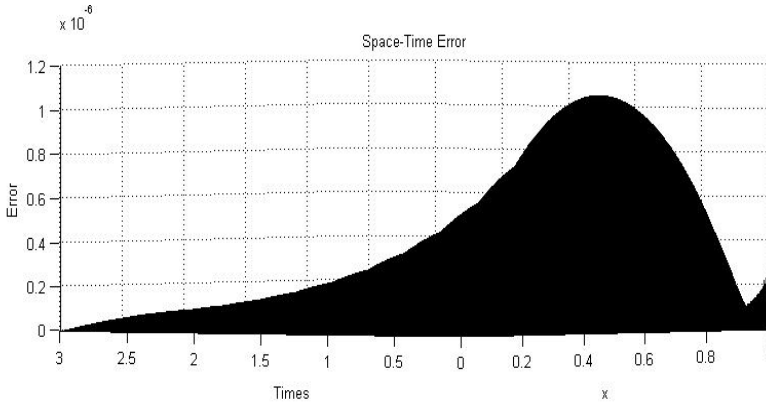
Fig. 1. The space-time graph of L_∞ error for T=3 using IMQ (c=0.01) as the radical basis function with $dx = 0.01$ and $dt = 0.001$ for example 1

We also give the analysis of the parameter c in IMQ for the results.

In Table (2), the L_∞, L_2 errors and RMS of errors with different c at time t=3 are presented.

Table 2. Numerical errors using IMQ represented with different parameter c at time t=3, where $dx = 0.01$, $dt = 0.001$, $\theta = 1/2$

c	L_∞ - error	L_2 - error	RMS
0.010	4.154×10^{-7}	3.541×10^{-7}	3.489×10^{-7}
0.030	2.178×10^{-6}	1.591×10^{-6}	1.568×10^{-6}
0.050	2.163×10^{-6}	1.582×10^{-6}	1.559×10^{-6}
0.070	2.160×10^{-6}	1.581×10^{-6}	1.557×10^{-6}
0.090	2.159×10^{-6}	1.580×10^{-6}	1.557×10^{-6}
0.093	0.001	2.485×10^{-4}	2.449×10^{-4}
0.095	8.991	1.959	1.930

Example 2. In this example, we consider the nonlinear telegraph equation takes the following form:

$$u_{tt} + u_t = 2u_{xx} + u^3 - 2u, \quad 0 < x < 1, t > 0,$$

the analytical solution by [6] is

$$u(x,t) = \frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2} \coth\left(x + \frac{3t}{2} + 5\right),$$

we extract the boundary function and the initial conditions from the exact solution.

We use the radial basis functions (TPS, IMQ) for the discussed scheme. These results are obtained for $dx = 0.01$ and $dt = 0.001$.

In Fig. 2 we give the space-time graph of L_∞ error for T=2 with TPS (m=2) radical basis function.
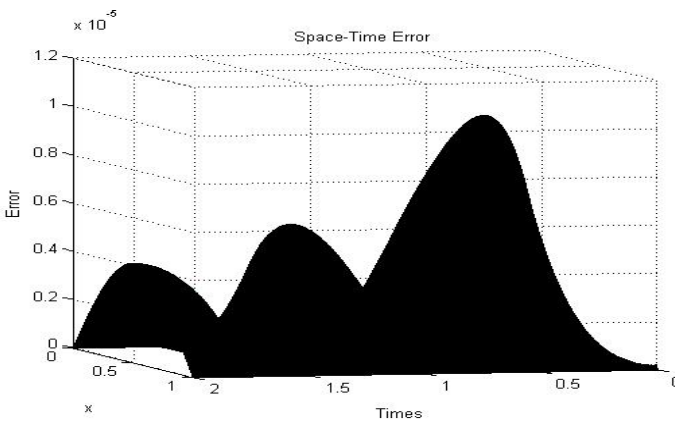
Fig. 2. The space-time graph of L_∞ error for T=2 using TPS (m=2) as the radical basis function with $dx = 0.01$ and $dt = 0.001$ for example 2

The L_∞, L_2 errors and RMS of errors are obtained in Table (3) for t=1, 2, 3, 4, 5. The results obtained in the table show the very good accuracy.

Table 3. Numerical errors using TPS (m=2) and IMQ (c=0.03) represented at different times, where dx =0.01, dt =0.001, θ =1/2

t	L_∞ - error		L_2 - error		RMS	
	TPS	IMQ	TPS	IMQ	TPS	IMQ
1	1.728×10^{-6}	2.199×10^{-6}	1.032×10^{-6}	1.340×10^{-6}	1.017×10^{-6}	1.320×10^{-6}
2	4.132×10^{-6}	2.302×10^{-6}	2.877×10^{-6}	1.610×10^{-6}	2.835×10^{-6}	1.587×10^{-6}
3	5.782×10^{-7}	2.399×10^{-6}	3.783×10^{-7}	1.662×10^{-6}	3.728×10^{-7}	1.637×10^{-6}
4	1.387×10^{-6}	1.133×10^{-6}	9.397×10^{-7}	8.134×10^{-7}	9.260×10^{-7}	8.014×10^{-7}
5	6.159×10^{-7}	1.563×10^{-7}	4.083×10^{-7}	9.434×10^{-8}	4.024×10^{-7}	9.296×10^{-8}

Similar to the previous example, we also give the analysis of the parameter c in IMQ for the results.

In Table (4), the L_∞, L_2 errors and RMS of errors with different c at time t=5 are presented.

Table 4. Numerical errors using IMQ represented with different parameter c at time t=5, where dx =0.01, dt =0.001, θ =1/2

c	L_∞ - error	L_2 - error	RMS
0.010	5.848×10^{-7}	4.512×10^{-7}	4.446×10^{-7}
0.030	1.563×10^{-7}	9.134×10^{-8}	9.296×10^{-8}
0.050	1.968×10^{-7}	1.191×10^{-7}	1.173×10^{-7}
0.070	2.046×10^{-7}	1.242×10^{-7}	1.224×10^{-7}
0.090	2.061×10^{-7}	1.252×10^{-7}	1.234×10^{-7}
0.095	0.394	0.090	0.089

4   Conclusion

In this paper, we proposed a numerical scheme to solve the nonlinear telegraph equation using the Kansa's method using TPS and IMQ radial basis functions. The numerical results given in the previous section demonstrate the good accuracy of this scheme and also see that the choice of parameter c is very important for the solution accuracy. The main advantage of this method over finite difference techniques is that, latter methods provide the solution of the problem on mesh points only.

References

1. Lai, S.Y.: The asymptotic theory of solutions for a perturbed telegraph wave equation and its application. Appl. Math. Mech. 43(7), 657–662 (1997)
2. Kolesov, A.Y., Rozov, N.K.: Parametric excitation of high-mode oscillations for a nonlinear telegraph equation. Sbornik Mathematics 191(7-8), 1147–1169 (2001)
3. Dehghan, M.: A numerical method for solving the hyperbolic telegraph equation. InterScience 24, 1080–1093 (2008)
4. Gao, F., Chi, C.: Unconditionally stable difference schemes for a one space dimensional linear hyperbolic equation. Appl. Math. Comput. 187, 1272–1276 (2007)
5. Alonso, J.M., Mawhi, J., Ortega, R.: Bounded solutions of second order semilinear evolution equations and applications to the telegraph equation. J. Math. Pures Appl. 78, 49–63 (1999)
6. Shang, Y.D.: Explicit and exact solutions for a class of nonlinear wave equations. Acta Math. Appl. Sinica 23, 21–30 (2000)
7. Fan, E.G., Zhang, H.Q.: The solitary wave solutions for a class of nonlinear wave equation. Chin. Phys. Soc. 46, 1245–1248 (1997)
8. Dehghan, M.: Parameter determination in a partial differential equation from the overspecified data. Math. Comput. Model. 41, 196–213 (2005)
9. Dehghan, M.: Implicit collocation technique for heat equation with non-classic initial condition. Int. J Non-Linear Sci. Numer. Simul. 7, 447–450 (2006)
10. Dehghan, M.: Finite difference procedures for solving a problem arising in modeling and design of certain optoelectronic devices. Math. Comput. Simulation 71, 16–30 (2006)
11. Liu, G.R., Gu, Y.T.: Boundary meshfree methods based on the boundary point methods. Eng. Anal. Bound. Elem. 28, 475–487 (2004)
12. Jiang, T.S., Li, M., Chen, C.S.: The Method of Particular Solutions for Solving Inverse Problems of a Nonhomogeneous Convection-Diffusion Equation with Variable Coefficients. Numerical Heat Transfer, Part A: Applications 61(5), 338–352 (2012)
13. Jiang, T.S., Jiang, Z.L., Jospeh, K.: A numerical method for one-dimensional time-dependent Schrodinger equation using radial basis functions. Internatioal Journal of Computational Methods (accepted, 2013)
14. Li, M., Jiang, T.S., Hon, Y.C.: A meshless method based on RBFs method for nonhomogeneous backward heat conduction problem. Engineering Analysis with Boundary Elements 34(9), 785–792 (2010)
15. Liu, G.R.: Mesh Free Method: Moving Beyond the Finite Element Methods. CRC Press, Florida (2002)
16. Liu, G.R., Liu, M.B.: Smooth Pariticle Hydrodynamics-a meshfree particle method. World Scientific, Singapore (2003)
17. Schaback, R.: Error estimates and condition numbers for radical basis function interpolation. J. Advances in Computational Mathematics 3(7), 251–264 (1995)
18. Wu, Z.M., Schaback, R.: Local error estimates for radical basis function interpolation of scattered data. J. IMA Journal of Numerical Analysis 13(1), 13–27 (1993)

Adaptive Replica Management Model
for Data-Intensive Application

Tian Tian[1], Liu Dong[2], and He Yi[1]

[1] Automation Department Beijing Oil Research Institute, Beijing 100000, China
[2] Academy of Equipment, Beijing 101416, China
oraclepla@126.com, {ld5m,grant.he}@163.com

Abstract. With the explosive increasing of data, data storage and management in data-intensive system are facing challenge. The paper put forward an adaptive replica management model for data-intensive applications. The demands for replica management in data-intensive application were firstly analyzed. Then, some concepts of replica management service were presented, where the operations about replica were described by services. Finally, the adaptive replica management model was proposed with the detailed implementations of replica management functions. The paper researched the replica management problem from a top of viewpoint. The presented model makes technical implementation guidance for replica management. And it is also a reference for data storage and management in data-intensive system.

Keywords: data, replica, management, data-intensive application.

1   Introduction

With the development of informationization and the increment of data, the applications of large-scale network present the characters of distribution, heterogeneousness, and intensiveness [1][2]. The journals Nature and Science published special issues of Big Data and Dealing with Data respectively, discussing the special topics in mass data processing and applications from internet, supercomputing, environmental science and other multi-aspects [3][4]. As one of the main contents in mass data storage and management, replica management technology is becoming a key research point in data-intensive application field.

At present, there have been some researches on replica management technology in data-intensive application. Megastore, Google's distributed memory system, and Spanner, the system suggested in recent years, treated replica management as a key technology, where chubby lock service and classical Paxos algorithm were used to coordinate and deploy replica [5]. Amazon's Dynamo adopted an improved distributed hashing algorithm to deal with data distribution and replication problem [6]. Besides, China Southeast University's Computer Science and Engineering College pointed out that more attention should be paid to the replica management strategy corresponding to decentralized cooperated data centers [7]. The mentioned literatures mainly introduced some specific technologies or presented individual

conceptions, and did not consider the replica management problem from a system point of view.

In this paper, we will analyze the replica management demands in data-intensive application in detail, and present the concept of replica management service. The main propose of the paper is to propose an adaptive replica management model which is adapted for different environments.

2   Demand of Replica Management

2.1   Characters of Data-Intensive System

Large-scale distributed systems often use data redundancy to improve reliability and performance. This brings the data explosive increment in both quantity and complexity, together with the problems and challenges for mass data storage and management [8][9].

• Dynamic and Scalability

The system for data-intensive application is mostly composed of general computers with various performances, which make the node failure becoming normality. The dynamic joining of a new node or the leaving of a failed node brings the difficulties to data management and data processing.

• Data complexity

Conventional structured data employs two-dimension relation table or <key, value> structure for storage. But in data-intensive system, the traditional method for storage and management is no longer valid, because of the coexisting of the structured data, semi-structured data, and unstructured data. We need new centralized management platform and technique to deal with different types of data.

• Decentralization

Data-intensive application usually needs the cooperating among decentralized data centers, which brings the high cost of data transmission and maintenance. We have to pay more attention to the data assignment and management strategy in order to decrease the data transmission among different centers and to save network cost.

2.2   Demand Analysis of Replica Management

According to the mentioned characters, we made a summarization of the replica management demands in data-intensive application:

• Scalability

The replica management system should take the mass replications in control. If the number of node and replica have changed, it should be ensured that the time for replica managing and processing be acceptable.

• Dynamic

The joining or leaving of a node should be dynamically solved. Different strategy for replica creating, locating, selecting, and consistency maintenance processes should also be dynamically adjusted according to user requirements and system parameters.

• Flexibility

Mass structured data, semi-structured data, and unstructured data should be managed in a common way. And the management should be transparent to user.

• High efficiency

When replica is being managed, the resource and time cost in layout and scheduling process must be concerned.

• Safety

The unsafe factors in management process should be monitored. For example, the user authority should be assigned for different operations.

2.3   Importance of Replica Management

The introducing of replica is to improve the usability and optimize the access performance of data. In data-intensive environment, the existence of massive data nodes, wide range distribution and multi-source heterogeneous data structure make the replica management be a challenging problem.

The factors that influence data usability include node failure rate, network connectivity, replica placement strategy, replica redundance rate, replica consistency strategy, replica location mechanism, replica selection mechanism, network load balancing, data transfer ability, and so on.

Replica provide an efficient way to improve data access performance. Replica placement strategy is an important factor involved. An appropriate placement method store frequently-used data in the local nodes or the nodes adjacent to local, which could decrease remote procedure call and reduce network flow. Another factor is the replica redundancy rate. If the redundancy rate is too low, data accessing time will become unacceptable. On the other hand, the high redundancy rate will increase maintainance cost. Replica location and selection strategy are also important factors that influence accessing performance. Good replica location and selection strategy balance system load, which could avoid the concentrating access request on a single node and reduce schedule time.

The system for data-intensive applications require replica management to adjust redundancy rate and placement strategy dynamically according to the system runtime states, so as to reduce network transfer cost. It is particularly important for the circumstances when a node fail and data have to be replicated to new nodes. The system also prefer the replica strategy, which could realize the balanced replica location according to the dynamically changed node capacity, performance and network parameters, so as to avoid network congestion and reduce access cost.

Replica management is an important aspect that influence the work performance of cluster systems.

2.4   Strategies of Replica Management

There have been many replica management strategies for different requirements, such as master-slave mode, hierarchical mode, peer-to-peer mode, graph-based mode.

• Master-slave mode

Master-slave replica management strategy uses one or several replicas as master replica, and the others as slave replicas [10]. Master replicas are in charges of managing replica information. If data are changed and update happens, slave replicas have to communicate with master replicas for updated data. Master-slave strategy is a simple method for its high efficiency in replica consistency maintenance. However, like most concentrated mechanisms, master-slave strategy has to be faced with the problems such as system bottleneck, single point failure and poor scalability.

• Hierarchical mode

Hierarchical replica management strategy organizes all the data replicas into several levels [11]. The replicas in different levels have different functions. And each replica can only communicate and exchange information with the replicas in its neighboring level. Hierarchical strategy can take advantage of the diversities among different data nodes so as to improve the system performance. However the coordination among the levels is necessary in order to reach consistent state.

• Peer-to-peer mode

In peer-to-peer (P2P) replica management strategy, a replica can communicate with any other replicas with no restraint [12]. This method is also referred to as topology independence replica management. The nodes in P2P strategy have high autonomy, which is the characteristic of P2P systems. However, excessive dispersion and autonomy increase the complexity of data consistence maintenances.

• Graph-based mode

Graph-based replica management strategy organizes replicas according to the topology of a specific graph [13]. Only the connected two replicas in the topology can communicate and exchange data. Tree structure and net structure are the two topologies usually adopted. Graph-based strategy can organize a large number of discrete nodes effectively so as to get an optimized maintenance. Moreover, group broadcast technology can be used to improve efficiency. However, if some parts of the topology fail, the graph will be divided into several independent sub-graphs, and communication path will be broken off for the inherent graph structure.

As the service oriented thought matures, more and more large-scale distributed systems, such as grid and cloud systems, provide their system management functions using the service form. This method assures the looser coupling, the stronger autonomy and flexibility of system functions. Replica management service is to provide replica creation, replica location, replica selection and replica consistency maintenance for users transparently. Since the replica management system in data-intensive applications requires more functions, it is significant to design adaptive replica management service so as to support big data application environment.

3   Replica Management Service

After the analysis of replica management demands in data-intensive applications, we present the concept of replica management service.

The service can be understood as the union of some functions. These functions can be accessed individually or collectively. And they can also compose a new complex application with other services [14].
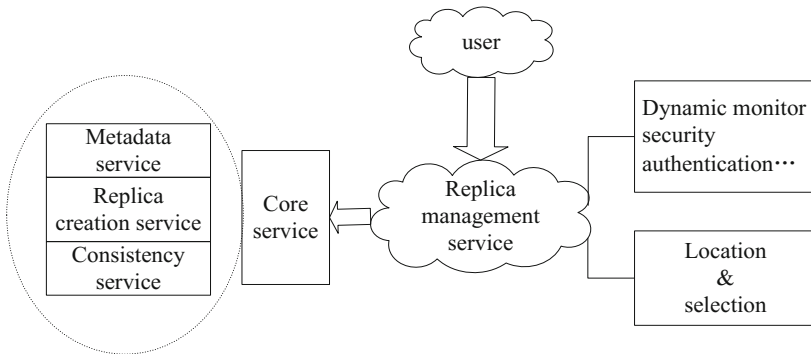
Fig. 1. Replica management service

In this paper, we define the replica management in data-intensive application as a service. It transparently provides replica creation, maintenance, location, selection, and other optimization functions for users (see Fig. 1).

● The core service

The service mainly includes metadata maintenance, replica creation, and consistency maintenance.

■ Metadata service is in charge of storing the metadata related to data object, such as the data location, creation time, update time, etc. These metadata characterize the data's attribute correctly.

■ Replica creation service includes replica redundancy, placement strategy, repeated data deletion, and other operations. The service dynamically creates, deletes, or schedules replica according to the changes of network and user requests.

■ Consistency service ensures the data consistency among replicas, and between the metadata and data object, in order to dynamically support node joining or leaving. Besides, it checks the inconsistent data and finishes the consistency task.

● Location and selection service

There may be lots of physical replicas, with the same logical data name, storing at different location in system. The location function decides the location of the physical replica according to the logical name. The selection function chooses the best replica to users, balancing the network load and avoiding congestion.

● There are also auxiliary services, such as dynamic monitor, security authentication, data transmission, and access recording.

■ Dynamic monitor service collects relevant information, such as network bandwidth, node utilization ratio, to provide references for replica management service.

■ Security authentication service is used to identify the status of users in order to prevent unauthorized tampering of useful data.

■ Data transmission service is in charge of evaluating the expenses in the scheduling and transporting of replica.

■ Access recording service collects the system information, such as access frequency, access time, data status, etc, so as to provide references for replica creating and deleting operations.

4   Adaptive Replica Management Model

As the features of the data-intensive environment are complicated and varied, on the one hand, the replica management method should be able to dynamically modulate the replica management strategy according to some parameters like network environment; on the other hand, the method should timely deal with the emergencies like the data reconstruction after node failure. Only the adaptive management could achieve the purposes to guarantee an acceptable data availability and access performance. Therefore, we propose an adaptive replica management model in data-intensive application (shown in Fig. 2).
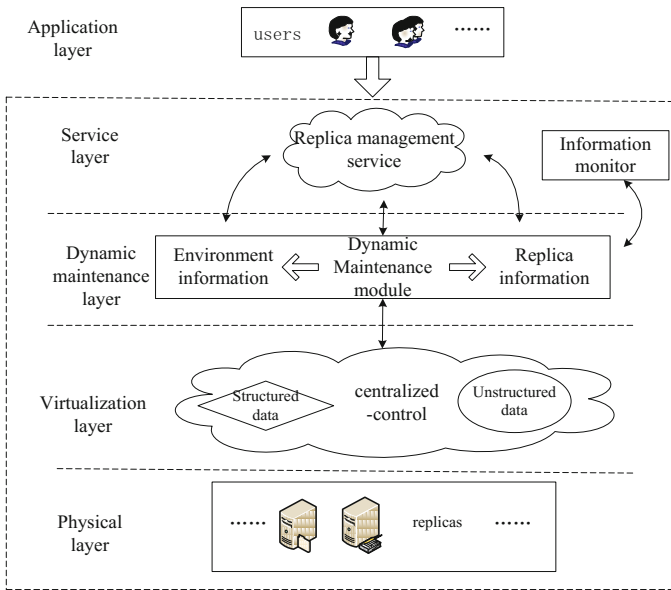
Fig. 2. Dynamic replica management model

The model consists of physical layer, virtualization layer, dynamic maintenance layer, service layer and application layer.
  ● Physical layer
Normally, the layer refers to the physical devices which store server file, physical replica, metadata, etc.
  ● Virtualization layer
The layer deals with mass structured data, semi- structured data, and unstructured data in a common strategy.
  ● Dynamic maintenance layer
The layer consists of environment information sub-service, replica information sub-service, and adaptive maintenance module. It conducts the replica management service by providing necessary parameters through the information in database and monitor.

■ Environment information sub-service: It periodically gathers system statuses, such as CPU idle ratio, memory usage, storage device free space, and network bandwidth, and records them in the *information database* (Info. DB), which will be used by the dynamic replica management.

■ Replica information sub-service: It stores the data information, such as logical data name, data size, physical location, time of creation, and replica access frequency, in the *data information database* (Data Info. DB). Users can use the replica information service to search for desired data and the closest sites in the system where the data are stored.

● Service layer (Fig. 3)

The layer is the implementation of the management services, including replica creation, consistency maintenance, location and selection sub-services, etc. Besides, it includes an information monitor module, which gathers the status information of network, tasks, applications, and data. The content of service layer is shown in Fig. 3.

■ Replica creation sub-service: It dynamically replicates, migrates, and deletes replica according to the changes of environment. It aims to reduce data access time, promote system stability, and improve the usage efficiency of storage device.

■ Replica consistency maintenance sub-service: It keeps data being consistent with duplicates stored in distributed nodes. When one data in a node is updated, it will notify the other nodes that have the same data to update to the newest version.

■ Location and selection sub-service: It gathers relevant information from the data information database to determine which sites are better to download data.

Replica management service dynamically adjusts replica maintenance mechanisms according to the parameters supplied by information monitor. It gives the conveniences for users to parallel download the same desired data by dynamically adjusting download speeds according to the network bandwidths between server nodes and client nodes, and balancing site loadings. The information monitor service includes recording sub-service and monitoring sub-service.

■ Recording sub-service: It provides data or file information before data downloading or adjusting.

■ Monitoring sub-service: It examines the variations during job processing.

● Application layer

The layer is the interface between users and replica management service. Usually, human-machine interactive interfaces are necessary for users to submit or download data.
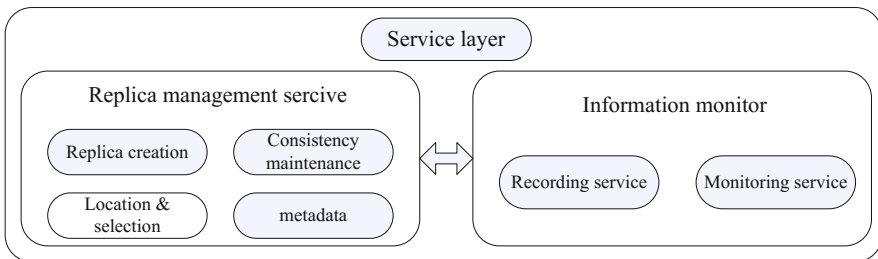
Fig. 3. Service layer

The work-flow of the adaptive replica management is shown in Fig. 4.
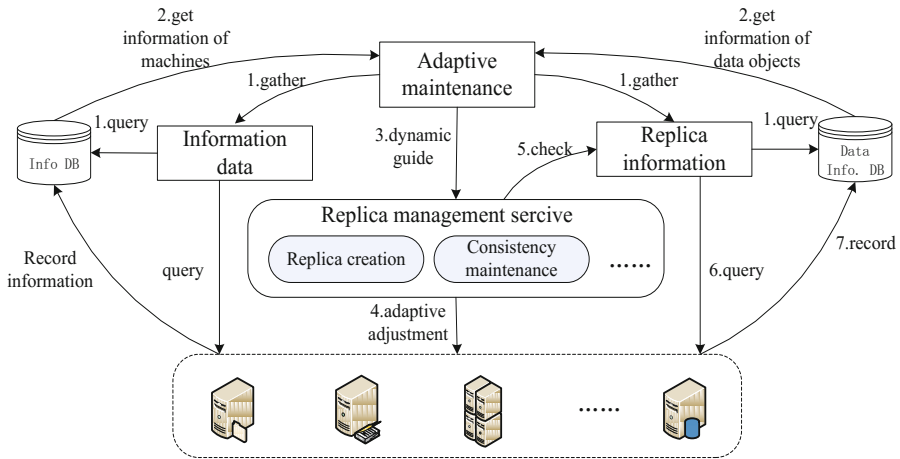
Fig. 4. Adaptive management flow

(1) Before applying the sub-service of replica creation and consistency maintenance, the adaptive maintenance module gathers system state information and data parameters from environment information sub-service and data information sub-service.

(2) Environment information sub-service and data information sub-service get system states and data parameters from environment information database and data information database respectively.

(3) System states and data parameters provide dynamic guidance to the replica management service.

(4) When the replica management service decides to make strategy adjustment, it directly applies the replication creation or other sub-services to perform the tasks adaptively.

(5) After the adjustment is finished, the replica management service checks data information sub-service

(6) Data information sub-service queries the system parameters that have been updated last.

(7) Then, all the latest information will be recorded into the data information database.

5   Conclusions

In this paper, we analyzed the characters of data-intensive application, introduced the replica management demands, and defined the functions of replica management in the form of service. The most important work is the adaptive replica management model for data-intensive application, whose components, functions, and work-flow were

represented in detail. The model provides guidance for the forward researches in replica management techniques, such as replica creation, location, selection, and consistency maintenance technologies.

References

1. Hai, Z., Zehua, Z., XueJie, Z.: A Dynamic Replica Management Strategy Based on Data Grid. In: 2010 9th International Conference on Grid and Cooperative Computing (GCC), Nanjing, China, November 1-5, pp. 18–23 (2010)
2. Wen-Hao, L., Yun, Y., Jinjun, C., Dong, Y.: A Cost-Effective Mechanism for Cloud Data Reliability Management Based on Proactive Replica Checking. In: 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), Ottawa, ON, May 13-16, pp. 564–571 (2012)
3. Lynch, C.: How do your data grow? Nature 455, 28–29 (2008)
4. Dealing with Data. Science 331 (February 2011), `http://www.sciencemag.org`
5. Peng, L.: Cloud Computing, 2nd edn., pp. 43–54. Publishing house of electronics industry, Beijing (2011)
6. DeCandia, G., Hastorun, D., et al.: Dynamo: Amazons Highly Available Key-value Store. In: Proceedings of Twenty-First ACM SIGOPS Symposium on Operating Systems Principles, pp. 205–220. ACM Press, New York (2007)
7. Junzhou, L., Jiahui, J., Aibo, S.: Cloud computing: architecture and key technologies. Journal on Communications 32(7), 3–21 (2011)
8. Bao, G., Chaojia, Y., Hong, Z.: The Model of Data Replica Adjust to the Need Based on HDFS Cluster. In: 2012 Fifth International Conference on Business Intelligence and Financial Engineering (BIFE), Lanzhou, China, August 18-21, pp. 532–535 (2012)
9. Amudhavel, J., Vengattaraman, T., Saleem Basha, M. S., Dhavachelvan, P.: Effective Maintenance of Replica in Distributed Network Environment Using DST. In: 2010 International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom), Kottayam, October 16-17, pp. 252–254 (2010)
10. Kistler, J., Satyanarayanan, M.: Disconnected Operation in the Coda File System. ACM Transaction on Computer Systems 10(1), 3–25 (1992)
11. Kubiatowicz, J., Bindel, D., Chen, Y., et al.: OceanStore: An Architecture for Global-scale Persistent Storage. ACM SIGARCH Computer Architecture News 28(5), 190–201 (2000)
12. Dahlin, M., Gao, L., Nayatc, A., et al.: PRACTI Replication for Large-scale Systems. Technical Report. University of Texas at Austin, Austin (2004)
13. Saito, Y., Karamanoli, C.: Replica Consistency Management in the Pangaea Wide-area File System. Technical Report. HP Labs (2002)
14. Linthicum, D.S.: Cloud Computing & SOA. Post and Telecommunication Press, Beijing (2011)

Study of EMC Problems with Vehicles

Hong Zhao[1,2], Goufeng Li[1], Ninghui Wang[1],
Shunli Zheng[2], Lijun Yu[2], and Yongli Chen[3]

[1] School of Electrical Engineering, Dalian University of Technology, Dalian, Liaoning, China
[2] Dalian Institute of Product Quality Supervision & Inspection, Dalian, Liaoning, China
[3] Dalian Boiler and Pressure Vessel Inspection Institute, Dalian, Liaoning, China
zhaohong@dlzjs.org

Abstract. Along with growing application of vehicle electronics, EMC problems with vehicles are getting more and more attention. In this paper overview is made about recent researches on EMC problems and the countermeasures, especially the problems relating to motor drive system, ignition system, wiper motor, entire vehicle and common vehicle electronics, also the problems with EMC tests of entire vehicle and some parts. Then briefly described are some relevant issues in EMC technology research for vehicles, as about the development of EMC standard system and the research on EMC requirements in the certification management of vehicle electronics.

Keywords: Electromagnetic compatibility, Vehicle electronics, Electric vehicle, Drive system, Ignition system, Wiper motor.

1   Introduction

In the development of vehicle technologies nowadays, the trend of electronization has gotten increasingly noticeable, and the costs of vehicle electronics have been accounting for more and more in the total of finished vehicle. Generally speaking, vehicle electronics are electronic control systems comprised of sensors, microprocessors, actuators, electronic components and parts. For application, those systems can be electronic control unit, safety control system, auto-body control system, entertainment and information system, etc [1]. Electronics application has been proved remarkably effective in improving the dynamic, economic and safety performances of vehicles, meanwhile leading to increased complexity of electromagnetic environment around vehicles, thus making the problem of electromagnetic compatibility (EMC) more and tougher to handle [2].

Chamon and others have studied the dielectric effects in EMC experiments for automotive vehicles [3]. By contrast, there have been several research subjects receiving more attention, involving EMC problems and countermeasures for drive system [4-8], ignition system [9-11], wiper motor [12,13], entire vehicle and common vehicle electronics [14-20], also the problems with EMC tests of entire vehicle and some parts [21-25]. For some relevant issues about vehicle EMC technology there are also researches having been made, as the status for vehicle EMC test standard

system building and research on EMC requirements in certification of automotive electronics [26-32].

2   Research on EMC Problem with Vehicle Drive System

From the review of the studies on EMC problem with drive system of electric vehicle (EV) [4], there are such issues requiring special notice as anti-interference performance of the control circuit and the troublesome electromagnetic interference (EMI) from high power or large current equipment. Further study of this subject will focus mainly on the following aspects.

1) Research on mechanism analysis and anti-interference methods for system EMI.
2) Research on EMC problems with DC/AC inverters or DC/DC convertors.
3) Research on conducted emissions of common-mode currents.
4) Design of filters.

Wang Bingyuan and others have studied the problem of radiated emission control for electric drive system of pure electric vehicle (PEV) [5]. On being aware of the nature of the radiation source, appropriate dealing with common-mode emission is the key to the radiated emission level control, and attention must be paid to the suppression of radiated emission from cables. Effective measures for radiated emission suppression may involve the introduction of filters and proper grounding and shielding of electronics.

According to existing experiences on EMC tests of entire EVs, most prototypes have to go repeatedly through rectification process to meet national standard provisions, with motor drive systems being the main EMI sources [6]. Given EV's specific characteristics, EMC performance of its drive system can be effectively improved by means of interference source suppressioin, system grounding, electromagnetic shielding, reasonable system layout and designing for anti-interference performances of the controller power source and the control panel.

Analyses have been made about how to reduce inductive and capacitive couplings in the wiring inside an EV [7], and about the sources and paths of common-mode interferences from within the drive system, with concrete measures for interference suppression suggested. Experiments have proven that by lowering the conducted common-mode currents in drive system, magnetic radiation emission from inside the EV can be effectively reduced, meanwhile the magnetic radiation from power lines can be suppressed through appropriate shielding.

In the analysis of conducted interference within motor windings, such influence must be taken into account as the noticeable distributed capacitances between the windings as well as between the winding and the motor enclosure. For the EMC modeling of direct current motor windings, Wang Xudong and others have made experimental measurements of the winding resistances, which then can be taken equivalent to a lumped parameter circuit [8]. To obtain the most optimal circuit parameters, ant colony algorithm with positive feedback mechanism has been utilized for the intelligent search and global optimization of the circuit parameters.

3   Research on EMC Problem with Vehicle Ignition System

Generated from starting the ignition system, high-voltage transient electromagnetic pulses can bring obvious electromagnetic interference on electronic devices inside the vehicle, meanwhile releasing strong electromagnetic radiation outside [9]. Circuit model for each part in ignition system has been built, then distribution parameters within each model can be determined from measurements, analytical method or finite element method and then integrated into equivalent circuit model of ignition system for system simulation to conduct, thus the effects of conducted electromagnetic interference from ignition system can be reasonably predicted. This method is verified by the measurement and simulation results in time domain as well as in frequency domain.

Wang Quandi and others have pointed the electromagnetic radiation from starting ignition system as the main cause resulting in most of misoperations of electronic control units [10]. Forming mechanism of EMI from ignition system has been analyzed, as well as the feasibility for photoelectric isolation technology to be utilized in EMI suppression. A testifying experiment has been carried out and proved the effectiveness of photoelectric isolation technology for electromagnetic noise suppression in the frequency range of 20MHz~100MHz.

Nowadays, with the growing application of vehicle electronics, electromagnetic interference has become the third largest form of vehicle pollution, following the exhaust emission and traffic noise. Among EMI sources in a vehicle electronics system, the spark plug discharge should be regarded as the strongest one [11]. Dynamic circuit model for simulation of spark plug discharge process has been built, and comparison between the results of model simulation and test measurement has been made, confirming the model simulation can be of great use in electromagnetic compatibility prediction for ignition sysem.

4   Research on EMC Problem with Vehicle Wiper Motors

It is of great significance and necessity to research on the issue of EMC problem with vehicle wiper motors [12]. Based on signal and system theories, the complicated problem of radiated interference from wiper motor can be examined by the calculation of the radiation field intensity by the transfer function of motor radiator system multiplying motor terminal voltage. This method has been proved effective in predicting and dealing with radiated EMI from wiper motor by an experiment conducted in frequency range of 30MHz~1000MHz.

According to He Jugang, reverse transient voltage up to several hundred volts may occur in a wiper motor during its operation and spark discharges take place between the brushes and commutator segments, which respectively constitute the main origin of conducted EMI and radiated EMI [13]. Forming mechanisms of electromagnetic interferences from vehicle wiper motors have been studied, as well as the test methods and suppression measures for such kinds of interferences.

5   Research on EMC Problems with Entire Vehicle and Common Vehicle Electronics

5.1   Literature Review

EMC test and design problems with new energy vehicles have been discussed, and several mandatory or voluntary standards at home and abroad for EMC tests of finished vehicles are listed by Du Jianfu [14]. Based on analyses about four common kinds of EMI transmitting paths, they have also pointed out several notable problems in the design of an entire vehicle. A case has been given about the engine throttle signal suffering electromagnetic radiation interference from high-voltage electrical components in a hybrid electric vehicle, then with anti-interference meausures properly taken; the throttle signals have restored stability.

By means of electronic design automation (EDA) simulation, Liu Ying and others have introduced a top-down EMC design flow for multi-vehicle communication system [15]. Taking some system involving four armored vehicles as example, a design flow from system level, device level to PCB (printed circuit board) level has been carried out, resorting to EDA simulation for quantitative evaluation and prediction about the impact of subsystem EMC performances on the whole system performance.

Ran Fangming and others have discussed some EMC problems with automotive CAN bus system design, making clear the sorts and sources of EMI firstly [16]. A new type of filter circuit has been introduced to lower common-mode interference and differential-mode interfercne that CAN bus system may possibly suffer from, features of the circuit being explained respectively.

Following a brief introduction of hardware makeup of electric vehicle AC charging point, Zou Qiang has described the requirements and methods of its EMC tests, nine kinds of tests included [17]. Then concisely outlined are differences among three of the current specifications for EMC tests of AC charging points.

In order to overcome the difficulty for the traditional method of conducted EMI measurement in diagnosing the EMI noise comprehensively, a conducted noise analyzer has been introduced, in which such functions have been integrated as noise measurement, modal separation, noise ayalysis and suppression [18]. By utilization of this analyzer in diagnosing the overall noise, quantities of common-mode noise and differential-mode noise can separately be ascertained, thus EMI filters may be accordingly designed to tackle the conducted EMI problems of electronics.

Based on EMC theory and the structure of EV, Pei Chunsong has made analyses of the electromagnetic compatibility of pure EV and also suggested some EMI suppression measures, whose effectivenesses have then been verified through specific EMC test [19].

Jin Songtao and others have analyzed the EMI problems with hybrid electric vehicle and explained the generating mechanism for electromagnetic interferences from HEV [20], then set forth some technical solutions for EMC problems with HEVs.

5.2   Case Study on EMI Suppression of an Onboard DVD Player

For some major automobile manufacturers around the world, the research on the electromagnetic compatibility of vehicle electronics is of great significance and thereby deserves great efforts, and this kind of research has become an important branch in EMC studying of electrical products. Therefore, aiming at providing technical services for vehicle manufacturers, Dalian institute of product quality supervision & inspection has planned for EMC tests of vehicle electronics. Illustrated in Fig.1 is a transient conduction immunity test arrangement for some piece of onboard electronics.
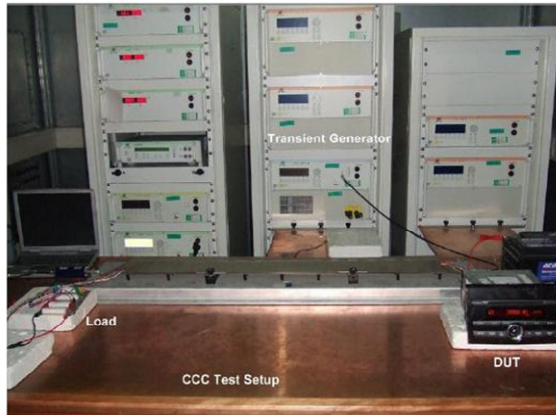
Fig. 1. Transient conduction immunity test arrangement for an onboard electrical equipment

Also we have made tests and analyses concerning EMC problems with some onboard audio & video products. For example, for the analysis about the causation of EMC problems with an onboard DVD, the probable disturbance sources can be identified layer by layer structurally, from inside to outside, see Fig. 2.

A: crystal oscillator/LSI/data line/address wire
⟶ main source of electromagnetic radiation

B: the connectives between PCBs
⟶ FFC wires/plug seats

C: outside parts
⟶ front and back panels/base plate

D: external connections
⟶ input and output cables
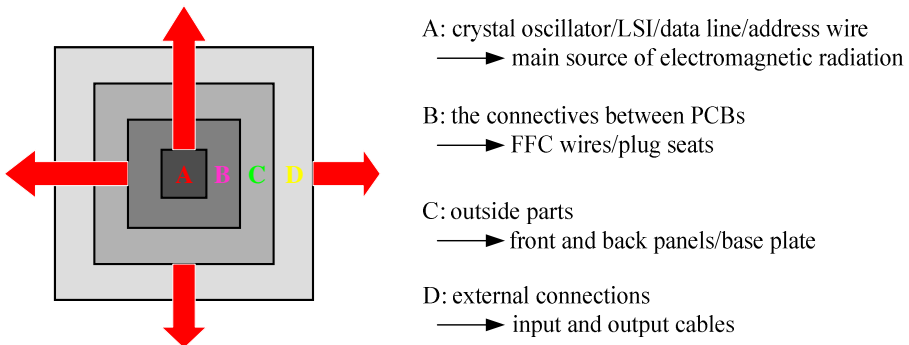
Fig. 2. Schematic diagram for production mechanism of electromagnetic radiation from onboard DVD player

Owing to the high frequency of crystal oscillator (up to hundreds of MHz) in operative condition, the radiation emission from it may be intense. With a sample of onboard DVD, radiation emission test has been made, in which two sets of emission limit have been involved.

(1) FCC (Federal Communications Commission) standard limit, as shown in Fig. 3 with the red solid line, which applies to enterprise products for exporting to North America.

(2) The limit set by product designer that is called "enterprise limit" here, 6dB below the FCC limit in this case, as shown in Fig. 3 with the red dotted line. Being a set of stricter limit, it has been set to ensure the compliance with FCC limit for the products from mass production.

For the radiation emission test of an onboard DVD in the frequency range of 30MHz~300MHz, test result has been presented as a data curve in Fig. 3. As shown in the curve, radiation emissions are more intense in the frequency range of 230MHz~270MHz, where there is an emission value less than the FCC limit for only 0.6dB at the frequency of 270MHz.

No.	FREQ	READING PEAK	ANT FACTOR	LOSS	GAIN	RESULT	LIMIT	MARGIN	ANTENNA	TABLE
	[MHz]	[dBuV]	[dB]	[dB]	[dB]	[dBuV/m]	[dBuV/m]	[dB]	[cm]	[DEG]

------ Horizontal ------

No.	FREQ	READING PEAK	ANT FACTOR	LOSS	GAIN	RESULT	LIMIT	MARGIN	ANTENNA	TABLE
1	189.030	47.6	17.4	1.8	27.3	39.5	43.5	4	100	64
2	229.531	49.7	18.3	2.0	27.1	42.9	46.0	3.1	399	88
3	243.031	47.2	18.5	2.0	27.0	40.7	46.0	5.3	399	104
4	270.031	50.6	19.7	2.1	27.0	45.4	46.0	.6	150	47
5	297.030	45.5	21.1	2.3	27.0	41.9	46.0	4.1	150	63

Fig. 3. Test data for electromagnetic radiation emission in between 30MHz~300MHz

With common electrical products with PCBs, the main sources of electromagnetic radiation emission may usually include the crystal oscillators, LSI (large scale integration), data lines and address wires on PCB. So when it comes to the limit-exceeding emission problem with onboard DVD, most attention may be paid about the PCB and its elements to find out about the problem causes and take countermeasures accordingly.

By thorough analysis it has been found out that imperfect design for GNDs (digital ground) on both sides of PCB accounts for the limit-exceeding emission problem. Poor GNDs can lead to high impedances for back currents, higher harmonic currents can consequently not be discharged to reference ground. Therefore certain measures have been intentionally taken, especially to increase the connectivity among the GNDs. With the connectivity among GNDs considerably improved, the back current impedance can be lowered, and the harmonic current noises can be easily discharged

to reference ground, thus electromagnetic radiation emission from PCB can be effectively suppressed. Final test afterwards the implementation of countermeasures turns out an expected result, as shown in Fig. 4.
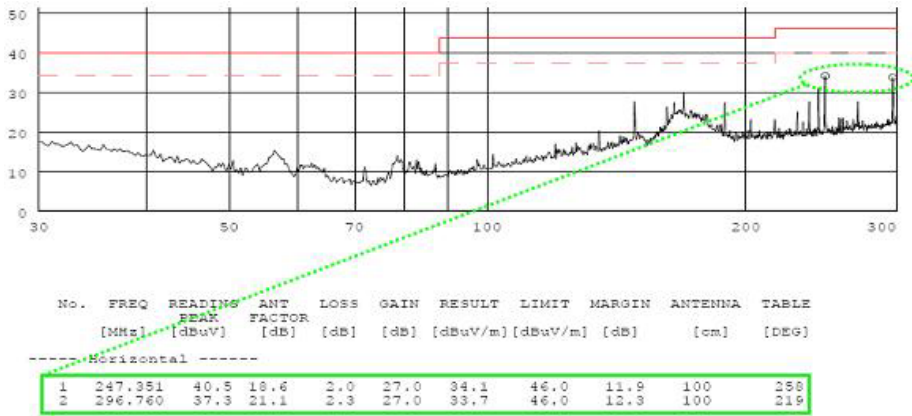
Fig. 4. Electromagnetic radiation test result after the countermeasure has been taken

6   Research on the Problems with EMC Tests of Entire Vehicle and Some Parts

For the remote control of operative condition of entire vehicle in its radiation immunity test, Li Xiao has designed the electronic control unit in an automated driving system based on PLC technology [21], with data communication carried out through optical fiber transmission and control signals transmitted to the pneumatic controller by PLC control unit receiving instructions from the upper computer.

Wu Qiong and Deng Junyong have made brief introduction about the radar-wave immunity test, suggesting it to be an effective means for the anti-interference capability verification for automotive components [22], and comparison has been made between radar-wave immunity test and the common radiation-field immunity tests in ways of test setup, calibration and signal modulation.

Deng Lingxiang and others have made analyses about the test devices and test methods of conducted and radiated emission tests for automotive parts [23]. Corresponding test circuits have been designed, and measurement results have been gained from the power line conducted emission test and horizontal polarization radiated emission test. The test means of anechoic chamber and that of large current injection have been separately taken for the radiation field immunity test and then compared with each other for the advantages and disadvantages. General principles for the EMC tests of automotive parts are then summarized.

Tan Jianrong has made a brief introduction of national standards for EMC tests of automotive instruments [24]. EMC test methods for such instruments have been illustrated as well as the problems that may occur during the tests and their probable reasons.

With the experience in EMC test for tire pressure monitoring system according to GB/T 21437.2, Sun Chengming and others have introduced the result and made analysis for a load dump transient conduction immunity test [25].

7   Research on the Development of EMC Test Standard System

Based on their rich experiences on EMC tests of EVs, Qin Yanming and his colleagues have analyzed the imperfections of current EMC test standards(see Table 1) in testing operability and consistency control of test results [26], then made further discussion about the future development in improving EMC test standard system for EV industry.

Table 1. Domestic fundamental EMC test standards for electric vehicle

Domestic standard number	Test type	Corresponding international standard	Application range
GB 14023-2006	Disturbance test	CISPR 12: 2005	Entire vehicle & parts
GB/T 18387-2008	Disturbance test	SAE J551-5 JAN2004	Entire vehicle & parts
GB/T 18655-2010	Disturbance test	CISPR 25: 2008	Entire vehicle & parts
GB/T 17619-1998	Immunity test	95/54/EC (1995)	Onboard parts
GB/T 21437.2-2008	Disturbance & Immunity test	ISO 7637-2: 2004	Onboard parts
GB/T 19951-2005	Immunity test	ISO 10605: 2001	Entire vehicle & parts
GB/T 22630-2008	Disturbance & Immunity test	2004/104/EC	Onboard parts

Li Tiehua has systematically analyzed the EMC standard system of automotive electronics, with applicable scope of each standard been described [27]. The status of the EMC standards of automotive electronics in China and some advices for improving the EMC standard system are presented.

With a view to the necessities of electrical safety, electromagnetic compatibility, braking system reliability and passenger protection, Liu Hui has introduced the EC directive and ECE regulations on the specific requirements of EVs [28]. Also described is the trend of standardization planning in European Union on relative technologies related to EVs.

Xiang Yunxiu and Qin Kaiyu have briefly introduced the automobile EMC standard system and the construction and verification of automobile EMC laboratory [29]. Advices have been presented about the improvement of EMC standard system and the establishment of automobile EMC laboratories.

Fang Cheng and Huang Xupeng have analyzed the automotive electromagnetic environment concerning EMI sources both inside and outside the vehicle [30]. Automotive EMC standards have been briefly introduced, along with some solutions for automotive EMC problems proposed.

Through the analysis of relative literature, technical information and research reports, Lan Bo and others have described the situation about automotive EMC technology research in view of automotive electromagnetic environment, international organizations and standards for automotive EMC, and testing technologies [31].

In such countries or regions as America, Japan and the European Union, EMC requirements have been included into quality certification management system for vehicle electronics [1]. Considering domestic conditions of EMC certification of vehicle electronics, Wang Zheng has suggested the necessity and urgency of EMC certification related technology researches, with the EMC test standard system building being most concerned.

Yang Yanzhang and others have analyzed and discussed the requirements of Ford Motor Company's standard EMC-CS-2009.1, concerning mainly the test projects, test methods and test setups [32]. The new requirements for upgrading the test capability of testing laboratory are also presented.

References

1. Wang, Z.: Research on EMC requirements in certification of automotive electronics. Safety & EMC (5), 9–11 (2010)
2. Jin, S., Ding, L., Liu, Q.: Importance of research on electromagnetic compatibility problem for vehicle. Bus & Coach Technology and Research (4), 1–5 (2011)
3. Chamon, L.E.A., Santos, C.H.G., dos Santos, K.M., et al.: Dielectric effects in electromagnetic compatibility experiments for automotive vehicles. In: 9th IEEE/IAS International Conference on Industry Applications, São Paulo, pp. 1–6 (2010)
4. Wang, H., Zhang, J., Yang, T.: Status quo of the EMC research on drive system in electric vehicle. Electronic Measurement Technology 34(6), 18–22, 59 (2011)
5. Wang, B., Zhang, J., Feng, H., et al.: Suppression of radiation emission on electric drive system for electric vehicles. Power Electronics 45(12), 75–77 (2011)
6. Dou, R., Wang, H., Gou, Y., et al.: Research of electromagnetic compatibility of motor drive system applied in electric vehicle. Journal of Tianjin Polytechnic University 30(6), 67–70 (2011)
7. Ji, S., Liu, D.: Study on conducted EMC of electric vehicle. Journal of Mechanical & Electrical Engineering 29(3), 359–364 (2012)
8. Wang, X., Liu, J., Yu, T.: Analysis of common mode EMI model of DC motor in electric vehicle. Power Electronics 45(12), 39–41 (2011)
9. Yu, J., Jia, J., Wang, Q., et al.: The simulation of conducted EMI in the ignition system for automotives. Journal of Chongqing University 34(1), 88–93 (2011)
10. Wang, Q., Liu, Q., Jia, J., et al.: The optical isolation method of suppressing electromagnetic interference of automotive ignition system. Journal of Chongqing University 34(2), 14–18 (2011)
11. Zheng, Y., Yu, J., Wang, Q., et al.: Dynamic circuit model of the spark plug for EMC prediction of ignition system. Transactions of China Electrotechnical Society 26(2), 8–13 (2011)

12. Wang, Q., Li, F., Zhou, S., et al.: Prediction model on the radiated electromagnetic interference of wiper motor in automobile. Journal of Chongqing University 33(4), 26–30 (2010)
13. He, J., Li, X., Cai, H., et al.: Research on electromagnetic interference testing and suppression methods for automotive wiper motor. Auto Electric Parts (10), 46–48, 52 (2010)
14. Du, J., Fu, D., Zhu, Z., et al.: EMC test and design of new energy vehicles. Agricultural Equipment & Vehicle Engineering (6), 22–25, 51 (2011)
15. Liu, Y., Xie, Y., Zhang, Y.: "Top-down" design flow and its applications in multi-vehicle communication system's EMC design. Journal of University of Electronic Science and Technology of China 39(5), 720–724 (2010)
16. Ruan, F., Sun, S., Dlugosz, T., et al.: Some consideration on electromagnetic compitibility in CAN bus design of automobile. In: Asia-Pacific International Symposium on Electromagnetic Compatibility, Beijing, pp. 1431–1434 (2010)
17. Zou, Q.: The electromagnetic compatibility test research of electric vehicle (EV) AC charging point. Electronics Quality (5), 75–78 (2011)
18. Wu, P., Liu, Y., Deng, L., et al.: The application of conducted noise analyzer in the EMC of automotive electronics. Journal of Nanjing Normal University (Engineering and Technology Edition) 11(4), 16–19, 23 (2011)
19. Pei, C.: Electromagnetic compatibility analysis and EMI suppression on pure EV. Auto Electric Parts (10), 59–63 (2011)
20. Jin, S., Liu, Q.: Electromagnetic compatibility technology research for hybrid electric vehicles. Bus Technology and Research (2), 45–47 (2010)
21. Li, X.: Design of PLC-based auto-driver electric control unit for EMC test. Process Automation Instrumentation 33(2), 68–72 (2012)
22. Wu, Q., Deng, J.: Testing method of immunity to radar waves for automotive components. Safety & EMC (5), 41–42, 51 (2010)
23. Deng, L., Feng, Z., Chen, D., et al.: Case study on automotive parts in EMC testing. Journal of Nanjing Normal University (Engineering and Technology Edition) 10(4), 12–16 (2010)
24. Tan, J.: Automotive instrument EMC testing and research. Automation & Instrumentation (4), 99–100 (2012)
25. Sun, C., Sun, Z., Chen, F., et al.: Test and analysis of transient conduction immunity for automobile load dump. Electronics Quality (10), 70–72, 75 (2010)
26. Qin, Y., Liu, Q., Li, B., et al.: Existing problems and development trend of EMC test standards of electric vehicle in China. Bus & Coach Technology and Research (1), 1–4, 13 (2012)
27. Li, T.: EMC standard system of automotive electronics and its application. Electronic Product Reliability and Environmental Testing 29(1), 50–54 (2011)
28. Liu, H.: Research of EU directives and regulations on electrical vehicle. Safety & EMC (3), 41–45 (2011)
29. Xiang, Y., Qin, K.: Electromagnetic compatibility of automobile: standards, situation and measures. Journal of Magnetic Materials and Devices (4), 10–13, 42 (2011)
30. Fang, C., Huang, X.: Automotive electromagnetic compatibility technology and its development trend. Automobile Parts (5), 75–77, 79 (2011)
31. Lan, B., Sun, Y., Zhang, Q.: The development status of automobile EMC research. Information Technology (8), 210–212, 215 (2010)
32. Yang, Y., Lin, B., Chen, Q., et al.: Improvement of measuring technology in EMC specifications of Ford motor company on electrical/electronic components and subsystems. Safety & EMC (4), 33–36 (2011)

RETRACTED CHAPTER: New Algorithms for Solving RFMLrR Circulant Linear Systems in Information

Jianhua Liu, Dengjie Wu, and Zhao-Lin Jiang[*]

Department of Mathematics, Linyi University, Linyi 276005, P.R. China
`jhliu@lyu.edu.cn, wdj20130313@163.com, jzh1208@sina.com`

Abstract. In this paper, fast algorithms for solving RFMLrR circulant linear systems are presented by the fast algorithm for computing polynomials. The extended algorithms are used to solve the RLMFrL circulant linear systems. Examples show the effectiveness of the algorithms.

Keywords: RFMLrR circulant matrix, linear system, Fast algorithm.

1   Introduction

The $f(x)$-circulant matrix has a wide application, especially on the generalized cyclic codes [1, 2]. The properties and structures of the $x^n + x - r$-circulant matrices, which are called RFMLrR circulant matrices, are better than those of the general $f(x)$-circulant matrices, so there are good algorithms for solving the RFMLrR circulant linear system. The advantage of the presented algorithm is that it can solve $AX = b$ whether the coefficient matrix is singular or nonsingular.

Definition 1. A row first-minus-last r-right (RFMLrR) circulant matrix with the first row $(a_0, a_1, \ldots, a_{n-1})$ is meant a square matrix over the complex field $\mathbf{C}$ of the form

$$A = RFMLRcir r(a_0, a_1, \ldots, a_{n-1})$$

$$= \begin{pmatrix}
a_0 & a_1 & \cdots & a_{n-2} & a_{n-1} \\
ra_{n-1} & a_0 - a_{n-1} & \cdots & a_{n-3} & a_{n-2} \\
ra_{n-2} & ra_{n-1} - a_{n-2} & \ddots & \cdots & a_{n-3} \\
\vdots & \cdots & \ddots & \ddots & \vdots \\
ra_1 & ra_2 - a_1 & \cdots & ra_{n-1} - a_{n-2} & a_0 - a_{n-1}
\end{pmatrix} \quad (1)$$

The paper starting on page 169 of this publication has been retracted because it does not differ sufficiently from the paper starting on page 59 of the same publication. One of the authors is the same. The Erratum to this chapter is available at DOI: 10.1007/978-3-642-53932-9_64

[*] Corresponding author.

In this paper, let r be a complex number and satisfy $r \neq 0$ and $r^{n-1} \neq -\frac{(n-1)^{n-1}}{n^n}$. It is easily verified that the polynominal $g(x) = x^n + x - r$ has no repeated roots in the complex field [3-5].

We define $\Theta_{(r,-1)}$ as the basic RFMLrR circulant matrix over $\mathbf{C}$, that is,

$$\Theta_{(r,-1)} = RFMrLRcirc_r fr(0,1,0,\ldots,0) \tag{2}$$

It is easily verified that the polynomial $g(x) = x^n + x - r$ is both the minimal polynomial and the characteristic polynomial of the matrix $\Theta_{(r,-1)}$, i.e., $\Theta_{(r,-1)}$ is nonsingular nonderogatory. In addition, $\Theta^n_{(r,-1)} = rI_n - \Theta_{(r,-1)}$.

In view of the structure of the powers of the basic RFMLrR circulant matrix $\Theta_{(r,-1)}$ over $\mathbf{C}$. it is clear that

$$A = RFMLRcirc_r fr(a_0, a_1, \ldots, a_{n-1}) = \sum_{i=0}^{n-1} a_i \Theta^i_{(r,-1)} \tag{3}$$

Thus, A is a RFMLrR circulant matrix over $\mathbf{C}$ if and only if $A = f(\Theta_{(r,-1)})$ for some polynomial $f(x)$ over $\mathbf{C}$. The polynomial $f(x) = \sum_{i=0}^{n-1} a_i x^i$ will be called the representer of the RFMLrR circulant matrix A over $\mathbf{C}$.

In addition, the product of two RFMLrR circulant matrices is a RFMLrR circulant matrix and the inverse of a nonsingular RFMLrR circulant matrix is also a RFMLrR circulant matrix. Furthermore, a RFMLrR circulant matrices commute under multiplication.

Definition 2. A row last minus-first r-left (RLMFrL) circulant matrix with the first row $(a_0, a_1, \ldots, a_{n-1})$ is meant a square matrix over the complex field $\mathbf{C}$ of the form

$$A = RLMFLcirc_r fr(a_1, a_2, \ldots, a_{n-1})$$

$$= \begin{pmatrix} a_0 & a_1 & \cdots & a_{n-2} & a_{n-1} \\ a_1 & a_2 & \cdots & a_{n-1} - a_0 & ra_0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ a_{n-2} & a_{n-1} - a_0 & \cdots & ra_{n-4} - a_{n-3} & ra_{n-3} \\ a_{n-1} - a_0 & ra_0 - a_1 & \cdots & ra_{n-3} - a_{n-2} & ra_{n-2} \end{pmatrix} \tag{4}$$

Lemma 1. Let $A = RFMLRcirc_r\, fr(a_0, a_1, ..., a_{n-1})$ be a RFML*r*R circulant matrix over $\mathbf{C}$ and $B = RLMFLcirc_r\, fr(a_{n-1}, a_{n-2}, ..., a_0)$ be a RLMF*r*L circulant matrix over $\mathbf{C}$. Then $B\hat{I}_n = A$ or $B = A\hat{I}_n$, where $\hat{I}_n$ is the backward identity matrix of the form

$$\hat{I}_n = LCirc(0, ..., 0, 1) \tag{5}$$

Lemma 2 ([3]). Let $\mathbf{C}[x]$ be the polynomial ring over a field $\mathbf{C}$, and let $f(x), g(x) \in \mathbf{C}[x]$. Suppose that the polynomial matrix $\begin{pmatrix} f(x) & 1 & 0 \\ g(x) & 0 & 1 \end{pmatrix}$ is changed into the polynomial matrix $\begin{pmatrix} d(x) & u(x) & v(x) \\ 0 & s(x) & t(x) \end{pmatrix}$ by a series of elementary row operations, then $(f(x), g(x)) = d(x)$ and $f(x)u(x) + g(x)v(x) = d(x)$.

Lemma 3. Let $A = RFMLRcirc_r\, fr(a_0, a_1, ..., a_{n-1})$ be a RFML*r*R circulant matrix over $\mathbf{C}$. Then A is nonsingular if and only if $(f(x), g(x)) = 1$, where $f(x) = \sum_{i=0}^{n-1} a_i x^i$ is the representer of A and $g(x) = x^n + x - r$.

In [4], for an $m \times n$ matrix A, a solution to the equation $AXA = A$ is called a generalized inverse of A. In addition, X satisfies $X = XAX$ then A and X are said to be semi-inverses $A^{\{1,2\}}$. In this paper we consider only square matrices A. If A has index 1, the generalized inverse X of A is called the group inverse $A^\#$ of A. Clearly A and X are group inverse if and only if they are semi-inverses and $AX = XA$.

2   Fast Algorithms for Solving the RFML*r*R and the RLMF*r*L Circulant Linear System

Consider the RFML*r*R circulant linear system

$$AX = b \tag{6}$$

where A is given in equation (**1**), $X = (x_1, x_2, ..., x_n)^T$, $b = (b_{n-1}, ..., b_1, b_0)^T$.

If A is nonsingular, then equation (**6**) has a unique solution $X = A^{-1}b$.

If A is singular and equation (**6**) has a solution, then the general solution of equation (**6**) is $X = A^{\{1,2\}}b + (I_n - A^{\{1,2\}}A)Y$, where Y is an arbitrary n-dimension column vector.

The key of the problem is how to find $A^{-1}b$, $A^{\{1,2\}}b$, $A^{\{1,2\}}A$, for this purpose, we first prove the following results.

Theorem 1. Let $A = RFMLRcirc_r fr(a_0, a_1, \ldots, a_{n-1})$ be a nonsingular RFMLrR circulant matrix of order n over $\mathbf{C}$ and $b = (b_{n-1}, \ldots, b_1, b_0)^T$. Then there exists a unique RFMLrR circulant matrix $C = RFMLRcirc_r fr(c_0, c_1, \ldots, c_{n-1})$ of order order n over $\mathbf{C}$ such that the unique solution of Equation (6) is the last column of $\mathbf{C}$, i.e.

$$X = (c_{n-1}, \ldots, c_1, c_0 - c_{n-1})^T$$

Proof. Since matrix $A = RFMLRcirc_r fr(a_0, a_1, \ldots, a_{n-1})$ is nonsingular, then by Lemma 3, we have $(f(x), g(x)) = 1$, where $f(x) = \sum_{i=0}^{n-1} a_i x^i$ is the representer of A and $g(x) = x^n + x - r$.

Let $B = RFMLRcirc_r fr(b_0 + b_{n-1}, b_1, \ldots, b_{n-1})$ be the RFMLrR circulant matrix of order n constructed by $b = (b_{n-1}, \ldots, b_1, b_0)^T$. Then the representer of B is $b(x) = (b_0 + b_{n-1}) + \sum_{i=1}^{n-1} b_i x^i$. Therefore, we can change the polynomial matrix $\begin{pmatrix} f(x) & 1 & 0 & b(x) \\ g(x) & 0 & 1 & 0 \end{pmatrix}$ into the polynomial matrix $\begin{pmatrix} 1 & u(x) & v(x) & c(x) \\ 0 & s(x) & t(x) & c_1(x) \end{pmatrix}$ by a series of elementary row operations. By Lemma 2, we have

$$\begin{pmatrix} u(x) & v(x) \\ s(x) & t(x) \end{pmatrix}\begin{pmatrix} f(x) \\ g(x) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} u(x) & v(x) \\ s(x) & t(x) \end{pmatrix}\begin{pmatrix} b(x) \\ 0 \end{pmatrix} = \begin{pmatrix} c(x) \\ c_1(x) \end{pmatrix}$$

i.e. $f(x)u(x) + g(x)v(x) = 1, u(x)b(x) = c(x)$

Substituting x by $\Theta_{(r,-1)}$ in the above two equations respectively, We have
$$f(\Theta_{(r,-1)})u(\Theta_{(r,-1)}) + g(\Theta_{(r,-1)})v(\Theta_{(r,-1)}) = I_n, u(\Theta_{(r,-1)})b(\Theta_{(r,-1)}) = c(\Theta_{(r,-1)})$$
Since $f(\Theta_{(r,-1)}) = A$, $g(\Theta_{(r,-1)}) = 0$ and $b(\Theta_{(r,-1)}) = B$, then

$$Au(\Theta_{(r,-1)}) = I_n \tag{7}$$

$$u(\Theta_{(r,-1)})B = c(\Theta_{(r,-1)}) \tag{8}$$

By equation (7) we know that $u(\Theta_{(r,-1)})$ is a unique inverse A^{-1} of A. According to equation (8) and the characters of the RFMLrR circulant matrix, we know that the last column of $\mathbf{C}$ is $X = (c_{n-1},\ldots,c_1,c_0 - c_{n-1})^T = A^{-1}b$. Since $AA^{-1}b = b$, then $A^{-1}b = (c_{n-1},\ldots,c_1,c_0 - c_{n-1})^T$ is the solution of equation (**6**). Since both A^{-1} and B are unique, then $A^{-1}B$ is also unique. So

$X = (c_{n-1},\ldots,c_1,c_0 - c_{n-1})^T = A^{-1}b$ is unique.

Theorem 2. Let $A = RFMLRcirc_r fr(a_0,a_1,\ldots,a_{n-1})$ be a singular RFMLrR circulant matrix of order n over $\mathbf{C}$ and $b = (b_{n-1},\ldots,b_1,b_0)^T$. If the solution of equation (**6**) exists, then there exist a unique RFMLrR circulant matrix

$$C = RFMLRcirc_r fr(c_0,c_1,\ldots,c_{n-1})$$

and a unique RFMLrR circulan matrix $E = RFMLRcirc_r fr(e_0,e_1,\ldots,e_{n-1})$ of order n over $\mathbf{C}$ such that $X_1 = (c_{n-1},\ldots,c_1,c_0 - c_{n-1})^T$ is a unique special solution of equation (**6**) and $X_2 = X_1 + (I_n - E)Y$ is a general solution of equation (**6**), where Y is an arbitrary n-dimension column vector.

Proof. Since $A = RFMLRcirc_r fr(a_0,a_1,\ldots,a_{n-1})$, then the representer of A is $f(x) = \sum_{i=0}^{n-1} a_i x^i$ and the characteristic polynomial of $\Theta_{(r,-1)}$ is $g(x) = x^n + x - r$. We can change the polynomial matrix $\begin{pmatrix} f(x) \\ g(x) \end{pmatrix}$ into the

polynomial matrix $\begin{pmatrix} d(x) \\ 0 \end{pmatrix}$ by a series of elementary row operations. Since A is singular, By Lemma 2 and Lemma 3, we know that $d(x)$ is a largest common factor, which is not equal to 1, of $f(x)$ and $g(x)$. Let $f(x) = d(x)f_1(x)$ and $g(x) = d(x)g_1(x)$, then $(f_1(x),g_1(x)) = 1$. Since $(d(x),g_1(x)) = 1$, we have $(f(x),g_1(x)) = (d(x)f_1(x),g_1(x)) = 1$. Since $(d(x),g_1(x)) = 1$ and $(f(x),g_1(x)) = 1$, we have $(f(x)d(x),g_1(x)) = 1$.

Let $B = RFMLRcirc_r fr(b_0 + b_{n-1},b_1,\ldots,b_{n-1})$ be the RFMLrR circulant matrix of order n constructed by $b = (b_{n-1},\ldots,b_1,b_0)^T$. Then the representer of B is $b(x) = (b_0 + b_{n-1}) + \sum_{i=1}^{n-1} b_i x^i$. Therefore, we can change the polynomial matrix

$$\begin{pmatrix} f(x)d(x) & \vdots & 1 & 0 & \vdots & d(x)b(x) & \vdots & d(x)f(x) \\ g_1(x) & \vdots & 0 & 1 & \vdots & 0 & \vdots & 0 \end{pmatrix}$$ into the polynomial

matrix $\begin{pmatrix} 1 & \vdots & u(x) & v(x) & \vdots & c(x) & \vdots & e(x) \\ 0 & \vdots & s(x) & t(x) & \vdots & c_1(x) & \vdots & e_1(x) \end{pmatrix}$ by a series of elementary row

operations. Then, by Lemma 2, we have

$$f(x)d(x)u(x) + g_1(x)v(x) = 1 \tag{9}$$

$$d(x)u(x)b(x) = c(x) \tag{10}$$

$$d(x)u(x)f(x) = e(x) \tag{11}$$

$$f(x)d(x)u(x)f(x) + g(x)f_1(x)v(x) = f(x) \tag{12}$$

Substituting x by $\Theta_{(r,-1)}$ in equations (10), (11) and (12), respectively, we have
$d(\Theta_{(r,-1)})u(\Theta_{(r,-1)})b(\Theta_{(r,-1)}) = c(\Theta_{(r,-1)}), d(\Theta_{(r,-1)})u(\Theta_{(r,-1)})f(\Theta_{(r,-1)}) = e(\Theta_{(r,-1)}),$
$f(\Theta_{(r,-1)})d(\Theta_{(r,-1)})u(\Theta_{(r,-1)})f(\Theta_{(r,-1)}) + g(\Theta_{(r,-1)})f_1(\Theta_{(r,-1)})v(\Theta_{(r,-1)}) = f(\Theta_{(r,-1)}).$

Since $f(\Theta_{(r,-1)}) = A$, $g(\Theta_{(r,-1)}) = 0$ and $b(\Theta_{(r,-1)}) = B$ then

$$d(\Theta_{(r,-1)})u(\Theta_{(r,-1)})B = c(\Theta_{(r,-1)}) \tag{13}$$

$$d(\Theta_{(r,-1)})u(\Theta_{(r,-1)})A = e(\Theta_{(r,-1)}) \tag{14}$$

$$d(\Theta_{(r,-1)})u(\Theta_{(r,-1)})A = A \tag{15}$$

In the same way , we have

$$d(\Theta_{(r,-1)})u(\Theta_{(r,-1)})Ad(\Theta_{(r,-1)})u(\Theta_{(r,-1)}) = d(\Theta_{(r,-1)})u(\Theta_{(r,-1)}) \tag{16}$$

By equation (15) and (16), we know $T = d(\Theta_{(r,-1)})u(\Theta_{(r,-1)})$ is a semi-inverses $A^{\{1,2\}}$ of A. Let $C = TB = c(\Theta_{(r,-1)}) = RFMLRcirc_r fr(c_0, c_1, \ldots, c_{n-1})$ and let $E = TA = e(\Theta_{(r,-1)}) = RFMLRcirc_r fr(e_0, e_1, \ldots, e_{n-1})$. According to equation **(13)** and the characters of the RFMLrR circulant matrix, we know that the last column of C is $(c_{n-1}, \ldots, c_1, c_0 - c_{n-1})^T = Tb$. Since $AX = b$ has a solution, then $ATb = b$, i.e., Tb is a solution of equation (6). We have

$A[Tb + (I_n - E)Y] = ATb + A(I_n - E)Y = b + AY - AEY = b + AY - ATAY = b + AY - AY = b$

So $X_2 = X_1 + (I_n - E)Y$ is a general solution of equation (6), where Y is an arbitrary n-dimension column vector and $X_1 = Tb$.

Since both A and T are RFML*r*R circulant matrices, then $AT = TA$. If there exists another RFML*r*R circulant matrix T_1 such that $AT_1A = A$, $T_1AT_1 = T_1$, $T_1A = AT_1$. Let $AT = TA = H$ and $AT_1 = T_1A = F$. Clearly $H^2 = H$ and $F^2 = F$. Thus we have $H = AT = AT_1AT = FH$, $F = T_1A = T_1ATA = FH$.

So $H = F$. Hence $T = TAT = HT = FT = T_1AT = T_1H = T_1F = T_1AT_1 = T_1$. So T is unique. Hence $TB = C$, $TA = E$ and Tb are also unique.

By Theorem 1 and Theorem 2, we have the following fast algorithm for solving the RFML*r*R circulant linear system (6) :

Algorithm 1:

Step 1. From the RFML*r*R circulant linear system (6), we get the polynomial
$$f(x) = \sum_{i=0}^{n-1} a_i x^i, \quad g(x) = x^n + x - r \text{ and } b(x) = (b_0 + b_{n-1}) + \sum_{i=1}^{n-1} b_i x^i;$$

Step 2. Change the polynomial matrix $\begin{pmatrix} f(x) & b(x) \\ g(x) & 0 \end{pmatrix}$ into the polynomial matrix $\begin{pmatrix} d(x) & c(x) \\ 0 & c_1(x) \end{pmatrix}$ by a series of elementary row operations;

Step 3. If $d(x) = 1$, then the RFML*r*R circulant linear system (6) has a unique solution. Substituting x by $\Theta_{(r,-1)}$ in polynomial $c(x)$, we obtain the RFML*r*R circulant matrix $C = c(\Theta_{(r,-1)}) = RFMLRcirc_r fr(c_0, c_1, \ldots, c_{n-1})$. So the unique solution of $AX = b$ is $(c_{n-1}, \ldots, c_1, c_0 - c_{n-1})^T$;

Step 4. If $d(x) \neq 1$, dividing $g(x)$ by $d(x)$, we get the quotient $g_1(x)$ and change the polynomial matrix
$$\begin{pmatrix} f(x)d(x) & d(x)b(x) & d(x)f(x) & d(x)f(x)b(x) \\ g_1(x) & 0 & 0 & 0 \end{pmatrix}$$ into the polynomial matrix $\begin{pmatrix} 1 & c(x) & e(x) & r(x) \\ 0 & c_1(x) & e_1(x) & r_1(x) \end{pmatrix}$ by a series of elementary row operations;

Step 5. Substituting x by $\Theta_{(r,-1)}$ in polynomial $r(x)$, we get the RFML*r*R circulant matrix $R = r(\Theta_{(r,-1)}) = AA^{\{1,2\}}B$. If the last column of R isn't b,

then $AX = b$ has no solution. Otherwise, the RFMLrR circulant linear system $AX = b$ has a solution. Substituting x by $\Theta_{(r,-1)}$ in polynomial $c(x)$ and $e(x)$, we have the RFMLrR circulant matrices $C = c(\Theta_{(r,-1)}) = A^{\{1,2\}}B$ and $E = e(\Theta_{(r,-1)}) = A^{\{1,2\}}A$. So the unique special solution of $AX = b$ is $X_1 = (c_{n-1},\ldots,c_1,c_0 - c_{n-1})^T$ and a general solution of $AX = b$ is $X_2 = X_1 + (I_n - E)Y$, where Y is an arbitrary n-dimension column vector.

By Lemma 1 and Theorem 1, we have the following theorem.

Theorem 3. Let $A = RLMFLcirc_r fr(a_{n-1}, a_{n-2}, \ldots, a_1, a_0)$ be a nonsingular RLMFrL circulant matrix of order n over $\mathbf{C}$ and $b = (b_{n-1}, \ldots, b_1, b_0)^T$. Then there exists a unique RFMLrR circulant matrix $C = RFMLRcirc_r fr(c_0, c_1, \ldots, c_{n-1})$ of order n over $\mathbf{C}$ such that the unique solution of $AX = b$ is $X = (c_0 - c_{n-1}, c_1, \ldots, c_{n-2}, c_{n-1})^T$.

By Lemma 1 and Theorem 2, we have the following theorem.

Theorem 4. Let $A = RLMFLcirc_r fr(a_{n-1}, a_{n-2}, \ldots, a_1, a_0)$ be a singular RLMFrL circulant matrix of order n and $b = (b_{n-1}, \ldots, b_1, b_0)^T$. If the solution of $AX = b$ exists, then there exist a unique RFMLrR circulant matrix $C = RFMLRcirc_r fr(c_0, c_1, \ldots, c_{n-1})$ and a unique RFMLrR circulant matrix $E = RFMLRcirc_r fr(e_0, e_1, \ldots, e_{n-1})$ of order n such that

$X_1 = (c_0 - c_{n-1}, c_1, \ldots, c_{n-2}, c_{n-1})^T$ is a unique special solution of $AX = b$ and $X_2 = X_1 + \hat{I}_n (I_n - E)Y$ is a general solution of $AX = b$, where Y is an arbitrary n-dimension column vector and $\hat{I}_n$ is given in equation (**5**).

By Theorem 3, Theorem 4 and Algorithm 1, we can get the fast algorithm for solving the RLMFrL circulant linear system $AX = b$.

Algorithm 2:

Step 1. From the RLMFrL circulant linear system (6), we get the polynomial

$$f(x) = \sum_{i=0}^{n-1} a_i x^i, \quad g(x) = x^n + x - r \text{ and } b(x) = (b_0 + b_{n-1}) + \sum_{i=1}^{n-1} b_i x^i;$$

Step 2. Change the polynomial matrix $\begin{pmatrix} f(x) & b(x) \\ g(x) & 0 \end{pmatrix}$ into the polynomial matrix

$\begin{pmatrix} d(x) & c(x) \\ 0 & c_1(x) \end{pmatrix}$ by a series of elementary row operations;

Step 3. If $d(x) = 1$, then the RLMF*r*L circulant linear system has a unique solution. Substituting x by $\Theta_{(r,-1)}$ in polynomial $c(x)$, we obtain the RFML*r*R circulant matrix $C = c(\Theta_{(r,-1)}) = RFMLRcirc_r fr(c_0, c_1, \ldots, c_{n-1})$. So the unique solution of $AX = b$ is $(c_0 - c_{n-1}, c_1, \ldots, c_{n-2}, c_{n-1})^T$;

Step 4. If $d(x) \neq 1$, dividing $g(x)$ by $d(x)$, we get the quotient $g_1(x)$ and change the polynomial matrix

$\begin{pmatrix} f(x)d(x) & d(x)b(x) & d(x)f(x) & d(x)f(x)b(x) \\ g_1(x) & 0 & 0 & 0 \end{pmatrix}$ into the polynomial

matrix $\begin{pmatrix} 1 & c(x) & e(x) & r(x) \\ 0 & c_1(x) & e_1(x) & r_1(x) \end{pmatrix}$ by a series of elementary row operations;

Step 5. Substituting x by $\Theta_{(r,-1)}$ in polynomial $r(x)$, we get the RFML*r*R circulant matrix $R = r(\Theta_{(r,-1)}) = A^{\{1,2\}}B$. If the last column of R isn't b, then $AX = b$ has no solution. Otherwise, the RLMF*r*L circulant linear system $AX = b$ has a solution. Substituting x by $\Theta_{(r,-1)}$ in polynomial $c(x)$ and $e(x)$, we have the RFML*r*R circulant matrices $C = c(\Theta_{(r,-1)}) = A^{\{1,2\}}B$ and $E = e(\Theta_{(r,-1)}) = A^{\{1,2\}}A$. So the unique special solution of $AX = b$ is $X_1 = (c_0 - c_{n-1}, c_1, \ldots, c_{n-2}, c_{n-1})^T$ and a general solution of $AX = b$ is $X_2 = X_1 + (I_n - E)Y$, where Y is an arbitrary n-dimension column vector.

3   Examples

Examples 1. Solve the RFML3R circulant linear system $AX = b$, where $A = RFMLRcirc_3 fr(1,0,3,2)$ and $b = (3,0,0,-2)^T$.

From $A = RFMLRcirc_3 fr(1,0,3,2)$ and $b = (3,0,0,-2)^T$, we get the polynomial $f(x) = 1 + 3x^2 + 2x^3$, $g(x) = -3 + x + x^4$ and $b(x) = 1 + 3x^3$.

Then

$$A(x) = \begin{pmatrix} f(x) & b(x) \\ g(x) & 0 \end{pmatrix} = \begin{pmatrix} 1+3x^2+2x^3 & 1+3x^3 \\ -3+x+x^4 & 0 \end{pmatrix}$$

We transform the polynomial matrix $A(x)$ by a series of elementary row operations as follows:

$$A(x) = \begin{pmatrix} 1+3x^2+2x^3 & 1+3x^3 \\ -3+x+x^4 & 0 \end{pmatrix}$$

$$\xrightarrow{2(2)-x(1)} \begin{pmatrix} 1+3x^2+2x^3 & 1+3x^3 \\ -6+x-3x^3 & -x-3x^4 \end{pmatrix}$$

$$\xrightarrow{3(1)+2(2)} \begin{pmatrix} -9+2x+9x^2 & 3-2x+9x^3-6x^4 \\ -6+x-3x^3 & -x-3x^4 \end{pmatrix}$$

$$\xrightarrow{3(2)+x(1)} \begin{pmatrix} -9+2x+9x^2 & 3-2x+9x^3-6x^4 \\ -18-6x+2x^2 & -2x^2-6x^5 \end{pmatrix}$$

$$\xrightarrow{2(1)-9(2)} \begin{pmatrix} 144+58x & 6-4x+18x^2+18x^3-12x^4+54x^5 \\ -18-6x+2x^2 & -2x^2-6x^5 \end{pmatrix}$$

$$\xrightarrow{29(2)-x(1)} \begin{pmatrix} 144+58x & 6-4x+18x^2+18x^3-12x^4+54x^5 \\ -522-318x & 6x-54x^2-18x^3-18x^4-162x^5-54x^6 \end{pmatrix}$$

$$\xrightarrow{159(1)+29(2)} \begin{pmatrix} 7758 & P_1 \\ -522-318x & -6x-54x^2-18x^3-18x^4-162x^5-54x^6 \end{pmatrix}$$

$$\xrightarrow{\frac{1}{7758}(1)} \begin{pmatrix} 1 & P_2 \\ -522-318x & -6x-54x^2-18x^3-18x^4-162x^5-54x^6 \end{pmatrix}$$

Where
$$P_1 = 954 - 810x + 1296x^2 + 2340x^3 - 2430x^4 + 3888x^5 - 1566x^6,$$
$$P_2 = \frac{53}{431} - \frac{45}{431}x + \frac{72}{431}x^2 + \frac{130}{431}x^3 - \frac{135}{431}x^4 + \frac{216}{431}x^5 - \frac{87}{431}x^6.$$

Since $d(x) = 1$, then the RFML3R circulant linear system $AX = b$ has a unique solution. On the other hand,

$$c(x) = \frac{53}{431} - \frac{45}{431}x + \frac{72}{431}x^2 + \frac{130}{431}x^3 - \frac{135}{431}x^4 + \frac{216}{431}x^5 - \frac{87}{431}x^6.$$

Substituting x by $\Theta_{(3,-1)}$ in polynomial $c(x)$, we have the RFML3R circulant matrix $C = c(\Theta_{(3,-1)}) = \mathrm{RFMLRcirc}_3\mathrm{fr}(-\frac{352}{431}, \frac{738}{431}, -\frac{405}{431}, \frac{217}{431})$. So the unique solution of $AX = b$ is the last column of C, i.e. $X = (\frac{217}{431}, -\frac{405}{431}, \frac{738}{431}, -\frac{569}{431})^T$.

Examples 2. Solve the RFML2R circulant linear system $AX = b$, where

$A = \text{RFMLRcirc}_2\text{fr}(1,-1)$ and $b = (-1,2)^T$.

From $A = \text{RFMLRcirc}_2\text{fr}(1,-1)$ and $b = (-1,2)^T$, we get the polynomial

$f(x) = 1 - x$, $g(x) = -2 + x + x^2$ and $b(x) = 1 - x$. Then

$$A(x) = \begin{pmatrix} f(x) & b(x) \\ g(x) & 0 \end{pmatrix} = \begin{pmatrix} 1-x & 1-x \\ -2+x+x^2 & 0 \end{pmatrix}$$

We transform the polynomial matrix $A(x)$ by a series of elementary row operations as follows:

$$A(x) = \begin{pmatrix} 1-x & 1-x \\ -2+x+x^2 & 0 \end{pmatrix}$$

$$\xrightarrow{(2)+x(1)} \begin{pmatrix} 1-x & 1-x \\ -2+2x & x-x^2 \end{pmatrix}$$

$$\xrightarrow{(2)+2(1)} \begin{pmatrix} 1-x & 1-x \\ 0 & 2-x-x^2 \end{pmatrix}$$

$$\xrightarrow{-(1)} \begin{pmatrix} -1+x & -1+x \\ 0 & 2-x-x^2 \end{pmatrix}$$

It is obvious that $d(x) = -1 + x \neq 1$, it denoted that A is singular.

Then $g_1(x) = g(x)/d(x) = 2+x$, $d(x)f(x) = -1+2x-x^2$,

$d(x)b(x) = -1+2x-x^2$, $d(x)f(x)b(x) = -1+3x-3x^2+x^3$, Thus we structure matrix $B(x)$ and transform the polynomial matrix $B(x)$ by a series of elementary row operations as follows:

$$B(x) = \begin{pmatrix} d(x)f(x) & d(x)b(x) & d(x)f(x) & d(x)f(x)b(x) \\ g_1(x) & 0 & 0 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} -1+2x-x^2 & -1+2x-x^2 & -1+2x-x^2 & -1+3x-3x^2+x^3 \\ 2+x & 0 & 0 & 0 \end{pmatrix}$$

$$\xrightarrow{(1)+x(2)} \begin{pmatrix} -1+4x & -1+2x-x^2 & -1+2x-x^2 & -1+3x-3x^2+x^3 \\ 2+x & 0 & 0 & 0 \end{pmatrix}$$

$$\xrightarrow{(1)-4(2)} \begin{pmatrix} -9 & -1+2x-x^2 & -1+2x-x^2 & -1+3x-3x^2+x^3 \\ 2+x & 0 & 0 & 0 \end{pmatrix}$$

$$\xrightarrow{-\frac{1}{9}(1)} \begin{pmatrix} 1 & \frac{1}{9}-\frac{2}{9}x+\frac{1}{9}x^2 & \frac{1}{9}-\frac{2}{9}x+\frac{1}{9}x^2 & \frac{1}{9}-\frac{1}{3}x+\frac{1}{3}x^2-\frac{1}{9}x^3 \\ 2+x & 0 & 0 & 0 \end{pmatrix}$$

then $r(x) = \frac{1}{9}-\frac{1}{3}x+\frac{1}{3}x^2-\frac{1}{9}x^3$, Substituting x by $\Theta_{(2,-1)}$ in polynomial $r(x)$, we get the RFML2R circulant matrix $R = r(\Theta_{(2,-1)}) = AA^{\{1,2\}}B =$ RFMLRcirc$_2$fr$(1,-1)$. since the last column of R is b, then the RFML2R circulant linear system $AX = b$ has a solution. Substituting x by $\Theta_{(2,-1)}$ in polynomial $c(x) = \frac{1}{9}-\frac{2}{9}x+\frac{1}{9}x^2$ and $e(x) = \frac{1}{9}-\frac{2}{9}x+\frac{1}{9}x^2$, we have the RFML2R circulant matrix $C = c(\Theta_{(2,-1)}) = A^{\{1,2\}}B = $ RFMLRcirc$_2$fr$(\frac{1}{3},-\frac{1}{3})$

and $E = e(\Theta_{(2,-1)}) = A^{\{1,2\}}A = $ RFMLRcirc$_2$fr$(\frac{1}{3},-\frac{1}{3})$. So the unique special solution of $AX = b$ is C, i.e.

$$X_1 = (-\frac{1}{3},\frac{2}{3})^T,$$

and a general solution of $AX = b$ is

$$X_2 = X_1 + (I_n - E)Y = \frac{1}{3}(-1+2k_1+k_2, 2+2k_1+k_2)^T,$$

where $Y = (k_1, k_2)^T \in \mathbf{C}$.

References

1. Williams, F.M., Sloane, N.J.A.: The Theory Error-correcting Codes. North-Holland, Amsterdam (1981)
2. David, C.: Regular Representations of Semisimple Algebras, Separable Field Extensions, Group Characters, Generalized Circulants, and Generalized Cyclic Codes. Linear Algebra Appl. 218, 147–183 (1995)
3. He, C.Y.: On the Fast Solution of r-circulant Linear Systems. Syst. Sci. Math. 21(2), 182–189 (2001) (in Chinese)
4. Cline, R.E., Plemmons, R.J., Worm, G.: Generalized Inverses of Toeplitz Matrix. Linear Algebra Appl. 8, 25–33 (1974)
5. Chen, M.K.: On the Solution of Circulant Linear System. SIAM. J. Numer. Anal. 24(3), 668–683 (1987)
6. Jiang, Z.L., Xu, Z.B.: A New Algorithm for Computing the Inverse and Generalized Inverse of the Scaled Factor Circulant Matrix. J. Computa. Math. 26, 112–122 (2008)

7. Jiang, Z.L., Xu, Z.B.: Efficient Algorithm for Finding the Inverse and Group Inverse of FLS r-circulant Matrix. J. Appl. Math. Comput. 18, 45–57 (2005)
8. Jiang, Z.L.: Fast Algorithms for Solving FLS r-Circulant Linear Systems. In: SCET 2012 (Xi'an), pp. 141–144 (2012)
9. Jiang, Z., Lu, F.: The Sum and Product of Jacobsthal and Jacobsthal-Lucas Numbers Representation by Matrix Method. Ars Combinatoria CX, 143–151 (2013)
10. Shen, N., Jiang, Z., Li, J.: On Explicit Determinants of the RFMLR and RLMFL Circulant Matrices Involving Certain Famous Numbers. WSEAS Transactions on Mathematics 12(1), 42–53 (2013)
11. Jiang, Z.: Efficient Algorithms for Finding the Minimal Polynomials and the Inverses of Level-k FLS $(r_1, r_2, \ldots, r_k)$-Circulant Matrices. WSEAS Transactions on Mathematics 12(4), 374–384 (2013)
12. Li, J., Jiang, Z., Shen, N.: Explicit Determinants of the Fibonacci RFPL Circulant and Lucas RFPL Circulant Matrix. JP Journal of Algebra, Number Theory and Applications 28(2), 167–179 (2013)

Automatic Scoring of English Writing
Based on Joint of Lexical and Phrasal Features

Shili Ge

School of English for International Business, Guangdong University of Foreign Studies,
Guangzhou, 510420, China
geshili@gmail.com

Abstract. In order to determine the effect of phrasal features in automated essay scoring (AES) of English writing by Chinese college students, a multiple linear regression model is adopted involving lexical features and phrasal features. A corpus of college English writing containing 660 compositions was used as training and testing texts. The experimental results show that with phrasal features, all indexes including precision, recall, total accuracy and total wrong ratio of the AES model are improved. Especially obvious is the total accuracy, which improves from 38.64% to 43.18%. Though the regression model cannot be applied into practical use yet, it lays a solid foundation for further research in AES field.

Keywords: Automated Essay Scoring, Phrasal Features, Lexical Features, Multiple Linear Regression.

1 Introduction

In automated essay scoring (AES), lexical features are the most fundamental and all AES systems take advantage of these features. Yet, phrasal features are seldom noticed in this field. Therefore, this study aims to explore the application of phrasal features in AES on the basis of lexical features with the technique of text categorization. After training on a training set of 440 compositions, this model is applied in the scoring of a testing set of 220 compositions to test the validity of phrasal features in AES for Chinese college students' English writing.

Traditionally, people consider "language text as the result of a very large number of complex choices. […] This is probably the normal way of seeing and describing language. It is often called a 'slot-and-filler' model, envisaging texts as a series of slots which have to be filled from a lexicon which satisfies local restraints"[1]. These restraints are mainly grammatical. This is the so called "open-choice principle" in linguistic communication. This principle is constructed around lexicon and grammar, but Sinclair points out that words "do not occur at random in a text", "The choice of one word affects the choice of others in its vicinity. Collocation is one of the patterns of mutual choice, and idiom is another. The name given to this principle of organization [of language] is the idiom principle [1]." In other words, "the language

user has available to him a large number of preconstructed or semi-preconstructed phrases that constitute single choices, even though they appear to be analyzable into segments [2]." First language acquisition studies show that through the acquisition, holistic memorization, and application of prefabricated phrases in communication, children can summarize and then acquire the construction rules of these phrases, so as to form grammatical ability and store the whole prefabricated expressions in their mental lexicon [3]. The cornerstone of "idiom principle" is linguistic segments that combine different lexical and grammatical functions. There are still disagreements in the theoretical definition, identification and classification of this linguistic unit, even different terms are adopted in different researches, such as holophrases [4], lexical phrases [3], and chunks [5]. In this research, they are called "phrases" but different names are adopted in citation of different literature.

2   Background and Related Works

2.1   The Definition of Phrases

Phrases in this research include traditionally defined set phrases, such as "kick the bucket" and "for example", semi-phrases, such as "make up one's mind" and "in one's opinion", collocations, such as "earn money" and "kinds of", frame structure, such as "as/so far as…BE concerned" and "with the…of", and sentence patterns, such as "It is necessary/hard/possible/difficult/… for sb. to do sth." and "no matter what/where/how". Verb phrases formed by high frequency verbs plus particles are included, too, such as the phrases of verb "live": live in, live off, line on, live out, live through, live up to, live with, and live without.

Many researches emphasized the high frequency or frequent appearance in a corpus in defining phrases. The definition of phrase based on corpus statistics heavily rely on a certain corpus and there is no wide accepted threshold value of frequency, yet. In this research, phrases are confined to those expressions related to high frequency words in college English teaching in China, which are picked out based on teaching curriculum, dictionaries and teaching experience, mainly including different patterns of verbs and other patterns frequently appear in college English as well.

2.2   The Correlation between Phrase Use and Essay Score

In recent years, many researches about phrases involved essay quality or holistic language proficiency. Diao investigated English undergraduates and found that there was a significant difference in the proficiency of lexical bundles between lower grade students and higher grade ones, and there was a significant correlation between the proficiency of lexical bundles and holistic language proficiency for those higher grade students [6]. However, the relationship between the proficiency of lexical bundles and holistic language proficiency for lower grade students was not mentioned. Ding and Qi conducted a research on senior English majors and found that compared with grammatical knowledge, the knowledge of lexical bundles could predict essay scores more accurately [7]. The knowledge of lexical bundles in this research referred to the

numbers of phrases beginning with verbs. In their research, only phrases beginning with verbs and involving particles were calculated. Naturally, they reached the result that lexical bundles occupied only a minor percentage in essay writing. Wang and Zhang studied some English argumentative essays by English majors and discovered that Chinese students overused three-word phrases in their writing, the types of phases were few and an oral tendency existed in their writing [5]. Xu and Ying analyzed essays by English major freshmen and senior students [8]. They found that higher level students use more aggregative noun phrases while lower level students use more high frequency phrases and sentence structures. However, the variation of phrases used by lower level students is not as much as that of higher level students. In addition, lower level students have a strong tendency of oral English in their writing. For example, higher level students use many phrases like "senior citizen" while lower level students use more phrases like "I think/thought", "in my opinion", "I find/found" and "I guess". The repetition of phrases in higher level students' writing is less than that in lower level students. In this research, the type number of phrases considered is quite large, the following conclusion is drawn: "English learners, no matter how proficiency they are, use large amount of lexical phrases. The frequency of phrases used by higher level students is a little more than lower level students, but the higher level students use more different types of phrases."

All the mentioned research subjects are English majors. Among them, higher grade students, namely, higher English proficiency students and lower grade students, namely lower English proficiency students have many differences in phrase use. However, the relationship between the phrase use of lower English proficiency students and their essay quality is seldom studied and non-English majors, namely, college English writing and the use of phrases in the writing are ignored completely.

3   The Automated Identification of Phrases

There are two methods of phrase identification: statistical and rule-based. Statistical method has its advantages of energy saving and easy extension, but students' English writing is usually no more than a few hundred words and full of various errors. It is impossible to identify phrases statistically on a single piece of student's writing. In this case, rule-based method is more practical. Certainly, in the construction of rules, statistical analyses on large native English corpus are conducted as needed.

In addition, very large corpus is usually required in corpus-based statistical identification of phrases. In China, currently used corpora of college English writing are very small. For example, in CLEC, the most famous and widely used corpus of Chinese student English writing, the sub-corpus for Band-4 examination, st3 and the sub-corpus for Band-6 examination, st4 are only a little more than 200,000 words, respectively. There is no way to carry out a statistical study of phrases based on such a small corpus. Therefore, in order to study the use of phrases in college English writing, we have to resort to manual construction of phrase rules.

3.1   The Rule Construction of Frequently Used Phrases

In the construction of phrase rules, we selected more than 1100 phrasal verbs covering 300 verbs frequently used in college English writing and more than 200 non-verb phrases. All these phrases were formalized to construct more than 1300 phrase rules, which formed the phrase rules of correct usage in the rule base. For those phrases that students often use wrongly in their writing, we generated more than 200 wrong use rules based on composition analysis and teaching experience. All these rules are listed in sequence. They are matched from the very beginning in use. Once a rule is matched, the matching program will stop and this round of match is finished.

The base of phrase rules is composed of three parts: the list of fundamental patterns, phrase rules of correct usage and phrase rules of wrong usage.

3.2   Fundamental Patterns

The fundamental patterns are a list of root words and their variations, and some certain constituents in rules and their lexical substitutions. Such as:

ACT := act|acts|acted|acting
BE := am|is|are|was|were|been|being|be|\'m|\'re|\'s

Every item in the list is composed of three parts: the entry in phrase rules, namely, the capitalized root word or pattern constituent, such as "ACT"; the "or" (representing by "|") selection of all possible substitutions of the entry; and the symbol ":=", separating the entry and its substitutions. Before the application of the rules, the capitalized entry in rules will be substituted with the "or" selection string.

3.3   Phrase Rules of Correct Usage

1) Rules of phrasal verbs, such as:

@ACT
ACT/Vω as/ω→ACT as

Rules in every verb group include lexical entry composed of symbol @ and a capitalized root verb, and phrase rules with the verb as the keyword. Each rule is separated into two parts with the symbol "→"as in the above example. The former part is a match string composed of words and POS tags. The latter part is the phrasal verb that the match string represents, namely, the feedback of the rule.

2) Rules of adjective phrases, such as:

@difficult
BE/Vω (ω/RBω?){0,2} difficult/ω (for/ωω/ω)? to/TO ω/VB→be difficult to do sth

3) Rules of other phrases

Rules of prepositional phrases, such as:

@about

BE/Vω about/ω to/TO ω/VB→BE about to do sth

Rules of noun phrases, such as:

@shortage

BE/Vω (ω/RBω?)? in/ω (ω/JJ)? shortage/ω→BE in shortage

Rules of sentence structures, such as:

@SEN_PATTERN

it/ω ('s|is)/ω (ω/JJ)? FOR_PAT/ω (for/IN (ω/ω){1,3})? (not/RB)? to/TO ω/VB →it is JJ to do sth

3.4   Phrase Rules of Wrong Usage

Students' wrong use of phrases is also formalized into rules, such as:

! LOSE/ω ((ω/ω 's/POS)|ω/PRP$)   hearts?/ω to/TO ω/VB→LOSE one's heart to sb/sth

In this rule, the initial symbol, "!", means this is a phrase rule of wrong usage. Certainly, these rules of wrong usage must be set up before application. In present study, the rules of wrong phrase use mainly cover grammatical errors, such as constituent missing, constituent redundancy, and constituent substitution. Among them, the first two kinds of errors are the most difficult aspects of English grammar for language learners [9].

4   The Experiment of AES Involving Lexical and Phrasal Features

In order to test the effect of phrasal features in AES, the testing materials, namely, the compositions have to be prepared first. Then, phrases are extracted from the compositions automatically and the results are evaluated. Finally, together with lexical features, phrasal features are applied in the automated scoring of compositions. These are the steps in any machine learning algorithm such as [10] and [11].

First, composition set should be prepared.

The Chinese college students' compositions in this research come mainly from the sub-corpus, st3, of Chinese Learner English Corpus (CLEC) and partly from the composition collection by Guan and Chen [12].   There are totally 660 compositions including more than 106,000 words.   Word spelling errors and simple grammar errors, such as subject verb agreement, are identified and corrected in the phase of pre-processing.   All compositions are scored by two raters with inter-rater coefficient of 0.761. With the help of a third rater, the discrepancy of scoring is resolved. In pre-processing, compositions are also POS tagged with Stanford NLP Group Part-of-Speech tagger [13], in which the tag set of Penn Treebank is adopted.   The accuracy of POS tagging is 97.6%, which is higher than the acceptable POS tagging accuracy [14]. The 660 compositions include 60 compositions with a score level of 1 and 150 compositions with a score level of 2 to 5, respectively. All these compositions are randomly divided into training set, which includes 2/3 of all compositions in each

score level (440 compositions in sum), and testing set, which includes 1/3 of all compositions in each score level (220 compositions in sum), for model construction and testing of AES.

Second, phrasal features have to be extracted and evaluated.

The program of phrase identification is keyword based. When a POS tagged sentence is input, words are scanned one by one from the beginning and checked in the base. Once a word is found in the base as an entry, the matching program is started.

Every group of patterns in the base are listed in sequence and matched from the first to the last when needed. Once a pattern is matched, the matching program reports the match result and starts from the next word behind the matched string until the end of the sentence.   The detailed algorithm is as following:

```
1     for all compositions do
2                 for all sentences in a composition do
3                       for all words in a sentence do
4                             if a word in entry of rule set and the rule match the
sentence:
5                                   output the phrase and skip to the next
word behind the phrase
6                             else:
7                                   skip to next word
8                       end
9                 end
10    end
```

The evaluation criteria of phrase identification are precision and recall, which are estimated using counts of hits, false positives and misses. A hit occurs when the program identifies an actual phrase, no matter correct usage or wrong usage. A miss occurs when the program fails to identify a phrase. A false positive occurs when a non-phrase is identified as a phrase, a wrong used phrase is not identified, or a correct used phrase is identified as wrong use.

Precision is the proportion of identified items that are, in fact, phrases, no matter correct or wrong use.

$$\Pr ecision = \frac{Hits}{Hits + False\_Positives} \tag{1}$$

It measures how often the program is correct when it reports that a phrase has been found. Recall is the proportion of actual phrases that have been identified.

$$\mathrm{Re} call = \frac{Hits}{Hits + Misses} \tag{2}$$

It measures the program's coverage, i.e., the fraction of phrases that the program has identified.

The program identified 7,384 phrases in the experimental composition set including 6,662 correct usages and 722 wrong usages. In order to gain an accurate evaluation, all 660 compositions were read through by two experienced English teachers to count hits, false positives and misses. The kappa statistic [9] calculated is 0.71, which means that a substantial agreement exists between the two raters. With careful discussion, two raters settled the disagreement and the final results achieved are as follows: hits, 7109; false positives, 275; and misses, 1457. Then, the precision and recall calculated are 96.28% and 82.99%, respectively.

Third, the experiments of automated essay scoring (AES) are carried out.

The evaluation criteria of AES are set in the first step.

In this research, the results of automated scoring are compared with the calculation of the following parameters: precision and recall of each score level, and total precision and total wrong ratio. They are defined as following:

Precision of each score level = The number of score X compositions which are scored as X / The total number of compositions which are scored as X * 100

Recall of each score level = The number of score X compositions which are scored as X / The total number of compositions with score X * 100

Total accuracy = The number of accurately scored composition in the testing set / The total number of composition in the testing set * 100

Total wrong ration = The number of compositions with an AES score which is 2 points larger or smaller than its real score / The total number of composition in the testing set * 100

In the above formulae, X= 1, 2, 3, 4, or 5.

The higher the precision for each score level, the more reliable the score assigned by the AES model to the composition is. For example, the precision of a certain score level X is 50%. While composition A is scored X, then the probability that A really has a score X is 50%.

The higher the recall of each score level, the more contribution they have to the total accuracy.

The higher the total accuracy, the more accurate the model is and thus, the more practical and empirical to put into real application.

The lower the total wrong ratio, the less unacceptable error the AES model possesses. Thus, the acceptability of the AES model in real application is increased.

After this, the experiment of AES with lexical features is conducted.

There are usually 5 steps in an AES experiment: feature selection, algorithm design, model training, automated scoring, and model evaluation.

The most fundamental features applied in automated essay scoring are lexical features, which are adopted in all AES researches and systems, such as PEG, IEA, e-rater and Liang [15] etc. Therefore, lexical features are adopted to train a scoring model as the baseline of automated scoring for this research.

There are lots of lexical features involved in AES studies and the most frequently adopted ones are lexical distribution, average word length and composition length. These features are adopted in this research, too. Lexical distribution (LD) has long been believed to be closely related to writing quality [16]. LD is measured based on a certain word list. There are many different word lists and they have different validity for different types of writing. In this research, a specifically revised word list from

[16] is adopted (refer to [17]), which includes 4 categories of words, wd1 to wd4, from the most frequently used words to rare words. Average word length (AWL) is a key lexical feature adopted in e-rater [18]. It is also an important index for lexical complexity. Another lexical feature is the total number of words in a composition, that is, the composition length (CL). "Across many studies, the length of an essay strongly predicts the quality score a human rater will assign it" [18]. The features extracted in this experiment is obviously different from that in text clustering and classification as in [19] and speech feature extraction as in [20] since feature extraction has to serve for the aim of the study. The features extracted in this research have to reflect the writing ability of composition writers.

The algorithm of AES in this research is multiple linear regression, which is also the widely adopted algorithm in many AES researches. The model training steps are as following:

1) input compositions from the training set one by one;
2) segment composition into single words;
3) count total word number, namely, CL;
4) calculate the average word length (AWL);
5) count the number of words in different word level according to word list by Li and Ge [17] (LD);
6) output the composition score (CS), LD, AWL and CL;
7) take CS as dependent variable and LD, AWL and CL as independent variables input them into multiple linear regression to get the regression formula as AES model.

After extract all the values of above variables from the training set with self-coded Python programs, the regression equation is achieved as following:

score = -0.588 – 0.056 * wd2 - 0.004 * wd3 – 0.037 * wd4 + 0.132 * AWL + 0.057 * length

The key elements of the AES model are listed in table 1.

Table 1. The elements of regression equation of the training set with lexical features

The elements of regression equation	State of the element
Coefficient of determination	0.39
Adjusted coefficient of determination	0.381
ANOVA F	38.221
ANOVA P	0
Maximum partial correlation coefficient	Average word length
Outlier	Yes

Then, the model is applied to the scoring of the compositions in the testing set. Each composition is processed to extract the values of all the variables in the model and the machine assigned score is calculated according to the regression equation.

At last, the machine assigned scores are compared with the real scores of the compositions in the test set to evaluate the scoring model according to the criteria set in the previous section. The results are shown in table 2.

Table 2. The performance (%) of AES models with lexical features and phrasal features

Score level	Evaluation criteria	Lexical features only	Lexical and phrasal features
1	precision	80.00	83.33
	recall	20.00	25.00
2	precision	38.89	47.62
	recall	28.00	40.00
3	precision	34.04	45.78
	recall	64.00	76.00
4	precision	29.41	35.71
	recall	40.00	50.00
5	precision	88.24	78.95
	recall	30.00	30.00
Total accuracy		38.64	43.18
Total wrong ratio		15.91	13.18

The scoring result with lexical features is the baseline for adding in other features such as phrasal ones.

At last, the experiment of AES with both lexical and phrasal features is conducted.

Many studies show that the use of phrases is closely related to essay quality though few AES researches involve this feature. Based on the relatively accurate extraction of phrases from college English writing with the help of self-coded phrase identification rule set, an experiment of AES involving phrasal feature is conducted.

The phrasal features adopted in this research include the total number of phrases, the number of verb phrases, and the number of wrongly used phrases in a composition. The extraction of phrases is as described above and the model training steps are very similar to those listed in previous section, only that the phrasal features are added after step 5 and then output together with other lexical features for further model building. The regression model achieved is shown as following:

score = $- 1.392 - 0.027 * wd2 - 0.015 * wd3 + 0.024 * wd4 + 0.125 * AWL + 0.032 * length + 0.060 * ttlphrase + 0.017 * verbphrase - 0.144 * wrongphrase$

After applying this model to the testing set, each composition receives a machine assigned score and the score is evaluated based on its real score. The results are also listed in table2.

Table 2 shows that taking the scoring model of only lexical features as the base line, when the phrasal features are involved, the performance of the model is significantly improved. From table 2 we can see that all the recalls are obviously raised for every score level and all the precisions are raised, too, except for score level of 5. The total accuracy is raised over 4% than the baseline of only lexical features and the total wrong ratio decreases nearly 3%. These changes prove that the phrasal features have a relatively high predicable capacity for composition score.

5   Conclusion and Discussion

The review of related literature shows that phrases used in college English writing have a close relation with writing quality. With the help of college English writing corpus and large English corpus, phrase identification rules are coded, including verb phrases, non-verb phrase, and wrongly used phrases. Based on the rule set, phrases can be accurately identified from college English compositions, which form the solid base for further automated essay scoring research.

The AES research in this study includes two stages. The first stage is based solely on lexical features and the second mixed with phrasal features. Experiments show that when phrasal features are added in, the precisions and recalls for almost all score levels are increased, and certainly the total accuracy is increased, too. Though the total accuracy is not high enough for practical use in daily English teaching, this research laid a solid foundation for further AES research.

References

1. Sinclair, J.: Corpus, Concordance, Collocation. Oxford University Press, Oxford (1991)
2. Partington, A.: Patterns and meanings: Using corpora for English language research and teaching. John Benjamins, Amsterdam (1998)
3. Nattinger, J., DeCarrico, J.: Lexical Phrases and Language Teaching. Oxford University Press, Oxford (1992)
4. Corder, S.P.: Introducing Applied Linguistics. Penguin, London (1973)
5. Wang, L., Zhang, Y.: A Corpus-based Study on Chunks in English Argumentative Writing of Chinese EFL Learners. Computer-Assisted Foreign Language Education (4), 36–41 (2006)
6. Diao, L.: A Survey of Chinese English Majors' Chunk Competence. Journal of PLA University of Foreign Languages (4), 35–38 (2004)
7. Ding, Y., Qi, Y.: Use of Formulaic Language as a Predictor of L2 Oral and Written Performance. Journal of PLA University of Foreign Languages (3), 49–53 (2005)
8. Xu, F., Ying, Y.: The Comparison of Phrases in High and Low Proficiency Level Students' English Writing. Journal of Shaoxing University 27, 107–109 (2007)
9. Leacock, C., Chodorow, M., Gamon, M., Tetreault, J.: Automated Grammatical Error Detection for Language Learners. Morgan & Claypool Publishers (2010)
10. Hasan, T.M., Wu, X.: An Adaptive Algorithm for Improving the Fractal Image Compression (FIC). Journal of Multimedia 6, 477–485 (2011)
11. Zhang, Q., Wang, P., Yang, H.: Applications of Text Clustering Based on Semantic Body for Chinese Spam Filtering. Journal of Computers 7, 2612–2616 (2012)
12. Guan, X., Chen, J.: College English Writing. Jilin University Press, Jilin (2004)
13. Toutanova, K.: The Stanford NLP Group Part-of-Speech Tagger (2009), http://nlp.stanford.edu/software/tagger.shtml

14. de Haan, P.: Tagging non-native English with the TOSCA-ICLE tagger. In: Mair, C., Hundt, M. (eds.) Corpus Linguistics and Linguistic Theory, Freiburg im Breisgau, pp. 69–79 (2000)
15. Liang, M.: The Research and Development of Automated Essay Scoring System for Large Scale English Test. Higher Education Press, Beijing (2012)
16. Laufer, B., Nation, P.: Vocabulary Size and Use: Lexical Richness in L2 Written Production. Applied Linguistics 3, 307–322 (1995)
17. Li, Y., Ge, S.: The Validity of Word List in Automated Essay Scoring for College Students. Foreign Languages and Their Teaching (10), 48–52 (2008)
18. Enright, M.K., Quinlan, T.: Complementing human judgment of essays written by English langauge learners with e-rater® scoring. Language Testing 27, 317–334 (2010)
19. Beliakov, G., Yearwood, J., Kelarev, A.: Application of Rank Correlation, Clustering and Classification in Information Security. Journal of Networks 7, 935–945 (2012)
20. Zhao, H., Zhao, K., Liu, H., Yu, F.: Improved MFCC Feature Extraction Combining Symmetric ICA Algorithm for Robust Speech Recognition. Journal of Multimedia 7, 74–81 (2012)

Trustworthy Software Development Based on Model Driven Architecture

Yang Zhu[1], Lanhua Fei[3], and Nianhua Yang[2]

[1] Information Technology Center,
Shanghai University of International Business and Economics, Shanghai 201620, China
[2] School of Business Information Management,
Shanghai University of International Business and Economics, Shanghai 201620, China
[3] Materials Center, General Electric Company in China,
Beijing 100192, China
zy_shift@163.com, feilanhua@126.com, yangnianhua@suibe.edu.cn

Abstract. In the recent years, a lot of accidents caused by software have resulted in serious consequences. So software trustworthy has attracted attentions from the governments, enterprises and research institutions. In order to improve trustworthiness of software models in the early design phase, a trustworthy software development framework is proposed in this paper. It integrates characteristics of MDA (Model Driven Architecture, MDA), Petri net and temporal logic. MDA is used to improve model reusability, shorten software development cycle and meet ever-changing requirements. Petri nets are used to formally represent software functional models. Temporal logic is used to represent non-functional requirements of the model. Existing techniques can be used to formally analyze performance and check correctness of software models in the design phase. Thus, the purpose for enhancing software models' trustworthy in the design phase can be achieved.

Keywords: Model driven architecture, Petri net, temporal logic, trustworthy software.

1   Introduction

In the resent years, software development needs to meet the requirement of short time to market and constant evolution. Model driven architecture (MDA) [1] is an attractive approach to realizing this target. It separates the application description from the platform specific implementation. Thus the modeling complexity can be reduced. Furthermore, alternative solutions to the target system can be investigated and compared in the early phase. The development risks can also be minimized. So it is an effective mean to reduce software development time and adapt to frequent requirement changing. Thus, the software trustworthy can be improved under MDA architecture. Models are the most important elements for the software development based on MDA.

In the recent years, a lot of accidents caused by software have happened in the worldwide. These accidents usually resulted in serious consequences. So these incidents attracted attentions from the governments, enterprises and research institutions. They decided to establish some approaches to enhance software trustworthy [2-5]. Establishing trustworthy software model in the design phase is one of the main approaches. To achieve this goal, formal methods are usually used for analyzing and verifying models during the early development phases.

UML (Unified Modeling Language, UML) [6] has become the de facto standard in the software modeling field. UML 2 has proposed as modeling language for MDA by OMG (Object Management Group). UML is a semi-formal language. It describes syntax of diagrams using formal methods. Though it has been widely used in software design, it absents the ability of formally verifying and analyzing dynamic semantics of the described diagram. Formal methods [7] are effective approaches to analyzing and verifying systems' functional and non-functional properties and improving systems' trustworthy. In order to verify functional properties of the model and analyze satisfactions of non-functional properties under the model in the design stage, it is necessary to formally verifying and analyzing software models. Petri net [8] is a formal model based on strict mathematical theory. It is appropriate for describing software functional model. Temporal logic [9] has been wildly used to formally describe software requirement specifications. The combination of Petri net and temporal logic can be used to formally describe and verify software models [10].

This paper proposes a software development framework which composing MDA, Petri net and temporal logic. The framework can be used to shorten software developing time, adapt to constant requirement changing and formally verifying and analyzing system models in software design stage for improving software trustworthy. It firstly presents related concepts about MDA, Petri net and temporal logic. Then, it proposes the framework, which combing Petri net and temporal logic under MDA, to improve software trustworthy during development. And the software development process under this framework is presented. At last, related work to this paper is analyzed.

2   Trustworthy Software Development Framework

2.1   MDA

MDA separates business logics and their platform specific implementation techniques. It can produce design models and software codes automatically through models automatic transformation. Thus, software development process is simplified and software reusability is enhanced. Core concerns in MDA are models and transformations.

MDA provides four categories model types classed into three abstract levels, such as shown in the left part of Fig. 1, which describes software development process based on MDA. In MDA, CIM (computation independent model, CIM) is used to build software requirement model. PIM (platform independent model, PIM) is used to build platform independent software functional model which is used to realize one of CIMs. PDM (platform description model, PDM) is used to describe realization

techniques under a specific platform. These techniques include operating system, programming language, database, user interface, et. al.. PSM (platform specific model, PSM) combines PIMs and PDMs. In the MDA platform, PIM is used to describe business functions and behavior specifications according to CIM which is built by business domain engineers. PIM and PDM are combined and transformed into PSM. According to code producing rules, PSM is transformed into codes by code generator. The codes are applied to a specific platform which is described by PDM. By using MDA methods, general software developers only need to focus on how to build PIM with UML (Unified Modeling Language). PSM and codes can be generated automatically according to PIM and transformation rules. In PIM, non-functional properties, such as correctness, reliability, real-time, have become important factors in determining the success of software development. However, PIM is described by UML in MDA. And non-function properties of PIM can't be effectively verified and analyzed directly.
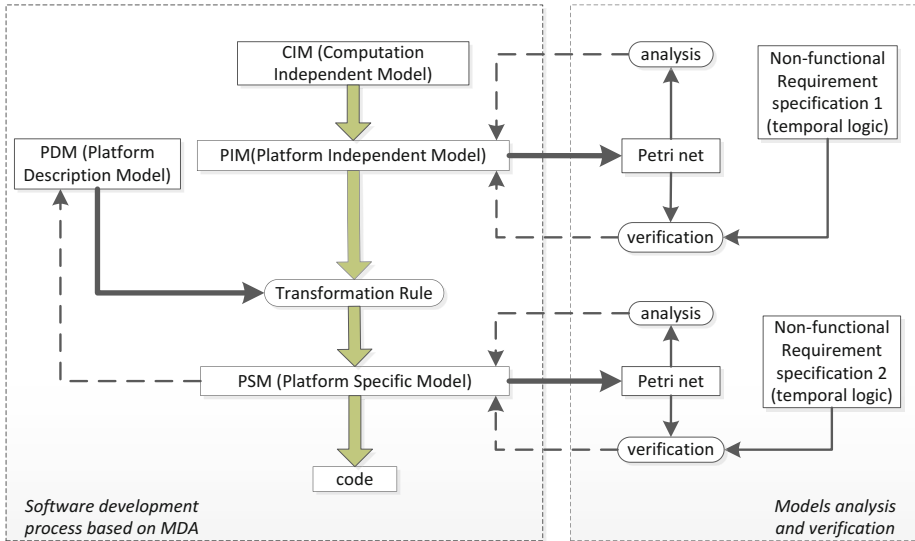
Fig. 1. Trustworthy software development framework based on MDA

2.2   Petri Net

A Petri net [8] can be described by a visualized diagram. It is also based on restrict mathematical theories. So it is an effective tool for modeling and analyzing. It can describe not only static structures but also dynamic behaviors of the system. In addition, it is also applied for modeling systems with concurrent, conflicts or synchronization relations. Therefore, Petri nets are appropriate for describing software functional models. Developed for nearly half a century, Petri nets have been equipped with abundant theories and also wildly used in the computer science and engineering field [11, 12]. Model analysis methods of Petri nets provide important approaches to analyzing properties of the target system.

A Petri net is a tuple $TCPNIA =< P, T, A, B, W, M_0 >$, where P is a finite set of places; T is a finite set of transitions; $A \subseteq (T \times P)$ is a set of output arcs; $B \subseteq (P \times T)$ is a set of input arcs; W is the weight of arcs; M_0 is the initial marking.

$^\circ t = \{ p_i \mid (p_i, t) \in B \}$ is used to represent the pre-set of a transition t. Similarly, $t^\circ = \{ p_j \mid (t, p_j) \in A \}$ represents the post-set of a transition t.

A transition $t \in T$ is said to be enabled if and only if the token number of each place p_i in the pre-set of t is larger than the weight of the arc (p_i, t), i. e. $\forall p_i \in \{ p_i \mid (p_i, t) \in B \} \rightarrow M(p_i) > w(p_i, t)$.

Supposing a transition t has a pre-set $^\circ t = \{ p_1, p_2, \cdots, p_a \}$. The firing of an enabled transition t changes the marking M into a new marking $M^{'}$. As a result of firing the transition t, next two events occur concurrently:

a) $\forall p_i \in {}^\circ t \Rightarrow M^{'}(p_i) = M(p_i) - w(p_i, t)$;

b) $\forall p_j \in t^\circ \Rightarrow M^{'}(p_j) = M(p_j) + w(t, p_j)$.

2.3   Temporal Logic

temporal logic [9] uses rules and symbolism for representing, and reasoning about propositions which are qualified in terms of time. We can use the temporal logic to express statements like "I am always busy", "I will eventually be busy", or "I will be busy until next week". It has been developed further by computer scientists, notably Amir Pnueli, and logicians.

Temporal logic has been used for formal verification, where it is used to state non-functional requirements of hardware or software systems. For instance, one may wish to state that whenever an event is occurred, another event will must be occurred at some time. Such a statement can conveniently be expressed in a temporal logic.

CTL (Computation Tree Logic, CTL) [13] is based on branching time statement logic which divides time into multi possible futures. The future is not determined. Any path in the future might be an actual one that is realized. It can be used in formal verification of software or hardware, typically by software applications known as model checkers which determine whether a given model satisfy safety or liveness properties. For example, CTL can specify that when some initial condition is satisfied, then all possible paths of a program avoid some undesirable condition. CTL is a class of temporal logics.

CTL formula consists of automatic propositions, Boolean connectors and temporal operators. A temporal operator is built by adding a path operator before time operator. Time operators consists of G, F, X and U, which represent all the future, a time in the future, the next time and until respectively. Path operators consists of A and E, which represents all the paths, existing a path respectively. CTL can easily describe a system's security, activity, equity, et al.. These non-functional properties could be verified by a model checker that explores all possible transitions out of program states satisfying the initial condition and ensures that all such executions satisfy the

properties. TCTL (Timed CTL, TCTL) [14] is an extension for CTL with real-time property. It adds quantitative constraints on temporal operators. TCTL can be used to describe real-time requirement specification of a system.

2.4   Software Development Framework

Fig. 1 shows a trustworthy software development framework which combines MDA and formal methods. A software designer describes software functional model with UML to produce PIM according to the CIM which built by domain engineer. A system engineer gives a PDM. PIM and PDM can be integrated into PSM according to a given transformation rules. Many contributions have been done for transforming PSM into codes.

Key issues for improving software reliability and reducing software development costs are formally modeling and analyzing PIM in the modeling stage under MDA framework. In the framework proposed in this paper, PIM in transformed into a proper Petri net after the CIM is transformed into a PIM. Different Petri nets can be used to describe PIM according to the requirement of the target software system. For example, timed colored Petri net [15] can be used to described a real-time system. According to related theories about Petri net, properties, such as activity and boundness, of the target Petri net model can be analyzed. If the analysis result is not consistent with requirement, the software designer should look for the cause in PIM, and alter the PIM. Analyzing and altering should be done repeatedly until the analysis result satisfy the requirement. System requirements, such as security, fairness and real-time, are described by temporal logic. In order to use existing model checker to checking whether these requirements are satisfied in the software functional model, Petri net should be transformed into the input language of the model checker. If the checking result shows that requirements described by temporal logic are not satisfied in the Petri net model corresponding to PIM, the software designer should alter the PIM, retransform the PIM into a Petri net, and rechecking the Petri net model repeatedly until the requirements are satisfied by the Petri net model. A dashed line with an arrow in Fig. 1 represents returning the checking result.

PIM is transformed into PSM automatically after its requirements have been confirmed by formal verification and analysis. PSM is also necessary to be verified and analyzed for possibly exists of artificial defects produced in the design stage or models in consists which are produced during the processes when PDM and PIM are combined and transformed into PSM. Because methods and procedures for verifying and analyzing PSM are similar with those for PIM, it's not necessary to describe them in detail. Software designer will check the PSM and PDM according to the verification results.

After the PSM is formally verified and analyzed, codes are produced from PSM automatically according to the methods of MDA. In this framework, models in different levels are analyzed and verified, and non-functional properties are ensured to be satisfied in the models. Therefore, the trustworthy of the software system can be enhanced.

3   Example

Fig. 2 shows a PDM which represents order processing described by UML activity diagram. The PDM should satisfy some non-functional properties. The order process should be closed in the system whether the merchant accept or reject the order after receiving it. An invoice should be produced after the merchant accept the order. In order to verify whether these properties are satisfied in the PDM model presented in Fig. 2, the PDM is transformed into a Petri net and non-functional properties are described with temporal logic according to the framework in Fig. 1. Model checker is used to verify whether these non-functional properties are satisfied in the Petri net model.
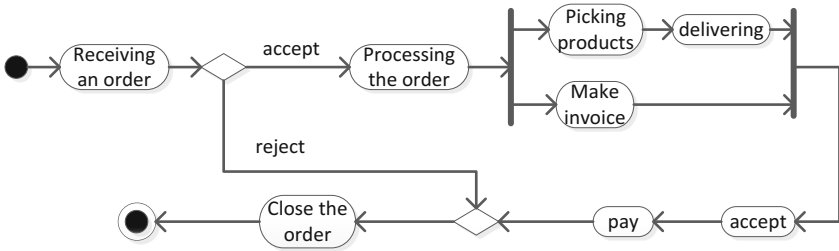
Fig. 2. PDM for order processing

The activity diagram in Fig. 2 is transformed into a Petri net model according to the method in [16]. Then it is simplified into the model shown in Fig. 3. In Fig. 3, t_1 represents that the merchant receive the order. t_2 represents that the merchant reject the order. t_3 represents that the merchant accept the order. t_4 represents the begin of order process by the merchant. t_5 represents the action of picking products for the order. t_7 represents the action of delivering for the order. t_6 represents the invoice producing for the order by the accounting department. t_8 represents that the customer receives the products and invoice. t_9 represents the pay action for the order. t_{10} represents the closure of the order. Based on the model in Fig.3, non-functional properties, such as activity, security, et. al., can be analyzed taking advantage of existing contributions related to Petri net.

Taking into account the Petri net in Fig. 3, the non-functional property, an order will be closed once the order is committed whether it was reject or accept by the merchant, can be represented as $AG(p_2 \rightarrow AF\ p_{11})$ using CTL. The formula indicates that once a token appeared in the place p_2, the place p_{11} will definitely contain a token at a time point in the future. Another non-functional property, an invoice for the order should be built once the merchant accept the order, can be represented as $AG(p_3 \rightarrow AF\ p_7)$ using CTL.
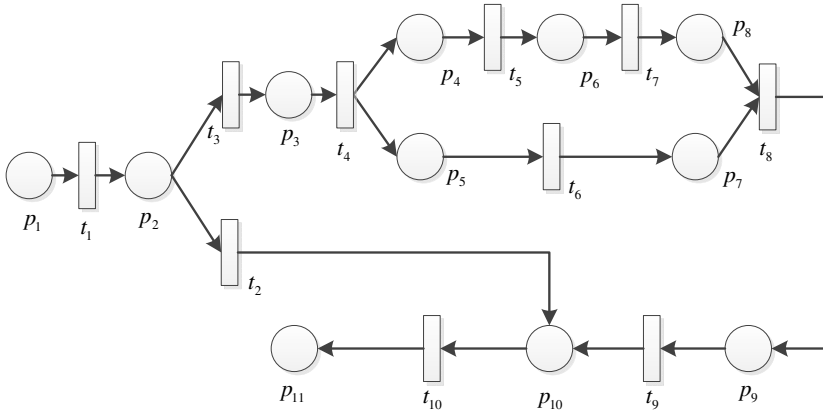
Fig. 3. The Petri net model for order processing

How to verify whether a non-functional property expressed by CTL is satisfied in the functional model presented by Petri net using the model checker, SMV [17], has been detailed in [18]. According to the method proposed in [18], model checking for verification has been taken. The result shows that the above mentioned non-functional properties are satisfied in the functional model. If some errors are appeared in the model checking process, the model checker will return the reversed error paths. The software designer can find the defect of the model following the reversed error paths. Thus, the model can be modified and reliability of the software can be improved.

Analysis and verification methods for PSM are similar to those for PDM. They will not be demonstrated in detail.

4   Related Work

Model transformation is one of the core techniques for software development based on MDA. The contribution in [19] studies methods for automatic model transformation from CIM to PIM. Much more researches concentrate on the transformation form PIM to PSM [20] or from PSM to software codes. Query/View/Transformation (QVT)[21] framework has been proposed by OMG to realize the transformation from models in MDA to codes. These researches worked for finding methods for model transformation between different levels in MDA.

Ameedeen et al. [22] propose rules for transforming a UML sequence diagram into a Petri net. Fernandes et al. [23] propose a model transformation method from a use case diagram into a colored Petri net. Choppy et al. [24] propose a method for transforming a UML state diagram into a colored Petri net and formal verification and analysis methods for the Petri net. López-Grao et al. [25] propose a transformation method from a UML activity diagram into a stochastic Petri net. Aspect orient method is also used to enhance model transforming correctness and effectiveness between UML models and Petri nets [26].

Temporal logic has been widely used to describe non-functional requirements. Ljungkrantz et al. [27] use temporal logic to describe time-related properties. Tun et al. [28] show how specification features may be formalized using a form of temporal logic called the Event Calculus, and prove their correctness using an off-the-shelf tool.

In order to realize automatic model checking, a lot of mature model checkers have been developed. Uppaal [29], as one of representative model checkers, not only applies to a common system, but also applies to real-time system. In addition, it provides graphical user interface for convenience. In order to checking Petri net model using Uppaal, Cassez et al. [30] propose a transforming method from time Petri net to timed automata, and give the correctness certification for the transformation.

5   Conclusion

In order to improve trustworthiness of software models in the early design phase, this paper proposes a trustworthy software development framework which combines MDA, Petri nets and temporal logics. An example demonstrates the feasibility and practicality of the framework. The framework can improve software reliability, shorten the software development cycle and reduce total costs of software development by using existing model transformation and analyzing methods. In the future, we will develop an automatic tool to realize the framework proposed in this paper.

References

1. OMG: MDA Guide Version 1.0.1 (2003), `http://www.omg.org/cgi-bin/doc?omg/03-06-01`
2. Tsai, W.-T., Zhou, X., Chen, Y., Bai, X.: On testing and evaluating service-oriented software. Computer 41(8), 40–46 (2008)
3. Li, Y., Song, Y.-D.: An adaptive and trustworthy software testing framework on the grid. The Journal of Supercomputing 46(2), 124–138 (2008)
4. Yan, Z., Prehofer, C.: An Adaptive Trust Control Model for a Trustworthy Component Software Platform. In: Xiao, B., Yang, L.T., Ma, J., Muller-Schloer, C., Hua, Y. (eds.) ATC 2007. LNCS, vol. 4610, pp. 226–238. Springer, Heidelberg (2007)
5. Kun, X., Yong, G., Xing, J.: A model of trusted software based on software gene. In: Proceedings of the International Conference on Computer Science and Service System (CSSS 2011), June 27-29, pp. 990–993 (2011), doi:10.1109/csss.2011.5974720

6. OMG: UML Superstructure, V2.1.2 (2007), `http://www.omg.org/spec/UML/2.1.2/Superstructure/PDF`

7. Woodcock, J., Larsen, P.G., Bicarregui, J., Fitzgerald, J.: Formal methods: Practice and experience. ACM Computing Surveys (CSUR) 41(4), 1–36 (2009)

8. Murata, T.: Petri Nets: Properties, Analysis and Applications. Proceedings of the IEEE 77(4), 541–580 (1989)

9. Pnueli, A.: The temporal logic of programs. In: Proceedings of the 18th Annual Symposium on Foundations of Computer Science (FOCS 1977), Providence, Rhode Island, USA, pp. 46–57. IEEE Computer Society (1977), doi:http://dx.doi.org/10.1109/SFCS.1977.32

10. Yu, H., He, X., Deng, Y., Mo, L.: Formal Analysis of Real-Time Systems with SAM. In: George, C.W., Miao, H. (eds.) ICFEM 2002. LNCS, vol. 2495, pp. 275–286. Springer, Heidelberg (2002)

11. Xiao, Z., Ming, Z.: A method of workflow scheduling based on colored Petri nets. Data & Knowledge Engineering 70(2), 230–247 (2011)

12. Wang, Y., Lin, C., Ungsunan, P.D., Huang, X.: Modeling and survivability analysis of service composition using Stochastic Petri Nets. The Journal of Supercomputing 56(1), 79–105 (2011)

13. Clarke, E.M., Emerson, E.A., Sistla, A.P.: Automatic Verification of Finite-State Concurrent Systems Using Temporal Logic Specifications. ACM Transactions on Programming Languages and Systems (TOPLAS) 8(2), 244–263 (1986)

14. Alur, R., Courcoubetis, C., Dill, D.: Model-checking for real-time systems. In: Proceedings of Fifth Annual IEEE Symposium on e Logic in Computer Science(LICS), Philadelphia, PA, USA, April 6-July 6, pp. 414–425. IEEE Computer Society (1990)

15. Jensen, K., Kristensen, L.M.: Timed Coloured Petri Nets. In: Coloured Petri Nets, pp. 231–255. Springer, Heidelberg (2009)

16. Yang, N., Yu, H., Sun, H., Qian, Z.: Modeling activity diagrams with extended Petri nets. Intelligent Automation and Soft. Computing 17(6), 725–735 (2011)

17. McMillan, K.L.: Symbolic model checking: an approach to the state explosion problem. Doctoral, Carnegie Mellon University, Pittsburgh, PA, USA (1992)

18. Yang, N., Yu, H.: Modeling and Verification of Embedded Systems Using Timed Colored Petri Net with Inhibitor Arcs. Journal of East China University of Science and Technology 36(3), 411–417 (2010) (in Chinese)

19. Kherraf, S., Lefebvre, É., Suryn, W.: Transformation from cim to pim using patterns and archetypes. In: Proceedings of the 19th Australian Conference on Software Engineering (ASWEC 2008), Perth, Australia, March 26-28, pp. 338–346. IEEE Computer Society (2008)

20. Zhang, T., Zhang, Y., Yu, X.-F., Wang, L.-Z., Li, X.-D.: MDA Based Design Patterns Modeling and Model Transformation. Journal of Software 19(9), 2203–2217 (2008) (in Chinese)

21. OMG: MOF 2.0 Query/View/Transformation (QVT) Specification (2008), `http://www.omg.org`

22. Ameedeen, M.A., Bordbar, B., Anane, R.: A Model Driven Approach to the Analysis of Timeliness Properties. In: Paige, R.F., Hartman, A., Rensink, A. (eds.) ECMDA-FA 2009. LNCS, vol. 5562, pp. 221–236. Springer, Heidelberg (2009)

23. Fernandes, J.M., Tjell, S., Jorgensen, J.B., Ribeiro, O.: Designing tool support for translating use cases and uml 2.0 sequence diagrams into a coloured petri net. In: Proceedings of the Sixth International Workshop on Scenarios and State Machines, Minneapolis, MN, USA, May 20-26. IEEE Computer Society (2007), doi:http://dx.doi.org/10.1109/SCESM.2007.1

24. Choppy, C., Klai, K., Zidani, H.: Formal verification of UML state diagrams: a Petri net based approach. ACM SIGSOFT Software Engineering Notes 36(1), 1–8 (2011)
25. López-Grao, J.P., Merseguer, J., Campos, J.: From UML activity diagrams to Stochastic Petri nets: application to software performance engineering. ACM SIGSOFT Software Engineering Notes 29(1), 25–36 (2004)
26. Yang, N., Yu, H., Sun, H.: Modelling UML sequence diagrams with aspect-oriented extended Petri nets. International Journal of Computer Applications in Technology 45(1), 57–65 (2012)
27. Ljungkrantz, O., Åkesson, K., Fabian, M., Yuan, C.: Formal Specification and Verification of Industrial Control Logic Components. IEEE Transactions on Automation Science and Engineering 7(3), 538–548 (2010)
28. Tun, T.T., Trew, T., Jackson, M., Laney, R., Nuseibeh, B.: Specifying features of an evolving software system. Software: Practice and Experience 39(11), 973–1002 (2009)
29. Larsen, K.G., Pettersson, P., Yi, W.: Uppaal in a nutshell. International Journal on Software Tools for Technology Transfer (STTT) 1(1), 134–152 (1997)
30. Cassez, F., Roux, O.-H.: Structural Translation from Time Petri Nets to Timed Automata. Electronic Notes in Theoretical Computer Science 128(6), 145–160 (2005)

Scalable Multipoint Videoconferencing Scheme without MCU

Xuan Zhang and Chongrong Li

Network Research Center, Tsinghua University
Beijing China, 100084
{zhangx,licr}@cernet.edu.cn

Abstract. Conventional multipoint videoconferencing is based on star-structure in which MCU is the bottleneck on bandwidth and signal processing power, causing this scheme not scalable and costly. In this paper, we propose one scalable multipoint videoconferencing scheme which based on the combination of mechanisms of the reflector/video-switching and P2P overlay multicast. The reflector works as the root of the overlay multicast tree and some stable participant nodes work as peer node to forward packets to children nodes. Each participant node sends two different bit rate video streams to the reflector with low and high definition resolutions using SVC, the reflector selectively distribute the upstream videos to all participants via overlay multicast tree. Only the speaking node's high bit-rate stream with high resolution and several low bit rate streams with low resolution are distributed by reflector to all users. This scheme avoids the reflector becoming the bottleneck on bandwidth and processing power. We implement the prototype system and compare it to the MCU based scheme on bandwidth usage. The results show that the scheme is feasible and could achieve more scalability than star-structure scheme based on MCU.

Keywords: multipoint videoconferencing, scalability, overlay multicast, SVC.

1   Introduction

For multipoint videoconferencing, all participants could communicate with each others using audio/video in real time. Now there are two main schemes to accomplish the multipoint communication functions. One is the MCU based star-configuration scheme which is known as H.323 system [1][2], the second is based on the video switching or reflector forwarding scheme[3] in which the media streams are distributed by the reflector without signal processing. The main difference between them is that whether the incoming audio/video packets are processed or mixed before distribution. In MCU scheme, the incoming videos/audios streams are decoded and mixed (composited) into one new frame, and to be re-encoded into new streams for distribution. This signal processing on MCU causes some problem, the first is the additional end-to-end delay caused by decoding and re-encoding. The second is that the decoding/re-encoding requires the MCU have considerable computational power

which cause the MCU costly and not scalable when the amount of incoming streaming is large.

The reflector or video switching based scheme need not decode/re-encode processing, but forward the audio/video packets directly, avoid the problems on additional delay and computational power. However it will cause large bandwidth consumption on reflector if not using SVC (scalable video coding or layered coding). If the video switched scheme is deployed based on multicast, the group communication based on multicast could save the bandwidth, but multicast is not well supported by ISP ubiquitously, so we need find one scalable scheme for multipoint videoconferencing via unicast..

In this paper, we propose one scalable scheme without MCU and without multicast supporting, in which the reflector and P2P based mechanisms [4][5][6][7] are combined to transmit the packets from sender participants to receiver participants. The reflector works as the root of the P2P overlay multicast tree and some of the participant nodes work as the peer node. The participants send their audio/video to the reflector, and the reflector selectively transmit the audio/video streams to all the participant nodes via the P2P overlay multicast tree.

To further improve the efficiency of bandwidth utilization, we employ simple SVC (scalable video coding) [8][9][10][11] strategy with two video resolutions. Most participants only need send their low resolution video streams with low bit rate to the reflector, only the chairman and the now speaking nodes send their high resolution video streams to the reflector. When the now speaking node changed, the sending and distribution of speaker's high resolution video stream would be changed too.

This scheme is especially suitable for those application scenarios, in which the amount of participant nodes is large, but the number of speaking node is relatively small, and the video resolution of the speaking node need be high with high bit rates ,otherwise the resolution should be low with low bit rates.

2   The Architecture of the Scalable Videoconferencing

In our scalable videoconferencing, the reflector and participant nodes are organized into one P2P overlay multicast tree, and the reflector works as the root of the overlay multicast tree. All the participant nodes send their audio/video streams to the reflector, and reflector selectively distribute the audio/video streams to all the participant nodes via overlay multicast tree. Fig.1 shows the topology of the overlay network for scalable videoconferencing.

The non-leaf nodes on the overlay multicast tree work as peer nodes to forward the audio/video streams from parent node to children nodes. The peer nodes have the functions of two aspects, one is the terminal function as one common videoconferencing terminal, the other is the P2P forwarding function which forwarding packets from parent node to children nodes.

For saving bandwidth, the participant nodes encode their video into two layered video streams with low and high definition resolution using SVC. Most participant nodes send their low resolution video stream with low bit rates, only the speaking nodes send their high resolution video streams to the reflector.
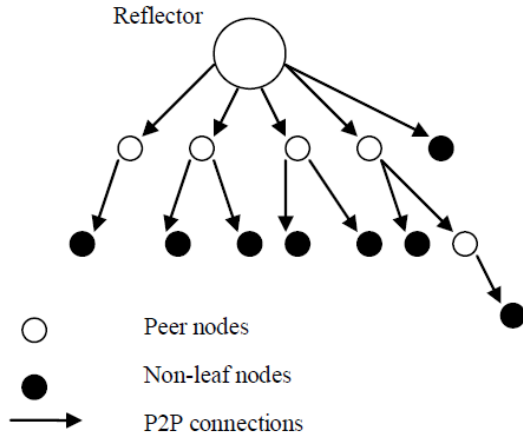
Fig. 1. The topology of the overlay multicast network for videoconferencing

3   The Audio/Video Data Exchanging and Distribution

The audio/video packets exchanging and distribution could be divided into two steps. The first step is the participant nodes capture/encode their audio/video and send their audio/video streams to the reflector. The second step is the reflector selectively distributes the incoming audio/video streams to all the participants via the P2P overlay multicast tree.
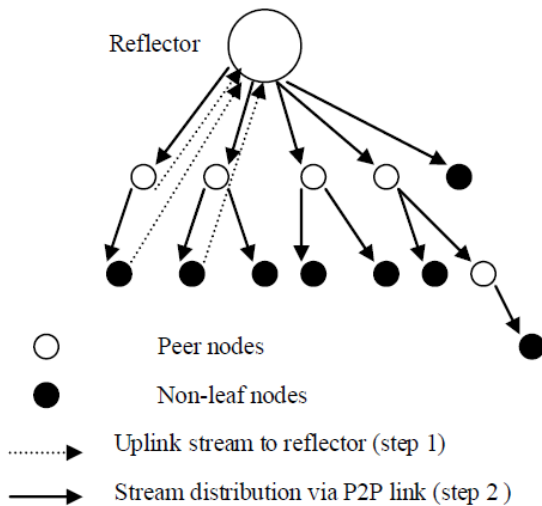
Fig. 2. The audio/video data exchange and distribution

Unlike the reflector based video switching architecture, the reflector in this paper does not distribute the incoming audio/video stream to all participants, only transmits

the streams to its children nodes, and the audio/video streams would be distributed to all the participant nodes via P2P overlay network as Fig.2 shows. This distribution way would decrease bandwidth and computational burdens on reflector. Additionally, two methods are employed for saving bandwidth. One is that most of the uplink video streams from the sender nodes to the reflector are low bit-rate with low video resolution, only the chairman and speaking nodes sending high definition video streams to the reflector with high bit-rate. The second it that the reflector only distributes the chairman/speaking node's video stream with high bit-rate and selectively distributes the no-speaking nodes' video streams with low bit-rates. Then the participant nodes could receive and display the multiple videos consists of high resolution of speaking/chairman node and low resolution of other non-speaking nodes. In most case, for the limitation of the screen area, the number of the low resolution video stream is up to 16. Figure.2 shows the audio/video data exchange and distribution.

When the speaking nodes are changed during the videoconferencing, the uplink video streams should be changed and the speaker's display window would be switched from the former speaker to the new speaker. To accomplish the function, the former speaking node should terminate its video stream sending and the new speaking node would send its high definition video stream to the reflector.

The switching process of the speaking nodes' high definition video streams distribution could be described as figure.3. In the process, when the former speaker finish its speaking, it would send its leave apply message (APPLY_LEAVE_MSG) to the reflector, then the reflector reply it with ACK message (ACK_LEAVE) to confirm the leave applying. And at the same time the reflector terminates the stream distribution for the former speaking node's high resolution video
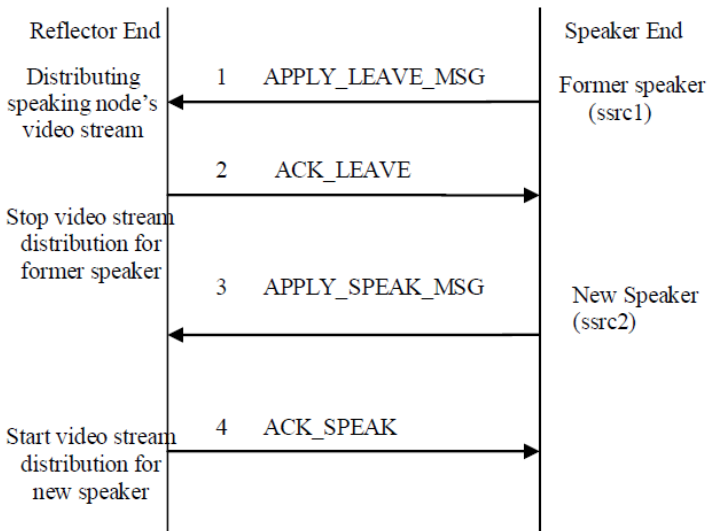
Fig. 3. The switching process of speaking nodes' high definition video streams distribution

When the new speaking node begins to speak, it would apply for speaking via sending applying message (APPLY_SPEAK_MSG) to the reflector. If the reflector agrees with the applying, the reflector would reply to the new speaker with ACK message (ACK_SPEAK) to agree the upstream transmission. Then the new speaker sends its high definition video stream to the reflector. And the reflector starts to distribute the new speaker's high definition video stream to all the participants.

By this switching way only one now speaking node's high definition video stream is distributed by the reflector at one time during the conferencing, the bandwidth could be saved largely comparing to distributing all participant nodes' high definition video streams.

4   The Session Management and Connection Building

4.1   Session Management

We establish the SMS (Session Management Service) to manage the conferencing session and participant nodes. The service may be run on the reflector machine. The SMS includes two functions. The first is to establish the virtual venue and assign parameters for each session, including the application port for audio/video communication channels and message channels, and enabling the data forwarding threads on reflector for audio/video stream packets and control messages. The second to store and manage the participant nodes information for each session, and provide service for the participant nodes' join and depart for each session. When one new node wants to join the session, it could get the session information and get the candidate parent nodes' list information from the session manage server for building connections.

The link relationships between the nodes are maintained by the participant nodes' recording and updating its parent node and its children nodes list.   During the session process of the videoconferencing, parent nodes and children nodes would send heart-beat message to each other periodically to keep the live link state.

4.2   Participant Nodes' Join and Connections Building

1) User registration and login

If one new node wants to join the videoconferencing, it need register and login to the session manage server, the manage server would authenticate the applying nodes and store the information of the new participant nodes such as name, IP address, computational power and uplink bandwidth ,which may be useful for the P2P connections building .

2) Apply for candidate parents nodes list

Secondly, the new participant node applies to the session manage server for the list of candidate parent nodes to make connection. The candidate parent nodes are the nodes which have joined the session with surplus computational power and uplink bandwidth for building new P2P connections. If the new participant node gets the list of the candidate parent nodes from the server, the information of the candidate parent nodes could be gotten accordingly.

3) Choose the suitable parent node

The new participant node would measure the end-to-end delay to the candidate parent nodes one by one, and the least delay node would be selected to measure the end-to-end bandwidth between it and the new node. If the bandwidth could meet the least bandwidth requirement MIN_BW for P2P data forwarding, the candidate parent node would be chosen as the new node's parent and the new P2P connection would be built. If the bandwidth is less than MIN_BW, the bandwidth between the new node and second least delay node would be measured. The bandwidth measurement would be continued until the end-to-end bandwidth meets the requirement.

The process of the new node's parent node chosen could be presented as following:

$$Min\{ \text{ delay}(n, p) \} . B(n, p) > MIN\_BW \tag{1}$$

In the expression, n denotes the new nodes, p denotes the candidate parent nodes, and B(n,p) presents the end-to-end bandwidth between n and p.

4) Connection building

After the parent node is chosen, we could build the connection between the new participant node and the parent node. The parent node would add the new node to its children node list, and forward the audio/video streaming data to the new child node. The new participant node is marked as one session member on session manage server and its node information may be useful for the later connection building as peer node. To keep the live link state between the parent nodes and children nodes, the heart-beat messages are sent between them periodically.

4.3   The Nodes' Depart

The departures of leaf nodes and non-leaf nodes are different. When the leaf nodes depart, they would not affect the other nodes for they have not children nodes, the only thing should do is that its parents node delete it from children node list and stop forwarding data to it.   The non-leaf nodes are peer nodes, if the peer nodes leave the session, it should send leave applying message to its parent node or session management server, the management server would launch the process to rebuild the P2P links for its children nodes. In practical, the peer nodes should not leave the session, if the node would leave the session with largely probability, it should not   be chosen as peer node but as leaf node.

5   The Prototype System Implementation and Experiments on Scalability Performance

We implement the prototype system according the new architecture proposed in this paper. The H.264 and G.711 are employed as the video/audio codec, and RTP/RTCP protocols are employed as the network transport protocols. In order to realizing the function that the speaking nodes could send both low and high bit-rate video streams with different resolutions, the raw video captured from the camera are duplicated and

encoded to two different compressed video streams respectively with low and high definition resolutions. To implement the P2P functions, the peer node terminal is designed as the combination of audio/video tools module and reflector module. The audio/video tools module accomplish the functions of capturing, encoding and sending the audio/video streams, and the functions of decoding and rendering the received audio/video streams at the same time. On the other hand, the reflector module receives the audio/video streams from the parent peer nodes or reflector, and forwards them to the children nodes. Fig.4 shows the composition of peer terminal modules.
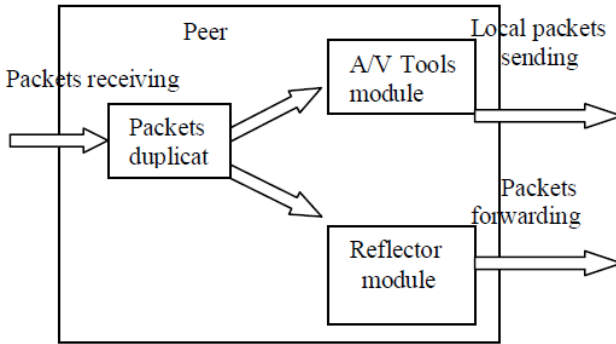
Fig. 4. The composition of peer terminal modules

We run the prototype system application in our laboratory and test it among several universities in Beijing city as figure 5. The experiments and applications show that the scheme is feasible when all peer nodes are stable. If the peer nodes leave the session during the videoconferencing, the children nodes would be affected with short audio/video switching. .

Fig. 5. The videoconferencing application of prototype system

We compare the scalability performance on bandwidth between MCU scheme and the new scheme in this paper by experiments. We deploy 20 end-host nodes and one MCU, the bit rate of one high definition video with 1080p resolution is about 3Mbps, and the bit rate of low definition with CIF resolution is about 100kbps. The layout is one large video window with 7 small video windows. In this case, only 7 low bit rate video streams and one large bit rate HD video stream need to be distributed by the reflector at one time.

For the MCU scheme, the re-encoded video are distributed to all the participants with high definition resolution, then the outgoing bit rate would be to about 3*20 = 60Mbps. For the P2P based scheme, suppose the overlay multicast tree have the depth 2 with the peer's average out-degree 4, the reflector need send 4 copies to its children peers with 4*(0.1*7+3) = 14.8Mbps. If the depth increased, the average out-degree would decrease with low bandwidth requirement for the reflector and peers. We could compute the average out-degree according to the node number and tree depth with logarithm. Table 1 show the changes of average out-degree and outgoing traffic load on reflector when the tree depth increases.

Table 1. The average out-degree and outgoing traffic load on reflector to tree depth (num 20)

Tree depth	Average out-degree of tree	Outgoing traffic load on reflector(/Mbps)
1	20	74
2	4	14.8
3	2.3	8.51

We could get that the overlay based scheme could decrease the traffic burden on the reflector when the depth of overlay multicast tree is larger than 1. In fact the out-degree of the peers is determined by the outgoing bandwidth of the participant nodes, some nodes with high outgoing bandwidth would have the large out-degree and the peers with low outgoing bandwidth would have the small out-degree.

From the comparison of the video streams traffic load on MCU and root reflector, we could see that the root reflector of P2P based scheme bears much less traffic burden than MCU. The bandwidth bottleneck and computational power bottleneck on MCU are avoided. This advantage would make the P2P based scheme much more scalable than MCU based scheme. Furthermore, the mechanism of encoding speaker nodes' video streams to dual bit-rate with low and high resolution and distributing them selectively could save the bandwidth largely.

6   Conclusion

In this paper we propose one scalable multipoint video conferencing scheme based on the combination of mechanisms of reflector forwarding and P2P overlay multicast distribution. The participant nodes' video are encoded into low and high bit-rate streams and selectively distributed by reflector according to requirements. Only the

now speaking nodes' video streams are transmitted using high bit rate streams with high resolution. We implement the prototype system and compare the traffic burden on reflector or MCU between the overlay multicast based scheme and MCU based scheme. The results show that the overlay multicast based scheme is feasible and could achieve more scalability than MCU based scheme on bandwidth and computational power. It should be noted that the peer nodes' stable is important for the reliability of the scheme. In practice, we choose the stable participant nodes as the peer nodes which would not leave the session during videoconferencing.

References

1. http://www.protocols.com/pbook/h323.htm
2. http://www.packetizer.com/ipmc/h323/standards.htmlreflector
3. http://www.cs.ucl.ac.uk/staff/s.bhatti/teaching/z02/reflecto r.html
4. Fiandrotti, A., Sheikh, A.M., Magli, E.: Towards a P2P videoconferencing system based on low-delay network coding. In: EUSIPCO 2012 (2012)
5. Liang, C., Zhao, M., Liu, Y.: Optimal Bandwidth Sharing in Multi-Swarm Multi-Party P2P Video Conferencing Systems. IEEE/ACM Transactions on Networking 19(6) (December 2011)
6. Hu, H., Guo, Y., Liu, Y.: Peer-to-Peer Streaming of Layered Video: Efficiency, Fairness and Incentive. IEEE Transactions on Circuits and Systems for Video Technologies 21(8) (2011); Conference Version Appeared in the Proceedings of NOSSDAV (August 2010)
7. Alalousi, A., Osman, A., Noori, S.: A Study on Video Conferencing using Overlay Network. European Journal of Scientific Research 59(3), 284–294 (2011) ISSN 1450-216X
8. Reichel, J., Schwartz, H., Wien, M.: Draft of Scalable Video Coding—Working Draft 4. Joint Video Team, Doc.JVT-Q201. Nice, France (2005a)
9. Eleftheriadis, A., Reha Civanlar, M., Shapiro, O.: Multipoint videoconferencing with scalable video coding. Journal of Zhejiang University - Science A 7(5), 696–705 (2006)
10. Abboud, O., Zinner, T., Pussep, K., Oechsner, S., Steinmetz, R., Tran-Gia, P.: A QoE-Aware P2P Streaming System Using Scalable Video Coding. In: IEEE International Conference on Peer-to-Peer Computing (2010)
11. Ponec, M., Sengupta, S., Chen, M., Li, J., Chou, P.A.: Multirate peer-to-peer video conferencing: A distributed approach using scalable coding. In: Proceedings of 2009 IEEE International Conference on Multimedia & Expo (ICME), New York (June 2009)

Reliability and Validity Assessment of Cluster Sampling on Multinomial Sensitive Question by Monte Carlo Simulation

Jia-chen Shi[1], Xiang-yu Chen[1], Yun-hua Zhou[1], Ying Fu[1], Lei Wang[2], and Ge Gao[1]

[1] School of Public Health, Medical College of Soochow University, Suzhou 215123, China
[2] Xuhui Center For Disease Control And Prevention, Shanghai 200030, China
gaoge@suda.edu.cn

Abstract. Objective: To assess the reliability and validity of RRT method of cluster sampling survey on multinomial sensitive question. **Method:** Monte Carlo Simulation was applied to simulate the investigation and compare the results got from RRT method and simulated direct questioning (criterion) by chi-square test. **Results:** 99 results are not statistically significant different from criterion (α=0.05). **Conclusion:** The method and corresponding formulae are valid and reliable.

Keywords: Multinomial Sensitive questions, Randomized Response Technique, Monte Carlo simulation, Validity, Reliability.

1   Introduction

Sensitive question is a question that a person doesn't want to be revealed to others in relation to the personal privacy [1-3], such as drug abuse, prostitution, premarital sex, AIDS and homosexuality. Sensitive questions are divided into two categories, multinomial sensitive questions and quantitative sensitive questions [4]. In dealing with surveys involving sensitive questions, one has to guard against the possibility of social desirability bias (SDB) [5]. Due to this bias, respondents often provide socially desirable responses rather than accurate ones. Researchers have suggested various methods to overcome this bias. Randomized Response Technique (RRT) is thought to be the most credible one [6-8].

This paper is a part of the study sampling design on sensitive questions investigation which is supported by National Natural Science Foundation of China [3, 9-11], in which we not only focus on statistical method of RRT model of cluster sampling on multinomial sensitive question investigation, but also examine the reliability and validity with Monte Carlo simulation method. Meanwhile, it provides a scientific and reliable method for the sensitive question investigation in large and complex sampling.

2   Methods

2.1   Concept of Cluster Sampling

The population were divided into several clusters (primary units), containing a certain number of second-stage units. Then we randomly drew part of clusters and investigate the second-stage units of these clusters with RRT model for multinomial sensitive questions [12-15].

2.2   RRT Model for Multinomial Sensitive Question

Suppose a sensitive question consisted of t kinds of incompatible categories and there was a randomization device which contained some cards written $0,1,2...,t$. The probability of drawing each kind of cards was $P_0:P_1:P_2...P_t(P_0+ P_1+ P_2+...+ P_t =1)$. If the respondents randomly picked up the *"0"* card, they should answer the number of sensitive question they truly belong to, otherwise, the number they picked up.

2.3   Statistical Formulae

Suppose there were N clusters and the ith cluster contained M_i second-stage units. In the investigation n clusters were randomly taken out. π_{ij} and π_j standed for the proportion of jth sensitive question of ith cluster and whole population. λ_{ij} denoted the probability of answering jth sensitive question in the ith cluster. a_{ij} meant the second-stage unit number of the ith cluster, jth category. Let $\hat{\pi}_j$ be the estimator of π_j. According to the total probability formula [4] and formulae given by Cochran W.G [5], we got:

$$\pi_{ij} = \frac{\lambda_{ij} - P_j}{P_0}, \quad i=1,2,...,n; j=1,2,...,t \tag{1}$$

$$a_{ij} = M_i \pi_{ij}, \quad i=1,2,...,n; j=1,2,...,t \tag{2}$$

$$\hat{\pi}_j = \frac{\sum_{i=1}^{n} a_{ij}}{\sum_{i=1}^{n} M_i}; \quad j=1, \cdots, t \tag{3}$$

$$v(\hat{\pi}_j) = \frac{1-f}{n\bar{M}^2} \frac{\sum_{i=1}^{n} a_{ij}^2 - 2\hat{\pi}_j \sum_{i=1}^{n} a_{ij}M_i + \hat{\pi}_j^2 \sum_{i=1}^{n} M_i^2}{n-1} \ ; \text{j=1,...,t} \tag{4}$$

2.4   Sampling and Investigation

There were 404 female sex workers (second-stage units) who were selected randomly in Xichang during May 2011 to July 2011 with cluster sampling method (19 districts were randomly selected from 57 districts). In the study, sample size was calculated by the formula of the relevant literatures [16-20]. Then RRT model was successfully applied to obtain their multinomial sensitive characters. One of the characters was the results of sexually transmitted diseases (STD) test, classified into five types: not checked (j=1), checked with uncertain results (j=2), checked with results unknown (j=3), negative results (j=4) and positive results (j=5). According to formulae (1)(2)(3)(4), we could get the estimate of five results($\hat{\pi}_j$) was 0.448894, 0.043131, 0.110406, 0.335066, 0.062502 and the variance: 0.002276, 0.000348, 0.000359, 0.335066, 0.000918.

3   Monte Carlo Simulation

3.1   Procedure

Monte Carlo method is a general term for computer experiments using random numbers. It is used in a variety of fields such as physics, chemistry, engineering, economics administration and mathematics. According to the relevant research [21-24], the random number produced by windows time function is proved to have perfect simulation calculating effects.

To fulfill the simulation, firstly, we built a simulated population consisted of 60 clusters (1200 simulated individuals, in 2.4 the sampling ratio is 1/3) which was close to the real situation and set the parameters of simulated population. Each parameter of multinomial sensitive characters of female sex workers' STD test was "0.448894, 0.043131, 0.110406, 0.335066, 0.062502" respectively. Then we regrouped the simulated individuals with random number generated by computer to define other parameters of simulated population, containing sensitive character number and cluster number. Finally, to simulate the procedure of cluster sampling and application of RRT model, we set cluster number as 20, number of simulated individuals as 400, 5 cards written 0, 5 cards written 1 to 5 respectively. New variables (sample size) were assigned as the proportion of cards in the RRT model and were sorted again with random number to finish one sampling.

3.2   Part of Program Codes

```
b(i) = 1
End If
Next i
For i = 0 To CountN - 1
Randomize Timer
M = Int(Rnd() * (CountN - 1 - i))
Temp = b(CountN - 1 - i)
b(CountN - 1 - i) = b(M)
b(M) = Temp
Next i
N = Int(CountN / ClusterNumber)
For i = 1 To ClusterNumber
For j = i * N - N To N * i
c(j) = i
Next j
Next i
For i = 0 To CountN - 1
xlSheet.Cells(i + 2, 1).value = a(i)
xlSheet.Cells(i + 2, 2).value = b(i)
xlSheet.Cells(i + 2, 3).value = c(i)
Next i
For i = 0 To SampleN / N - 2
For j = i + 1 To SampleN / N - 1
If c(i) > c(j) Then
Temp = c(i)
c(i) = c(j)
c(j) = Temp
End If
Next j
Next i
For i = 1 To 6
For j = 0 To SCN - 1
e2(i) = e2(i) + ((e3(j, i) / N) - P(i)) / P(0)
e22(i) = e22(i) + e32(j, i) / N
Next j
For j = 0 To SCN - 1
```

```
e4(i) = e4(i) + (((e3(j, i) / N - P(i)) / P(0) - e2(i) / SCN)
* (((e3(j, i) / N - P(i)) / P(0) - e2(i) / SCN)))
e42(i) = e42(i) + ((e32(j, i) / N - e22(i) / SCN) * ((e32(j,
i) / N - e22(i) / SCN)))
Next j
xlSheet.Cells(i + 1, (g * 4) - 2).value = (e22(i) / SCN)
xlSheet.Cells(i + 1, (g * 4) - 1).value = (((1 - SCN /
ClusterNumber) / (SCN * (SCN - 1))) * e42(i))
xlSheet.Cells(i + 1, (g * 4)).value = (e2(i) / SCN)
xlSheet.Cells(i + 1, (g * 4 + 1)).value = (((1 - SCN /
ClusterNumber) / (SCN * (SCN - 1))) * e4(i))
Next i
```

3.3   Results

We repeated sampling 100 times under .NET program and calculated the response value of RRT model and simulated direct questioning (DQ) according to formulae (1)(2)(3)(4). Part of the results is given in Table 1.

Table 1. Simulation Results of RRT and Direct Questioning

No.	$j=1$		$j=2$		$j=3$	
	RRT	DQ	RRT	DQ	RRT	DQ
1	0.425	0.435	0.045	0.045	0.115	0.1225
2	0.46	0.4575	0.045	0.0375	0.095	0.1075
3	0.5	0.48	0.04	0.04	0.115	0.1075
4	0.49	0.44	0.025	0.05	0.09	0.1125
5	0.46	0.475	0.045	0.04	0.15	0.14
6	0.45	0.4625	0.03	0.0325	0.145	0.1225
7	0.43	0.44	0.04	0.0475	0.135	0.1225
8	0.43	0.44	0.05	0.0475	0.115	0.1275
9	0.41	0.4325	0.025	0.05	0.13	0.1225
10	0.48	0.435	0.035	0.04	0.115	0.1075
…	…	…	…	…	…	…
99	0.485	0.5125	0.02	0.0275	0.115	0.1025
100	0.495	0.5125	0.04	0.0425	0.09	0.09

Table 1. (*continued*)

| NO. | j=4 | | j=5 | |
	RRT	DQ	RRT	DQ
1	0.345	0.335	0.07	0.0625
2	0.32	0.33	0.08	0.0675
3	0.305	0.3175	0.04	0.055
4	0.35	0.3525	0.045	0.045
5	0.29	0.29	0.055	0.055
6	0.335	0.325	0.04	0.0575
7	0.31	0.32	0.085	0.07
8	0.35	0.33	0.055	0.055
9	0.355	0.3225	0.08	0.0725
10	0.3	0.3425	0.07	0.075
…	…	…	…	…
99	0.33	0.3125	0.05	0.045
100	0.305	0.3	0.07	0.055

3.4   Reliability and Validity Assessment

Reliability is defined as the extent to which the measurement yields the same results when repeating the analysis, while validity refers to the degree to which the measurement describes the specific concepts that one is attempting to describe [25-29]. When assessing an investigation method, reliability and validity is the primary condition.

In this paper, Monte Carlo method was used to simulate the real situation and repeatedly sampling 100 times in same theoretical population. Then we obtained 100 samples and investigated everyone by using RRT model. At the same time we investigated same samples through simulated direct questioning and took its average value as criterion. Finally, to assess the validity and reliability, we compared the results got from RRT model with criterion by using chi-square test ($\alpha=0.05$). The results are given in Table 2.

Table 2. Results of Assessment of Reliability and Validity

Simulation Times	P-Value	Simulation Times	P-Value	Simulation Times	P-Value
1	0.9645	36	0.8597	71	0.9109
2	0.8251	37	0.6584	72	0.5268
3	0.4532	38	0.9523	73	0.3527
4	0.3347	39	0.0003	74	0.2603
5	0.4031	40	0.1872	75	0.23
6	0.3129	41	0.3889	76	0.5212
7	0.559	42	0.8427	77	0.7384
8	0.9395	43	0.0725	78	0.7601
9	0.3648	44	0.1027	79	0.4163
10	0.7974	45	0.9874	80	0.2713
11	0.518	46	0.5621	81	0.986
12	0.1809	47	0.9427	82	0.9765
13	0.0602	48	0.19	83	0.6183
14	0.2995	49	0.227	84	0.2843
15	0.643	50	0.7819	85	0.6361
16	0.7231	51	0.1832	86	0.8822
17	0.6785	52	0.2783	87	0.215
18	0.1795	53	0.5647	88	0.9743
19	0.4856	54	0.6746	89	0.1102
20	0.1989	55	0.3109	90	0.8158
21	0.9755	56	0.2644	91	0.9345
22	0.9545	57	0.2329	92	0.2581
23	0.4271	58	0.6922	93	0.2753
24	0.7571	59	0.1169	94	0.4841
25	0.5135	60	0.9684	95	0.1742
26	0.0953	61	0.5037	96	0.3886
27	0.4581	62	0.535	97	0.3571
28	0.1881	63	0.8798	98	0.6591
29	0.7531	64	0.9197	99	0.1102
30	0.8337	65	0.2698	100	0.2753
31	0.914	66	0.1929		
32	0.5561	67	0.8234		
33	0.3529	68	0.6311		
34	0.7257	69	0.8505		
35	0.6217	70	0.6358		

From Table 1, 99 results (RRT) are not statistically significant different from the criterion (simulated direct questioning), indicating that the method and formulae are of high validity (Chi-square test, $\alpha=0.05$). Simultaneously, nearly all the results are very close to the same level, illustrating that the method and formulae are of high reliability.

4   Discussion

Monte Carlo method is a kind of computational algorithm relying on repeated random sampling to obtain numerical results. It is often used in physical and mathematical issues. In recent years, the method is even used in physics, pharmacy, biology, economics, general statistics, and econometrics. However, there are also some problems that exist in the application of a Monte Carlo method [9].One is that the calculation for large values of simulations may not be practical. Another problem is that high quality of algorithms for random number is necessary for the obtainment of reliable simulation results. For this trial is aiming at obtaining estimated mean value, 100 simulations are available [10]. Otherwise, random numbers generated by computer time function (section 3.1) proved to have perfect simulation effects. Therefore, simulation results in this paper are reliable.

Effectiveness and efficiency of investigation methods and formulae are highly required. However, when evaluating whether an investigation method is feasible, numerous repeated researches are needed to guarantee the accuracy of assessment. Monte Carlo method based on powerful calculating ability of computer, is the most appropriate way to solve the problem.

Randomized Response Technique (RRT) was firstly introduced by Warner in 1965[11]. Subsequently, Warner's method was modified by Greenberg [12], Raghavrao [13], Franklin [14], Arnab [15], Kim and Warde [16]. RRT methods prevent against the possibility of disclosure to third parties and diminish the need to give socially desirable answers by totally protecting the respondents' privacy [17]. So when conducting sensitive questions investigation, RRT methods should take precedence.

Reliability and validity is important indicator to assess an investigation method. A recent meta-analysis of RRT research compared the RRT with other interview methods such as direct questioning [18]. Another study investigated the viability of the RRT with respect to the measurement of socially sensitive attitudes [19]. In our recent research [20-22], reliability or validity of different RRT models was accessed, but few studies evaluate RRT through a large number of repeated investigations. In this paper, Monte Carlo method, employed under cluster sampling survey, designed for multinomial sensitive question, indicated that the method and formulae are of high reliability and validity, which had certain innovation and high-applied value, providing scientific basis for policy makers to draw up effective control of HIV/AIDS among high risk group and future survey of sensitive questions.

References

1. Liu, W., Gao, G., Li, X.D.: Stratified Random Sampling on Simmons Model for Sensitive Question Survey. J. Suzhou University Journal of Medical Science 30(4), 759–762, 776 (2010)
2. Sat, G., Javid, S., Supriti, S.: Mean and Sensitivity Estimation in Optional Randomized Response Models. Journal of Statistical Planning and Inference 140, 2870–2874 (2010)
3. Li, W., Gao, G., He, Z.L.: Statistical Methods of Two-Stage Sampling on Simmons Model for Sensitive Question Survey with and Its Application. Studies in Mathematical Sciences, 46–51 (2011)
4. Su, L.J.: Advanced Mathematical Statistic. Beijing University Press (2007)
5. Cochran, W.G.: Sampling Techniques, 3rd edn., vol. 87, pp. 93–95, 432, 491. Wiley, New York (1977)
6. Wang, J.F., Gao, G., Fan, Y.B., Chen, L.L., Liu, S.X., Jin, Y.L., Yu, J.G.: The Estimation of Sample Size in Multi-stage Sampling and Its Application in Medical Survey. Applied Mathematics and Computation 178, 239–249 (2006)
7. Dai, Y., Ji, K.: A New Method to Produce Normal Distribution Random Number-Based on the Windows Time Function. Surveying and Mapping of Geology and Mineral Resource 20(2), 7–8 (2004)
8. Terje, A., Bjørnar, H.: Reliability and Validity of Risk Analysis. Reliability Engineering and System Safety 94, 1862–1868 (2009)
9. Esward, T.J., Ginestous, A., Harris, P.M.: A Monte Carlo Method for Uncertainty Evaluation Implemented on a Distributed Computing System. Metrologia 44, 319–326 (2007)
10. Bonate, P.L.: A Brief Introduction to Monte Carlo Simulation. Clinical Pharmacokinetics 40(1), 15–22 (2001)
11. Warner, S.L.: Randomize Response: A Survey Technique for Eliminating Evasive Answer Bias. Journal of the American Statistical Association 60, 63–69 (1965)
12. Greenberg, B.D., Abul-Ela, A.L.A., Simmons, W.R., Horvitz, D.G.: The Unrelated Question Randomized Response Model: Theoretical Frame Work. American Statistical Association 64, 520–539 (1969)
13. Raghavrao, D.: On an Estimation Problem in Warner's Randomized Response Technique. Biometrics 34, 87–90 (1978)
14. Franklin, L.A.: A Comparison of Estimators for Randomized Response Sampling with Continuous Distributions from A Dichotomous population. Communications in Statistics-Theory and Methods 18(2), 489–505 (1989)
15. Arnab, R.: Optional Randomized Response Techniques for Complex Survey Designs. Biometrical Journal 46, 114–124 (2004)
16. Kim, J., Warde, W.D.: A Stratified Warner's Randomized Response Model. Journal of Statistical Planning and Inference 120, 155–165 (2004)
17. Gerty, J.L.M., Lensvelt, M., Hennie, R.B.: Evaluating Compliance with A Computer Assisted Randomized Response Technique: A Qualitative Study Into the Origins of Lying and Cheating. Computers in Human Behavior 23, 591–608 (2007)
18. Lensvelt, M., Hox, J.J., Mass, C.J.M.: Meta-analysis of Randomized Response Research. Sociological Methods and Research 33, 319–348 (2005)
19. Ivar, K.: Estimating the Rrevalence of Xenophobia and Anti-semitism in Germany: A Comparison of Randomized Response and Direct Questioning. Social Science Research 41, 1387–1403 (2012)

20. Gao, G., Fan, Y.B.: Stratified Cluster Sampling and Its Application on the Improved RRT Model for Sensitive Question Surve. Suzhou University Journal of Medical Science 28(5), 750–754 (2008)
21. Gao, G., Fan, Y.B.: The Research of Stratified Cluster Sampling on Simmons Model for Sensitive Question Survey. Chinese Journal of Health Statistics 25(6), 562–565, 569 (2008)
22. Wang, L., Gao, G., Yu, M.R.: Stratified Two-stage Cluster Sampling and Its Application on the Greenberg Model for Sensitive Question Survey. Chinese Journal of Health Statistics 28(1), 37–39 (2011)
23. Miller, K.W., et al.: Estimating the Probability of Failure when Testing Reveals no Failures. IEEE Trans. Software Engineering 18(1), 33–43 (1992)
24. Moller, K., Paulish, D.: An Empirical Investigation of Software Fault Distribution. In: Proc. 1st IEEE Int'l Software Metrics Symp., pp. 82–90 (1993)
25. Ostrand, T.J., Weyuker, E.J.: The Distribution of Faults in a Large Industrial Software System. In: Proc. ACM Int'l Symp. Software Testing and Analysis, pp. 55–64 (2002)
26. Podgurski, A., et al.: Automated Support for Classifying Software Failure Reports. In: Proc. 25th Int'l Conf. Software Engineering, pp. 465–475 (2003)
27. Robert, C.P., Casella, G.: Monte Carlo Statistical Methods, 2nd edn. Springer (2000)
28. Singh, H., Cortellessa, V., Cukic, B., Guntel, E., Bharadwaj, V.: A Bayesian Approach to Reliability Prediction and Assessment of Component Based Systems. In: Proc. 12th Int'l Symp. Software Reliability Engineering, pp. 12–21 (2001)
29. Yacoub, S., Cukic, B., Ammar, H.: Scenario-Based Reliability Analysis of Component-based Software. In: Proc. 10th Int'l Symp. Software Reliability Engineering, pp. 22–31 (1999)

Control Pass Check Technology of Switches Based on Expert System

Zhonghua Cai[1], Tingting An[1], and Hongtu Zhang[2]

[1] Beijing University of Chemical Technology, Beijing, China
[2] Northeast Dianli Universities, Jilin, China
`caizh@mail.buct.edu.cn, {att200851012,zhtpld}@163.com`

Abstract. To improve the security and stability of the operation in High-end Switch, a technology of Control Pass Check (CPC) is designed in this paper. Based on the function requirement and the features of CPC, the frame structure and functional module of the technology is built in this paper. At the same time, the CPC plays an important role in the Diagnostic Mechanism of the Switch System. Most of the data for Diagnostic comes from CPC and the action of CPC needs Diagnostic to perform. Then the basic principles and processes of CPC are built in this paper, and to optimize performance, the CPC use hardware interrupt to improve its real-time.

Keywords: Security, Switches, Expert system, Control Pass Check.

1 Introduction

The 21st century is the age of all kinds of abundant information. Information technology and industry is growing rapidly which presents an unprecedented prosperous picture [1-3]. And the events that endanger information security are occurring constantly, so that maintaining information security situation is most severe. Information security mainly includes the following four aspects: information equipment security, data security, content security and behavior-based safety. Hereinto, the information system hardware structure security and the operating system security is the basis of information system security, password and network security technology is the key point [4]. Only adopted the security measures from the bottom of information system hardware and software, on the whole, can more effectively ensure the security of information system [5].

As switches are the hub of information communication, thus the high-end switches for the provincial and national network domain will be bearing more and more pressures and responsibilities. However, presently, all the high-end switches, which operate at an overload for a long time, will inevitably lead to a software or hardware breakdown, and the breakdown will often lead to the mistake of the entire system, or even the business interruption, which then will bring great economic losses. Currently, however, switches malfunction is often directly troubleshooting by the technical personnel to the site or by remote control. No matter which method, it all

needs human input, so that product maintenance has been accounted for a significant proportion of total input. The artificial intelligent fault diagnosis technology of current mainstream switches manufacturers (such as Cisco and other companies) is not mature yet [6]. If there is a new highly efficient artificial intelligent fault diagnosis technology which the equipment can deal with some problems with themselves, it will greatly improve the quality of the product. In order to avoid unnecessary loss, experts and scholars spend a lot to the self-diagnosis function of switches. This paper makes some corresponding improvement—when the equipment appears some fault it can ruled out by itself—for equipment which is aimed at fault diagnosis system. Diagnosis system is composed of expert system and timing processor system.

2   Theory Introductions

2.1   The Diagnostic System

The diagnostic system includes timing processor system and expert system. The timer processor system is responsible for equipment fault collection and finishing work. And the expert system is responsible for fault analysis and ruled out. The timer processing system is responsible for sending test message to auxiliary equipment regularly. According to the response message, check whether the main equipment and various auxiliary equipment real-time communication is normal. Collected the fault information and then sent to the expert system through the way of writing queue. The sending movement of detecting message in the control channel on the main control board is processed in timed processor. And its closing action is articulated in the hard interrupt to be processed. Next in the interface board, the closing action of interface board is articulated in the soft interrupt to be conducted. The diagnostic system is shown in figure 1.
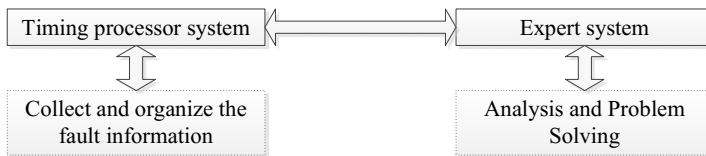
Fig. 1. The diagnostic system

2.2   Expert System

The expert system is an application system with a large number of specialized knowledge and experience. It applies to the artificial intelligence technology and computer technology. It accords to the knowledge and experience of the experts to reasoning and judgments. Simulate human experts in the decision-making process, in order to solve those needs human experts to deal with complex problems. In this paper, the expert system is based on the rules and logic. The characteristics of the expert system are predictive, diagnostic, debug and maintainability. Predictive namely to infer object according to the object's past and present situation of the future

evolution of the results. Diagnostic object is based on the input information to find faults and defects. Debug namely to determine the fault eliminate scheme by itself. Maintainability is specified and implemented corrective plan of a kind of fault.

Figure 2 shows the structure of expert system.
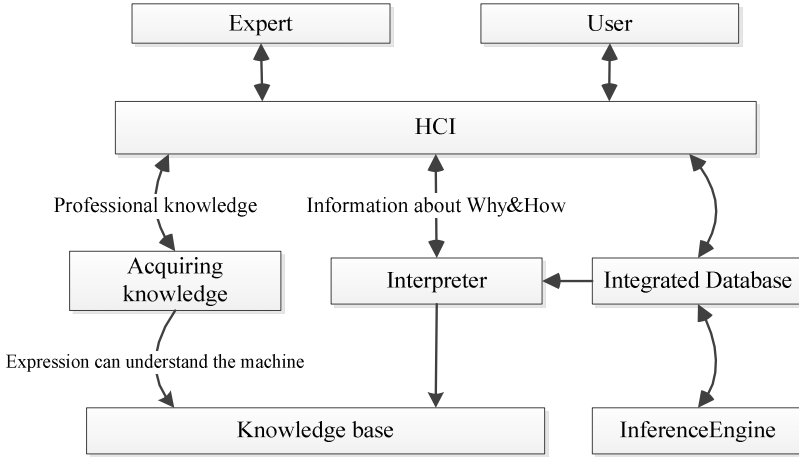
Fig. 2. The structure of expert system

2.3   The Timer System

The timer system is a multitasking timed reminder system. Diagnosis mechanism is a relatively complicated mechanism, and control channel detection is one module of the diagnostic function services and also one kind of diagnostic technique. The control channel detection described in this paper refers to monitoring the situation of communication between the high-end switches' mother board and others' ports board at real time, and finding the abnormal data, then analyzing and preserving the data, finally dealing with the data accordingly. There are not only very large and various data packets through the high-end switches, but also the different priorities, even if the most powerful processors cannot process all the instant messages at a moment. However, the control channel detector has the characteristic of strong real-time, thus will process the data through the highest priority hardware interrupt request. The final test results show that the design is the most reasonable. And it not only meets the requirement of real-time, but also ensures the realization of the diagnostic function.

3   Key Technologies

The core of timing processor system is to detect the sending and receiving of packets. The process flow is shown in Figure 3. The master device is responsible for regularly sending packets detected on the handshake to each auxiliary device and the sending interval is 300ms. Auxiliary device uses the hard interrupt to receive the detection message of master device and uses the soft interrupt to send a reply message to control board. The handshake number of detection is maintained by the auxiliary

device. When the auxiliary device receives a packet, the handshake number will be plus one, then the master device receives the reply message of auxiliary device using the hard interrupt. Because of the latency for the reading time of soft interrupt queen, thus there is a time difference when the master device receives the packets.

Master device can make a real-time detection for all auxiliary device sending handshake detection packets with the polling method, as shown in Figure 4. The recording cycle of timing processor system is 10s and the transmission cycle of handshake detection packet of timer task is 400ms, which means the system will send 25 pieces of detection packets within a record period. All kinds of detecting abnormal information that main device records are configured in accordance with the following rules:

(1)   When the number of lost packets is 10 in 10 seconds in a discontinuous manner, the internet will conduct packet loss without notifying diagnosis module, only to print message to tell the user.

(2)   When the number of lost packets is 15 in 10 seconds, the internet will give an alarm. Firstly, the loss time will be warned with a sign of log. Then it gives an early warning to diagnosis module.

(3)   When the diagnosis module lost 25 data packets in 10 seconds, the internet will put forth its hand in token of failure.
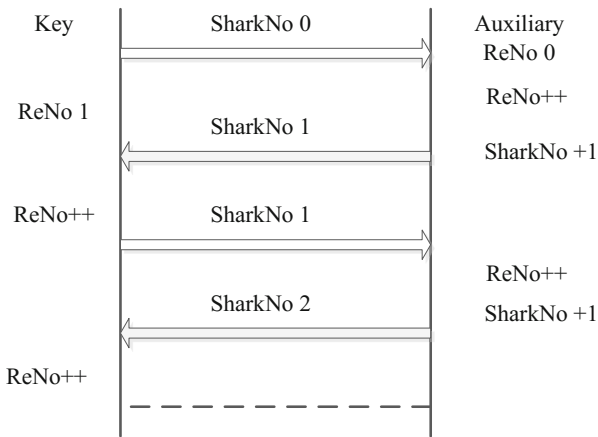
Fig. 3. Packet processing
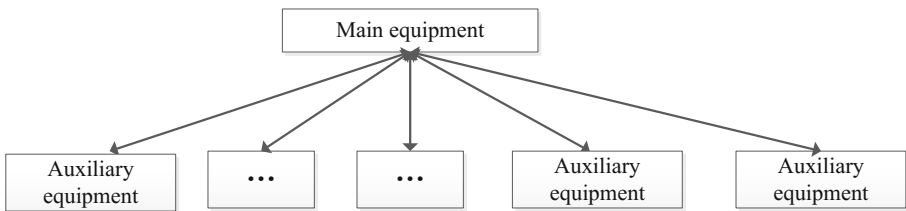
Fig. 4. Send and receive packets flow

4   Achieve the System Architecture

Aiming at the characteristics of Comware V7 platform of the company, the system of the product side is divided into two parts, user mode and kernel mode. Comware software platform is a company's core software platform. Comware software platform builds the foundation on a full range of IP networks products for a company. Based on the latest platform version, all function associated with the user operational will set in user mode, while other function needing to be transparent to the user will set in kernel mode. This could not only increase the security of the system, but also make the system structure more clear and reasonable. For the control channel detection module, a variety of commands that users send are directly related to the user action, so placed in the user mode. For the operations that are transparent to the user are set in the kernel mode implementation. Figure 5 is the functional block diagram of Control Pass Check.
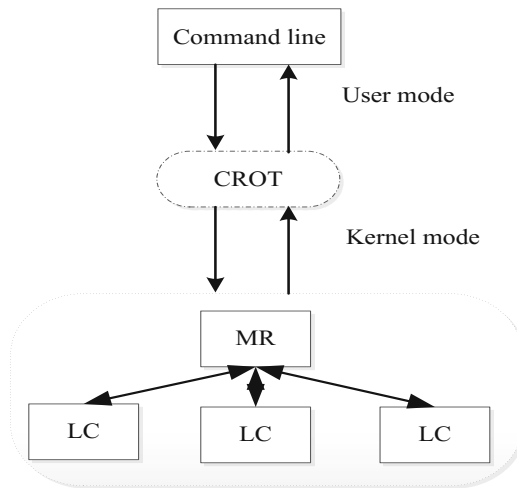
Fig. 5. System function block diagram

Command line: Operation of various commands is issued by the user, including setting and querying controlling channel detection switches and results. CROT (Controller Input/output communication control interface): Provide interface for transmission of data and command of user mode and kernel mode. Through this interface, the user mode sends orders to kernel mode, while kernel mode will return the result data back to the user mode through this interface. The CROT is organized and displayed by the user mode. The interface is implemented by the platform. Drive side can only see interfaces. MR (Main Board): Master control board would open or close their control channel detection through the command sent by the user mode. Meanwhile the main board sends detection packets to all of the interface board constantly during the initialization of the main control board by creating tasks and a timer. When detecting abnormality, it will inform the diagnostic module to make the

appropriate treatment. LC (Line Connector Interface Board): being similar to the basic functions of the main control board MR, LC and MR use the semaphore mechanism to keep pace.

Control channel check is only used to detect the passage between active main board and interface board, not used to detect the control channel of the master, slave control board and network board. Standby main board is the backup of main control board. When the main control board breaks down, make the changeover between main board and standby main board directly. The active main board (AMB) does restart as the standby main board (SMB) and at the same time SMB directly promotes to be AMB. The relationship between AMB and SMB means that they would not affect the business forwarding flow of the entire equipment package, so there is no need to do testing between AMB and SMB. Due to the forwarding characteristics of screen board, as long as there is a screen board exists, it will not affect the forwarding flow of the equipment package. The relationship of the screen board in the entire forwarding process is shown in Figure 6. In IRF（Intelligent Resilient Framework) mode control channel detection is carried out in the frame, and the abnormality detected by the spare frame is sent to the main frame of the main control board through inter-board communication for processing.
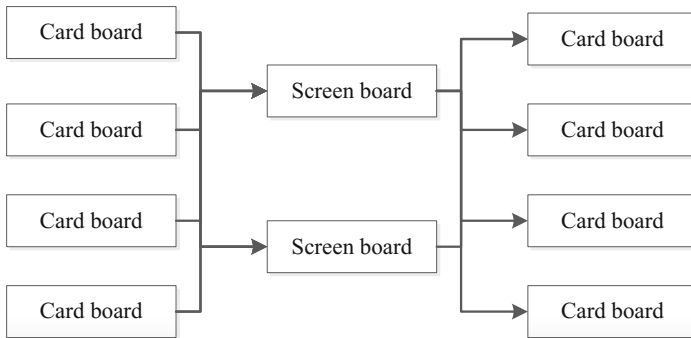
Fig. 6. Processing of Screen board

5 Data Processing and Analysis

5.1 Timer Data Processing Tasks

Message sending by main control board is finished in the control channel detection task timer. And timer control channel detection processing is the key of message transmission and abnormal information collection and processing. Timing processor operates as follows:

(1) The main target of the timer operation include: whether control channel detection function is manually disabled, whether the channel has started, handshake timeout counter, the number of packet loss in 10s record cycle, the record time value in 10s record cycle of packet loss, counting once in 400ms to a total of 25 times in 10s, detecting messages and detecting packet length;

(2)   Firstly start the timer with 400ms of start-up cycle, every time you start the following actions will be done once;

(3)   Check the handshake timeout counter, when the main control board or interface board receives the packet handshake timeout counter will be set to zero, so if the handshake timeout counter is not zero in this check, it indicates that the handshake cycle has not received the packet, thus the count of lost packet plus one. In the same time conduct the simultaneous detection of packet loss count, if the packet loss count in 10s reaches the limit of 10, it will notice the packet loss processing module to handle, and at this point the count may not have been finished in the 10s or the number of packet loss is less, so the packet loss processing module would only make packet loss notification processing;

(4)   The record time value in 10s record cycle of packet loss plus one, every time you start the timer, this object will be incremented by one, so when the count reaches the threshold of 25, it indicates the counting to 10s and a record cycle finishes, then clear the object and also the number of packet loss target during the 10s recording period;

(5)   Handshake timeout count plus one, then send handshake detection packets (at this moment interface board doesn't send test packets) and check the handshake timeout counter, if the count reaches the threshold of 25, it illustrates that the handshake with the board fails, under the circumstance the first thing to do is recording the failed veneer, while writing queue notify the handshake failing processing module to handle in the task; if the handshake timeout counter reaches the abnormal warning threshold count of 15, write queue to notice warning processing module to deal with in the task.

For the above action control board and interface board will execute once during the period of 400 MS, but the packets sending of interface board is not performed in the timer.

Compare with the packets sending of control board, the packets sending of interface board is only real-time; send the ACK packet immediately while receiving the handshake detection packet of main control board and packet loss would not be cared. Timeout processing and other operations are all performed in the main control board.

5.2   Analysis of Test Results

The system regards 10 seconds as a record period, and the transmission cycle of the task timer is 400ms. The system can send 25 pieces of text message within a record period. All kinds of detecting abnormal information that main control board records are configured in accordance with the following rules:

(1)   When the number of lost packets is 10 in 10 seconds in a discontinuous manner, the internet will conduct packet loss without notifying diagnosis module, only to print message to tell the user.

(2)   When the number of lost packets is 15 in 10 seconds, the internet will give an alarm. Firstly, the loss time will be warned with a sign of log. Then it gives an early warning to diagnosis module.

(3)   When the diagnosis module lost 25 data packets in 10 seconds, the internet will put forth its hand in token of failure.

The operations of main control panel interface board on packet warning and handshake failure are mainly handled by AI diagnosis. The process of handshake failure is as follows:

First, reported exception information is proposed to determine the type of problem. Currently supported questions include three types: main control board failure, interface board failure and unknown failure.

The three fault types are determined by the following rules:

(1)   Its an unknown failure type, if the number of interface board for control channel detection is 1;

(2)   Comparing the current number of interface board for control channel detection and the reported number of abnormal interface board, if the former is larger than the latter, then it's an interface board failure;

(3)   Comparing the current number of interface board for control channel detection and the reported number of abnormal interface board, if the former is equal to the latter, then it's a main control board failure.

Secondly, according to the fault types obtained to implement relative action, in the same time need to confirm the error information again in order to ensure shook hands abnormal indeed exists between main control board and interface board reported. Because the malfunction caused by the control channel detection creates highest processing levels and results in relatively large losses thus should be careful. Different processes will be performed depending on different types of failure in the phase of implementation action:

(1)   When current fault is the main control board failure, to notify AI diagnostic information recording module, and AI diagnostic information recording module will collect fault information, and then call the shots standby switchover. If the fault occurs in the active main board of backup box, then can directly restart the board and notify the AI diagnostic action execution module to perform the appropriate action;

(2)   When current fault type is the interface board type, then notify AI diagnostic information recording module, AI diagnostic information recording module will collect fault information, and then notify the AI diagnostic action execution module to perform the restart operations for the board;

(3)   When current fault is unknown failure, to notify AI diagnostic information recording module, AI diagnostic information recording module will collect fault information, and then the final diagnosis module will deal with it differently according to the reported fault information.

5.3   User Mode and Kernel Mode Communication

User mode needs to convey commends to kernel mode including startup and shutdown control channel detection switches, check the current detection status, etc. The commands sent by user mode are transmitted to kernel mode through the interface CROT, for example, users send the command of opening the control channel

detection switch, which is the command issued only to the main control board, and then the main control board will issue the command. This is very different from other module command issued. First, the main control board opens all manual inhibit switches of main control board and interface board, and then starts its own control channel detection function, and finally opens all control channel detection interface of interface boards through the communication way among general boards.

6   Results

In practical applications, the control channel detector will find a lot of the machine failures and respond promptly, also restore the unnecessary losses. Due to the equipment failure probability is very small in a short time, and diagnostic module will record all devices abnormal in real time, so using the method of querying diagnostic information recording, checking the control channel detection test results. The following are the experimental results of opening and closing control channel detection switches, testing the current status of the equipment and querying the action performed by control channel detection:

Open or close the control channel detection switches:
_set driver ibc control-pass-check enable
Result:
The control pass check is starting;
_set driver ibc control-pass-check disable slot 3//Close the 3rd slot detection switch
Result:
The control pass check has stopped slot 3.
The above results show that the No. 3 slot switch were opened and shut down successfully.
Query the current state of the Control Pass Check switch
_display driver ibc control-pass-check state slot 0//Check the detection status of the MPU
Result：
slot	state	failcnt	Ischecking	Isenablechking
1	normal	0	N	Y
2	normal	1	Y	Y
3	normal	0	N	Y
The above results show that the 3rd slot interface board has stopped testing, and the 2nd slot interface board found a dropout error exception, in the same time, the No. 1 slot is a backup master which are not involved in the control channel detection.
c.   query control channel recording actions performed
_display driver diag hardware-diag-action information
Result:
Recent records:
--Executed action records: -
Slot 3:
2013-01-01 00:45:46 hot-rebooted by CONTROLPATH

Reason: CTRL_PATH, the action executed 1 times

The above results indicate the 3rd slot interface board has a thermal restart in at 0:45:46 on January 1, 2013, and the reboot is triggered by the control channel detection were restarted 1 times.

7 Conclusion

The experimental results showed that the device has been successfully performed different kinds of commands by the user mode, and the control channel detection function is operating normally and can detect the normality of communication between main control board and interface board at the real-time and the real-time is high. It can also make timely adjustment and decision when the abnormal happened, then to maintain the normal business processes of the equipment to meet the design requirements.

References

1. Zalis, K.: Application of expert systems in diagnostics of high voltage insulating systems. In: Proceedings of the 2004 IEEE International Conference on olid Dielectrics, ICSD 2004, July 5-9, vol. 2, pp. 691–694 (2004)
2. Ang, J.K., Leong, S.B., Lee, C.F., Yusof, U.K.: Requirement engineering techniques in developing expert systems. In: 2011 IEEE Symposium on Computers & Informatics (ISCI), March 20-23, pp. 640–645 (2011)
3. Bilal, K., Mohsin, S.: Muhadith: A Cloud Based Distributed Expert System for Classification of Ahadith. In: 2012 10th International Conference on Frontiers of Information Technology (FIT), December 17-19, pp. 73–78 (2012)
4. Wangphanich, P.: A simple Web-Based Expert System for a supplier assessment: A case of a JIT production environments. In: 2011 International Conference on System Science and Engineering (ICSSE), June 8-10, pp. 96–100 (2011)
5. Yuan, Y., Ling, H.: Designing an IDSS supported by heterogeneous Distributed Expert System. In: 2010 The 2nd IEEE International Conference on Information Management and Engineering (ICIME), April 16-18, pp. 439–443 (2010)
6. Girgis, A.A., Johns, M.B.: A Hybrid Expert System for Faulted Section Identification, Fault Type Classification and Selection of Fault Location Algorithms. IEEE Power Engineering Review 9(4), 56–57 (1989)

Evaluation Strategy and Translation
of Environment Calculus

Shin-ya Nishizaki

Department of Computer Science, Tokyo Institute of Technology
nisizaki@cs.titech.ac.jp

Abstract. The environment is one of the execution states in programming language processors. The first-class environment is a reflective programming facility which enables us to use meta-level environments as object-level entities, and inversely, to use the object-level environments as meta-level entities. We named a lambda calculus with first-class environments the environment calculus, and studied various systems of the environment calculus. PCF (Programming Computable Functions) is a simply typed lambda calculus with natural numbers, Boolean values, and recursive operators, which was proposed as a formal framework for the semantical study of evaluation strategies.In this paper, we introduce an extended system of PCF with first-class environments and study evaluation strategies, such as call-by-name and call-by-value strategies.

Keywords: lambda calculus, reflective programming, first-class environment, PCF, evaluation strategy.

1   Introduction

An environment is the mapping from variables to bound values in the research area of a programming language. The environment is a fundamental notion in theoretical and practical studies of functional programming languages.A first-class entity in a programming language is an object which can be passed to a function as an actual parameter and returned from a function as a result. For example, integers are first-class entities in many programming languages including programming language C. However, functions are not first-class entities in C but in functional languages like Haskell and Scheme [10, 3] and scripting languages like JavaScript.

Some implementations of Scheme, such as MIT/GNU-Scheme [3],provide first-class environments. In the terminology of reflective programming, we can reify a runtime environmentand reflect the reified environment to the runtime with Scheme's built-in procedure `the-environment`and`eval`, respectively (see Fig. 1)

We have studied the first-class environment in the framework of the lambda calculus [5, 6]. In order to formalize environments, we adopt a method called explicit environments, proposed by P.-L. Curien et al [1]. In this paper, we refer to this idea as"environment-as-substitution."

In the traditional lambda calculus, a substitution of variables is defined in the meta-level, but in the lambda-sigma calculus, based on the idea of explicit substitution, it is

defined in the object-level, in other words, the substitution operation is defined as a part of the reduction of the calculus. Accordingly, environments in the lambda-sigma calculus are formalized as substitutions defined in the object-level. However, the lambda-sigma calculus does not give us a first-class environment since the terms and the environments are categorized as different syntax classes. In our previous works [5, 6], we mix the terms and the environment together in the same syntax class, which enables us to have first-class environments in the calculus. We call a lambda calculus with a first-class environment an *environment calculus*.
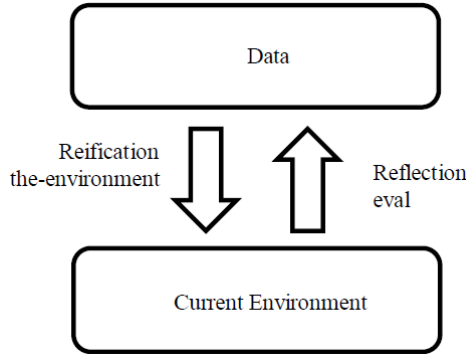
Fig. 1. Scheme's built-in procedure the-environmentandeval

In paper [5], we proposed a simply typed environment calculus and investigated fundamental properties, such as its confluence, strong normalization theorem, type inference algorithm, and principal typing theorem. In paper [6], we extended it to an ML-polymorphic type system and studied its type inference algorithm and principal typing theorem.

In this paper, we study evaluation strategies of the environment calculus, such as call-by-name and call-by-value evaluation strategies. In the paradigm of the lambda calculus, PCF(Programming language for Computable Function) has been used for study of evaluation strategies [7,9]. This is a simply typed calculus with natural numbers, Boolean values, and recursive operators. We here propose an environment calculus variant of PCF.

2   Environment PCF

Definition 1 (Types of PCFenv). Types of PCFenv are inductively defined by the following grammar:

$$A ::= \iota \mid o \mid (A \rightarrow B) \mid \{x_1 : A_1\} \cdots \{x_n : A_n\}.$$

The type ι represents a primitive type of natural number and o a primitive type of Boolean value. The type $(A \rightarrow B)$ is called a *function type,* whose domain is of type A and codomain of type B. The type $\{x_1 : A_1\} \cdots \{x_n : A_n\}$ is called an *environment type,* whose elements are the environment values mapping variables $x_1, \ldots, x_n$ to

values of type $A_1, \ldots, A_n$, respectively. We will use symbols such as E, E', H, H' for environment types.

Definition 2 (Terms of PCFenv). Terms of PCFenv are inductively defined by the following grammar:

$$M ::= 0 \mid tt \mid ff \mid succ(M) \mid pred(M) \mid zero?(M) \mid if\ L\ then\ M\ else\ N$$
$$\mid x \mid (MN) \mid \lambda x{:}A.M \mid \mu x{:}A.M \mid id \mid (M/x) \cdot N \mid (M \circ N)$$

The natural numbers are represented by the zero 0 and the *successor tionsucc*$(-)$. For example, a number 3 is encoded as $succ(succ(succ(0)))$. The primitive function $pred(-)$ is the predecessor function. The predicate $zero?\,(-)$ is the zero-testing predicate. If the value of a parameter is zero, then it returns the truth tt; otherwise, it returns the false ff. The term $(if\ L\ then\ M\ else\ N)$ is the conditional branch, that is, if the value of L is true, then it returns the value of M, and otherwise it returns the value of N. Variable x, lambda abstraction $\lambda x{:}A.M$, and recursive definition $(\mu x{:}A.M)$ are similar to the lambda calculus and PCF. The term id is called the identity environment, which corresponds to the identity substitution from the viewpoint of environment-as-substitution. Similarly, the term $(M \circ N)$ is considered as a composition of "substitutions" M and N. The term $(M/N) \cdot N$ is an extension of environment N by adding a binding of x to M. These three kinds of term are similar to those in environment calculus in the previous works [5,6]. We next give the type system to PCFenv.

Definition 3 (Typing Rules of PCFenv). Type judgment $E \vdash M{:}A$ is a ternary relation among environment type E, term M, and type A, which is inductively defined by the following *typing rules*.

$$\frac{}{\{x:A\}E \vdash x:A}\ \text{Var} \qquad \frac{\{x:A\}E \vdash M:B}{E \vdash \lambda x{:}A.M : A \to B}\ \text{Lam}$$

$$\frac{E \vdash M:A \to B \quad E \vdash M:A}{E \vdash (MN):B}\ \text{App}$$

$$\frac{}{E \vdash tt:o}\ \text{True} \qquad \frac{}{E \vdash ff:o}\ \text{False}$$

$$\frac{}{E \vdash 0:\iota}\ \text{Zero} \qquad \frac{E \vdash M:\iota}{E \vdash succ(M):\iota}\ \text{Succ} \qquad \frac{E \vdash M:\iota}{E \vdash pred(M):\iota}\ \text{Pred}$$

$$\frac{E \vdash M:\iota}{E \vdash zero?(M):\iota}\ \text{IsZero}$$

$$\frac{E \vdash L:o \quad E \vdash M:A \quad E \vdash N:A}{E \vdash if\ L\ then\ M\ else\ N:A}\ \text{If}$$

$$\frac{\{x:A\}E \vdash M:A}{E \vdash \mu x{:}A.M:A}\ \text{Mu}$$

$$\frac{}{E \vdash id:E}\ \text{Id}$$

$$\frac{E \vdash N:H \quad H \vdash M:A}{E \vdash (M \circ N):A}\ \text{Comp} \qquad \frac{E \vdash M:A \quad E \vdash N:H}{E \vdash (M/x) \cdot N : \{x:A\}H}\ \text{Extn}$$

Definition 4 (Equivalence Relation of PCFenv). The equivalence relation $M = N$ between terms M and N is inductively defined by the following rules.

$$pred(0) = 0, \quad \textbf{Pred1} \quad zero?(succ(\overline{n})) = f\!f, \quad \textbf{IsZero2}$$
$$pred(succ(\overline{n})) = \overline{n}, \; \textbf{Pred2} \quad (if \; tt \; then \; M \; else \; N) = M, \; \textbf{If1},$$
$$zero?(0) = tt, \quad \textbf{IsZero1} \quad (if \; f\!f \; then \; M \; else \; N) = N, \; \textbf{If2},$$

$$((\lambda x{:}A.M) \circ L)N = M \circ ((N/x) \cdot L) \; \textbf{Beta1},$$
$$(\lambda x{:}A.M)N = M \circ ((N/x) \cdot id) \quad \textbf{Beta2},$$
$$\mu x{:}A.M = M \circ ((\mu x{:}A.M/x) \cdot id), \quad \textbf{Mu},$$

$$id \circ M = M, \textbf{IdL} \quad M \circ id = M, \textbf{IdR}$$

$$((M/x) \cdot N) \circ L = (M \circ L/x) \cdot (N \circ L), \; \textbf{DExtn}$$
$$0 \circ M = 0, \; \textbf{Zero'} \quad succ(M) \circ N = succ(M \circ N), \quad \textbf{Succ'}$$
$$tt \circ M = tt, \; \textbf{True'} \quad pred(M) \circ N = pred(M \circ N), \quad \textbf{Pred'}$$
$$f\!f \circ M = f\!f, \; \textbf{False'} \quad zero?(M) \circ N = zero?(M \circ N) \; \textbf{Iszero'}$$
$$(if \; M \; then \; N \; else \; N') \circ L = (if \; M \circ L \; then \; N \circ L \; else \; N' \circ L) \; \textbf{If'}$$

$$(MN) \circ L = (M \circ L)(N \circ L) \; \textbf{DApp}$$

$$x \circ ((M/x) \cdot N) = M, \textbf{VarRef} \quad y \circ ((M/x) \cdot N) = y \circ N, \textbf{VarSkip}$$

$$(\mu x{:}A.M) \circ N = M \circ (((\mu x{:}A.M) \circ N/x) \cdot N), \; \textbf{Mu'}$$
$$(M \circ N) \circ L = M \circ (N \circ L), \quad \textbf{Assoc.}$$

(Equivalence)

$$\frac{}{M = N} \; \textbf{Refl} \qquad \frac{N = M}{M = N} \; \textbf{Sym} \qquad \frac{M = L \quad L = N}{M = N} \; \textbf{Trans}$$

(Congruence)

$$\frac{M = N}{succ(M) = succ(N)} \; \textbf{CongSucc} \qquad \frac{M = N}{pred(M) = pred(N)} \; \textbf{CongPred}$$

$$\frac{L = L' \quad M = M' \quad N = N'}{if \; L \; then \; M \; else \; N = if \; L' \; then \; M' \; else \; N'} \; \textbf{CongIf}$$

$$\frac{M = N}{zero?(M) = zero?(N)} \; \textbf{CongIsZero} \qquad \frac{M = M' \quad N = N'}{(MN) = (M'N')} \; \textbf{CongApp}$$

$$\frac{M = N}{\mu x{:}A.M = \mu x{:}A.N} \; \textbf{CongMu} \qquad \frac{M = N}{\lambda x{:}A.M = \lambda x{:}A.N} \; \textbf{CongLam}$$

$$\frac{M = M' \quad N = N'}{(M/x) \cdot N = (M'/x) \cdot N'} \; \textbf{CongExtn} \qquad \frac{M = M' \quad N = N'}{(M \circ N) = (M' \circ N')} \; \textbf{CongComp}$$

We next give call-by-name and call-by-value evaluation strategies to PCFenv.

Definition 5 (Call-by-name Evaluation Strategy of PCFenv). The *call-by-name evaluation strategy* of PCFenv, $M \downarrow^n N$, is defined as a binary relation between the terms M, N of PCFenv, by the following rules.

$$\frac{}{0 \downarrow^n 0} \text{ Zero} \qquad \frac{}{tt \downarrow^n tt} \text{ True} \qquad \frac{}{f\!f \downarrow^n f\!f} \text{ False} \qquad \frac{M \downarrow^n 0}{pred(M) \downarrow^n 0} \text{ Pred1}$$

$$\frac{M \downarrow^n succ(V)}{pred(M) \downarrow^n V} \text{ Pred2} \qquad \frac{M \downarrow^n V}{succ(M) \downarrow^n succ(V)} \text{ Succ} \qquad \frac{M \downarrow^n 0}{zero?(M) \downarrow^n tt} \text{ IsZero1}$$

$$\frac{M \downarrow^n succ(V)}{zero?(M) \downarrow^n f\!f} \text{ IsZero2} \qquad \frac{M \downarrow^n tt \quad N \downarrow^n V}{(if\ M\ then\ N\ else\ L) \downarrow^n V} \text{ If1}$$

$$\frac{M \downarrow^n f\!f \quad L \downarrow^n V}{(if\ M\ then\ N\ else\ L) \downarrow^n V} \text{ If2}$$

$$\frac{M \downarrow^n (\lambda x{:}A.M') \circ L \quad M' \circ ((N/x) \cdot L) \downarrow^n V}{(MN) \downarrow^n V} \text{ Beta1}$$

$$\frac{M \downarrow^n (\lambda x{:}A.M') \quad M' \circ ((N/x) \cdot id) \downarrow^n V}{(MN) \downarrow^n V} \text{ Beta2}$$

$$\frac{}{\lambda x{:}A.M \downarrow^n \lambda x{:}A.M} \text{ Lam} \qquad \frac{M \circ ((\mu x{:}A.M/x) \cdot id) \downarrow^n V}{\mu x{:}A.M \downarrow^n V} \text{ Mu}$$

$$\frac{}{id \downarrow^n id} \text{ Id} \qquad \frac{}{(M/x) \cdot N \downarrow^n (M/x) \cdot N} \text{ Extn} \qquad \frac{}{0 \circ M \downarrow^n 0} \text{ Zero'}$$

$$\frac{}{tt \circ M \downarrow^n tt} \text{ True'} \qquad \frac{}{f\!f \circ M \downarrow^n f\!f} \text{ False'}$$

$$\frac{M \downarrow^n (N/x) \cdot L \quad N \downarrow^n V}{x \circ M \downarrow^n V} \text{ VarRef}$$

$$\frac{M \downarrow^n (N/x) \cdot L \quad y \circ L \downarrow^n V \quad x \not\equiv y}{y \circ M \downarrow^n V} \text{ VarSkip}$$

$$\frac{succ(M \circ N) \downarrow^n V}{succ(M) \circ N \downarrow^n V} \text{ Succ'} \qquad \frac{pred(M \circ N) \downarrow^n V}{pred(M) \circ N \downarrow^n V} \text{ Pred'}$$

$$\frac{zero?(M \circ N) \downarrow^n V}{zero?(M) \circ N \downarrow^n V} \text{ IsZero'} \qquad \frac{if\ (M \circ L)\ then\ (N \circ L)\ else\ (N' \circ L) \downarrow^n V}{if\ M\ then\ N\ else\ N' \circ L \downarrow^n V} \text{ DIf}$$

$$\frac{}{(\lambda x{:}A.M) \circ N \downarrow^n (\lambda x{:}A.M) \circ N} \text{ Lam'}$$

$$\frac{M \circ ((\mu x{:}A.M/x) \cdot N) \downarrow^n V}{(\mu x{:}A.M) \circ N \downarrow^n V} \text{ Mu'}$$

$$\frac{(((M \circ L)/x) \cdot (N \circ L)) \downarrow^n V}{((M/x) \cdot N) \circ L \downarrow^n V} \text{ DExtn} \qquad \frac{M \circ (N \circ L) \downarrow^n V}{(M \circ N) \circ L \downarrow^n V} \text{ Assoc}$$

Definition 6 (Call-by-value Evaluation Strategy of PCFenv). The call-by-value evaluation strategy of PCFenv, $M \downarrow^v N$, is defined as a binary relation between the terms M, N of PCFenv, by the same rules except **Beta1** and **Beta2**. Instead of these two rules, we assume the following two rules.

$$\frac{M \downarrow^v (\lambda x{:}A.M') \circ L \quad N \downarrow^v W \quad M' \circ ((W/x) \cdot L) \downarrow^v V}{(MN) \downarrow^v V} \text{ \textbf{Beta1}}$$

$$\frac{M \downarrow^v (\lambda x{:}A.M') \quad N \downarrow^v W \quad M' \circ ((W/x) \cdot id) \downarrow^v V}{(MN) \downarrow^v V} \text{ \textbf{Beta2}}$$

We write PCFenv with a call-by-name evaluation as $\text{PCF}_{\text{env}}^{\text{cbn}}$ and the one with a call-by-value evaluation as $\text{PCF}_{\text{env}}^{\text{cbv}}$. If we do not need to distinguish between these two evaluations, we simply use PCFenv. We have the following property on values.

Proposition 1. Let V be a value. If $\vdash V{:}\iota$, then there is a natural number n satisfying that V is n. If $\vdash V{:}o$, then V is either tt or ff.

Proposition 2. If $M \downarrow^n V$, then $M = N$. If $M \downarrow^v N$, then $M = V$.

3   Translation Semantics of PCFenv

In this section, we present semantics of PCFenv, which is given as a translation of PCFenv to PCF. The translation consists of two mappings on types and on terms. We will give a definition of PCF in [2]. Before defining the semantics, we assume that the variable symbols are totally ordered. In this paper, if we represent a set of variables as an indexed set such as $\{x_1, x_2, x_3, \dots\}$, then it is implicitly assumed that $x_1 < x_2 < x_3 < \cdots$. We use a symbol $\equiv$ for a syntactical equivalence.

Definition 7 (Translation of Types). Let A be a type of PCFenv and $X \equiv \{x_1, \dots, x_N\}$ a set of the variables appearing in A. A *translation* of A, $\langle A \rangle_X$, is a type of PCF inductively defined by the following rules.

$$\langle \iota \rangle_X \equiv \iota, \quad \langle o \rangle_X \equiv o, \langle A \to B \rangle_X \equiv \langle A \rangle_X \to \langle B \rangle_X,$$

$$\langle E \rangle_X \equiv \sigma_1 \times (\sigma_2 \times (\cdots (\sigma_N \times \mathbf{1}))),$$

where $\sigma_i = \langle A_i \rangle_X$ (if $\{x_i{:}A_i\}$ occurs in E) $\sigma_i = \mathbf{1}$(otherwise).

Type $\mathbf{1}$ is a *unit type*, a type of PCF, inhabited by only one element. Its definition is given in [2]. An intuitive explanation of the translation is as follows. We assume that a set X of variables is large enough to cover the variables occurring in the type to be translated. Each environment type is translated as a product type. The unit type $\mathbf{1}$ is assigned to the unbound variables in the environment type to be translated. For example, assume that X is w, x, y, z and $w < x < y < z$. Then, $\langle \{x{:}\iota\}\{y{:}o\} \rangle_X$ is translated as $\mathbf{1} \times (\iota \times (\mathbf{1} \times (o \times \mathbf{1})))$. In order to represent environment values in PCFenv as product values in PCF, we prepare several abbreviations ("macros" in programming terminology) in PCF.

Definition 8. We assume that X is a finite set $\{x_1, \dots, x_n\}$ of variables. The *empty environment* on X, written as $\lceil x_i \rceil_X$ means a non-negative integer $(i - 1)$. An abbreviation $\text{lookup}_X(r, i)$ and $\text{update}_X(r, i, M)$ encodes replacement of the i-th component in a sequence M.

Proposition 3. It holds that

$$\mathrm{lookup}_X(\mathrm{update}_X(r,i,M),i) = M,$$
$$\mathrm{lookup}_X(\mathrm{update}_X(r,i,M),j) = \mathrm{lookup}_X(r,j) \quad (i \not\equiv j).$$

We next show a translation of terms of PCFenv into those of PCF.

Definition 9 (Translation of Terms). Let M be a term of PCFenv, X a finite set of variables, and r a term of PCF. *Translation of M* is a term of PCF defined as follows.

$$\langle\!\langle 0 \rangle\!\rangle_X(r) \equiv 0, \quad \langle\!\langle tt \rangle\!\rangle_X(r) \equiv tt, \quad \langle\!\langle f\!f \rangle\!\rangle_X(r) \equiv f\!f,$$

$$\langle\!\langle succ(M) \rangle\!\rangle_X(r) \equiv succ(\langle\!\langle M \rangle\!\rangle_X(r)), \quad \langle\!\langle pred(M) \rangle\!\rangle_X(r) \equiv pred(\langle\!\langle M \rangle\!\rangle_X(r)),$$

$$\langle\!\langle zero?(M) \rangle\!\rangle_X(r) \equiv zero?(\langle\!\langle M \rangle\!\rangle_X(r)),$$

$$\langle\!\langle if\ M\ then\ N\ else\ L \rangle\!\rangle_X(r) \equiv if\ \langle\!\langle M \rangle\!\rangle_X(r)\ then\ \langle\!\langle N \rangle\!\rangle_X(r)\ else\ \langle\!\langle L \rangle\!\rangle_X(r),$$

$$\langle\!\langle x \rangle\!\rangle_X(r) \equiv \mathrm{lookup}_X(r, \lceil x \rceil_X), \quad \langle\!\langle (MN) \rangle\!\rangle_X(r) \equiv (\langle\!\langle M \rangle\!\rangle_X(r)\ \langle\!\langle N \rangle\!\rangle_X(r)),$$

$$\langle\!\langle \lambda x{:}A.M \rangle\!\rangle_X(r) \equiv \lambda x'{:}\langle A \rangle_X.\langle\!\langle M \rangle\!\rangle_X(\mathrm{update}_X(r, \lceil x \rceil_X, x')),$$

$$\langle\!\langle \mu x{:}A.M \rangle\!\rangle_X(r) \equiv \mu x'{:}\langle A \rangle_X.\langle\!\langle M \rangle\!\rangle_X(\mathrm{update}_X(r, \lceil x \rceil_X, x')),$$

$$\langle\!\langle (M/x) \cdot N \rangle\!\rangle_X(r) \equiv \mathrm{update}_X(\langle\!\langle N \rangle\!\rangle_X(r), \lceil x \rceil_X, \langle\!\langle M \rangle\!\rangle_X(r))$$

$$\langle\!\langle M \circ N \rangle\!\rangle_X(r) \equiv \langle\!\langle M \rangle\!\rangle_X(\langle\!\langle N \rangle\!\rangle_X(r)), \quad \langle\!\langle id \rangle\!\rangle_X(r) \equiv r$$

We write $\langle\!\langle M \rangle\!\rangle_X(\emptyset_X)$ as $\langle\!\langle M \rangle\!\rangle_X$ for sake of simplicity.

Theorem 1 (Typing Preserving of Translation). For a term M of PCFenv, a term r of PCF, and X a finite set of variables, if $E \vdash M{:}A$, $\Gamma \vdash r{:}\langle E \rangle_X$, and X includes all the variables occurring in E, M, A, Γ, and r, then it holds that $\Gamma \vdash \langle\!\langle M \rangle\!\rangle_X(r){:}\langle A \rangle_X$.

This is proved by structural induction on term M. The values of PCFenv are translated to values of PCF.

Proposition 4. Let V be a value of PCFenv. If $\{\} \vdash V{:}A$, then $\langle\!\langle V \rangle\!\rangle_X$ is a value of PCF.

Proposition 5. Suppose that values V and W of PCFenv, and a base type A. If $\langle\!\langle V \rangle\!\rangle_X = \langle\!\langle W \rangle\!\rangle_X$, then $V \equiv W$.

4   Semantical Study of Environment PCF

In the previous section, we give a translation semantics to PCFenv. The soundness of the equivalence relation, the call-by-name and call-by-value evaluation strategies is formulated as follows.

Theorem 2 (Soundness of Equivalence Relation). Let M and N be terms of PCFenv satisftying that $E \vdash M : A$ and $E \vdash N : A$, and X a finite set of variables occurring in terms M and N. If $M = N$, then $\langle\!\langle M \rangle\!\rangle_X(r) = \langle\!\langle N \rangle\!\rangle_X(r)$ for any PCF-term r of type $\langle\!\langle E \rangle\!\rangle_X$.

Proof Sketch. We prove the theorem by structural induction on the relation $M = N$.

Case of Beta1. We assume that (1) $M \equiv \left((\lambda x : B . M_1) \circ M_2\right) M_3$, (2) $N \equiv M_1 \circ ((M_3/x) \cdot M_2)$, (3) $E \vdash M : A$, $E \vdash N : A$. Let r be a PCF-term of type $\langle E \rangle_X$. (1) and (2) imply

$$\langle\!\langle M \rangle\!\rangle_X(r) \equiv (\lambda x' : \langle B \rangle_X . \langle\!\langle M_1 \rangle\!\rangle_X(\text{update}_X(\langle\!\langle M_2 \rangle\!\rangle_X(r), \lceil x \rceil, x')))$$

$$(\langle\!\langle M_3 \rangle\!\rangle_X(r))$$

and

$$\langle\!\langle N \rangle\!\rangle_X(r) \equiv \langle\!\langle M_1 \rangle\!\rangle_X(\text{update}_X(\langle\!\langle M_2 \rangle\!\rangle_X)(r), \lceil x \rceil_X, x')[x' := \langle\!\langle M_3 \rangle\!\rangle_X(r)],$$

respectively. By the rules of the equivalence relation of PCF, we have $\langle\!\langle M \rangle\!\rangle_X(r) = \langle\!\langle N \rangle\!\rangle_X(r)$.

Case of Beta2. This case is similar to the previous one.

Cases of Mu and Mu'. These cases are also similar to Beta1 and Beta2.

Case of VarRef. Suppose that (4) $M \equiv x \circ ((M_1/x) \cdot M_2)$, (5) $N \equiv M_1$, (6) $E \vdash M : A$, $E \vdash N : A$. (4) implies that

$$\langle\!\langle M \rangle\!\rangle_X(r) \equiv (\text{lookup}_X(\text{update}_X(\langle\!\langle M_2 \rangle\!\rangle_X(r), \lceil x \rceil, \langle\!\langle M_1 \rangle\!\rangle_X(r)), \lceil x \rceil)$$

By Proposition 3,

$$\langle\!\langle M \rangle\!\rangle_X(r) = \langle\!\langle M_1 \rangle\!\rangle_X(r))$$

(5) implies

$$\langle\!\langle M \rangle\!\rangle_X(r) = \langle\!\langle N \rangle\!\rangle_X(r).$$

The case of VarSkip is similarly proved using the latter half of Proposition 3. The other cases are proved straightforwardly.

End of Proof Sketch.

Corollary 1 (Soundness of Call-by-Name and Call-by-Value). Suppose that either $M \downarrow^n V$ or $M \downarrow^v V$, $E \vdash M : A$, $E \vdash V : A$, and X be a finite set of the variables occurring in M and V. Then, for any PCF-term r of type $\langle E \rangle_X$, it holds that $\langle\!\langle M \rangle\!\rangle_X(r) \equiv \langle\!\langle V \rangle\!\rangle_X(r)$.

We next show the converse of Theorem 1 restricted on terms of base types, which is called *computational adequacy*[2]. Due to lack of space, we only present it for the call-by-name evaluation strategy.

We prepare the following lemma to prove it.

Lemma 1. Suppose that $\langle\!\langle x \circ (M_1/y) \cdot M_2 \rangle\!\rangle_X(r) \downarrow^n V$. If $x \equiv y$, then $\langle\!\langle M_1 \rangle\!\rangle_X(r) \downarrow^n V$. Otherwise, $\langle\!\langle x \circ M_2 \rangle\!\rangle_X(r) \downarrow^n V$.

Lemma 2. Let M be a term of PCFenv, A a type of PCFenv, V a value of and V a finite set of variables satisfying that $\{\ \} \vdash M : A$ in PCFenv, $\langle\!\langle \{\ \} \rangle\!\rangle_X \vdash V : \langle A \rangle_X$ in PCF, $\langle M \rangle_X \downarrow^n V$, in PCF, X includes all the varialbes occurring in both M and A. Then, there exists a value W of PCFenv satisfying that $\{\ \} \vdash W : A$ in PCFenv, $M \downarrow^n W$ in PCFenv, $\langle W \rangle_X \downarrow^n V$ in PCF.

Theorem 3 (Computational Adequacy). Let M be a term, V a value, and A a base type of PCFenv satisfying that $\{\ \} \vdash M : A$, $\{\ \} \vdash V : A$, and X a finite set of the variables occurring in both M and V. if $\langle M \rangle_X = \langle V \rangle_X$, then it holds that $M \downarrow^n V$.

Proof. By Proposition 1, $\{\ \} \vdash \langle M \rangle_X : \langle A \rangle_X$, $\{\ \} \vdash \langle V \rangle_X : \langle A \rangle_X$. Since A is a base type of PCFenv, $\langle A \rangle_X$ is also a base type of PCF. Hence, by Theorem 4 (Computational Adequacy of PCF), $\langle M \rangle_X \downarrow^n \langle V \rangle_X$. By Lemma 2, there exists a value W satisfying that $M \downarrow^n W$ and $\langle W \rangle_X \downarrow^n \langle V \rangle_X$. By Proposition 4, $\langle V \rangle_X$ and $\langle W \rangle_X$ are values, and therefore, $\langle W \rangle_X \equiv \langle V \rangle_X$. By Proposition 5, $W \equiv V$. Hence, $M \downarrow^n V$.

Theorem 4. Let M be a term, V a value of PCFenv satisfying that $\{\ \} \vdash M : A$, $\{\ \} \vdash V : A$ and a finite set of variables occurring in M and V. Then, $M = V \Leftrightarrow M \downarrow^n V$.

End of Proof.

5   Conclusion

In this paper, we proposed PCF with first-class environments, PCFenv, and defined a translation of PCFenv on PCF as its semantics and studied theoretical properties such as the soundness of the equivalence relation, and evaluation strategies and computational adequacy for the call-by-name evaluation strategy.

Besides our system of environment calculus [5,6], there are other systems such as Sato-Sakurai-Burstall's $\lambda\epsilon$-calculus[8]. It is also interesting to apply our method to their calculus.

6   Programming Language PCF

We present definitions of several notions related to a computational formal system PCF. It is based on the system presented by Gunter [2]; and we add product types and pairs to it. Recursion is introduced by the recursive binder μ. The natural numbers are represented by a constant 0 and a constructor $\mathrm{Succ}(-)$.

Definition 9 (Types and Terms of PCF)
Types of PCF are defined inductively by the following grammar.

$$A ::= \iota \mid o \mid 1 \mid (A \rightarrow B) \mid (A \times B)$$

These are called *numeral type*, *boolean type*, *unit type*, respectively. Terms of PCF are defined inductively by the following grammar.

$$M ::= 0 \mid tt \mid ff \mid 1 \mid succ(M) \mid pred(M) \mid zero?(M)$$
$$\mid if\ L\ then\ M\ else\ N \mid x \mid (MN) \mid \lambda x{:}\sigma.M \mid \mu x{:}\sigma.M$$
$$\mid (M,N) \mid \pi^1(M) \mid \pi^2(M)$$

These terms are called zero, truth, false, unit, successor function, predecessor function, zero-testing, conditional branch, variable, function application, lambda-abstraction, recursive binding, pairing, first projection, and second projection, respectively.

For the sake of convenience, we use the same symbol **1** for the unit term and the unit type. If any confusion arises, we distinguish between them by writing 1_{type} and 1_{expr}.

In PCF, we use the same names of rules as those of PCFenv, for the sake of simplicity. If any confusion arises, we use e, e_1, e_2 for terms of PCF, σ, τ for types of PCF, Γ, Δ for type assignments of PCF (see below).

Definition 10 (Typing of PCF). A type assignment is a partial mapping from variable names to types, written as

$$\{x_1 : A_1\} \cdots \{x_n : A_n\}.$$

We use meta-variables Γ, Δ for type assignments of PCF. Typing of PCF is defined by a ternary relation (called a *typing judgment*) among type assignment E, term e, and type σ inductively defined by typing rules. Many typing rules are similar to the typing rules of PCFenv, the left-hand side of $\vdash$ is a type assignment in place of an environment type. For example,

$$\frac{}{\{x : \sigma\}\Gamma \vdash x : \sigma} \; \text{Var} \qquad \frac{\{x : \sigma\}\Gamma \vdash e : \tau}{\Gamma \vdash \lambda x{:}\sigma.e : \sigma \to \tau} \; \text{Lam}$$

$$\frac{\Gamma \vdash e_1 : A \to B \quad \Gamma \vdash e_2 : A}{\Gamma \vdash (e_1 e_2) : B} \; \text{App}$$

The other rules of PCFenv (**True, False, Zero, Succ, Pred, IsZero, If,** and **Mu**) are regarded as typing rules of PCF, replacing E, H to Γ, Δ, M, N to e_1, e_2, and A, B to σ, τ. The typing rules related to the product and unit types are different to PCFenv.

$$\frac{\Gamma \vdash e_1 : \sigma \quad \Gamma \vdash e_2 : \tau}{\Gamma \vdash (e_1, e_2) : \sigma \times \tau} \; Pair$$

$$\frac{\Gamma \vdash e : \sigma \times \tau}{\Gamma \vdash \pi^1(e) : \sigma} \; Proj1 \qquad \frac{E \vdash M : \sigma \times \tau}{\Gamma \vdash \pi^2(e) : \tau} \; Proj2 \qquad \frac{}{\Gamma \vdash 1_{expr} : 1_{type}} \; Unit$$

Definition 11 (Equivalence Relation of PCF). Equivalence relation of PCF is defined by the rule many of which are similar to those of PCFenv. The rules related to equivalence (**Refl, Sym, Trans**), to (**CongSucc, CongPred, CongIsZero, CongIf, CongMu, CongLam**), to primitive operations (Pred1, Pred2, IsZero1, IsZero2, If1, If2) are the same in PCF. The following rules related to congruence are different from PCFenv.

$$1 \downarrow^n 1 \; \; \text{Unit}$$

$$(e_1, e_2) \downarrow^n (e_1, e_2) \; \; \text{Pair}$$

$$\frac{e_1 \downarrow^n \lambda x{:}\sigma.e_3 \quad e_3[x{:=}e_2] \downarrow^n v}{(e_1 e_2) \downarrow^n v} \; \text{Beta}$$

$$\frac{e[x{:=}\mu x{:}\sigma.e] \downarrow^n v}{\mu x{:}\sigma.e \downarrow^n v} \; \text{Mu}$$

$$\frac{e \downarrow^n (e_1, e_2) \quad e_1 \downarrow^n v}{\pi^1(e) \downarrow^n v} \ \mathbf{Fst}$$

$$\frac{e \downarrow^n (e_1, e_2) \quad e_2 \downarrow^n v}{\pi^2(e) \downarrow^n v} \ \mathbf{Snd}$$

The following theorem appears in the other literature, e.g. [2,4].

Theorem 4 (Computational Adequacy of PCF). For a closed term M and a value V of a base type,

$M = V$ if and only if $M \downarrow^n V$.

References

1. Abadi, M., Cardelli, L., Curien, P.L., Lévy, J.J.: Explicit substitutions. Journal of Functional Programming 1(4), 375–416 (1991)
2. Gunter, C.A.: Semantics of programming languages: structures and techniques. The MIT Press (1992)
3. Hanson, C.: MIT/GNU Scheme Reference Manual (Release 9.1), 1.105 edn. (October 2011)
4. Mitchell, J.C.: Foundations for programming languages. The MIT Press (1996)
5. Nishizaki, S.: Simply typed lambda calculus with rst-class environments. Publication of Research Institute for Mathematical Sciences Kyoto University 30(6), 1055–1121 (1995)
6. Nishizaki, S.: Polymorphic environment calculus and its type inference algorithm. Higher-Order and Symbolic Computation 13(3) (2000)
7. Plotkin, G.: Call-by-name, call-by-value, and the -calculus. Theor. Comput. Sci. 1, 125–159 (1975)
8. Sato, M., Sakurai, T., Burstall, R.: Explicit environments. In: Girard, J.-Y. (ed.) TLCA 1999. LNCS, vol. 1581, pp. 340–354. Springer, Heidelberg (1999)
9. Scott, D.: A type-theoretical alternative to ISWIM, CUCH, OWHY. Theor. Comput. Sci. 121, 411–440 (1993)
10. Sperber, M., Dybvig, R.K., Flatt, M., van Straaten, A. (eds.): Revised [6] Reporton the Algorithmic Language Scheme. Cambridge University Press (2010)

Environmental Map Building and Location
for Mobile Robots in Intelligent Space

Yi Shen[1,2], Ping Liu[1,3], Mingxin Yuan[1,2], Shuai Chen[1], and Yafeng Jiang[1]

[1] School of Mechanical-Electronic and Automotive Engineering,
Jiangsu University of Science and Technology, Zhangjiagang 215600, China
[2] Suzhou Institutes of Technology, Jiangsu University of Science and Technology,
Zhangjiagang 215600, China
[3] Institute of Bio-Mechatronic Engineering, Jiangsu University, Zhenjiang 212013, China
Shenyi76@163.com,
{1019472877,429622974,396554112,1058220161}@qq.com

Abstract. In order to achieve the environmental information for the mobile robots during the autonomous navigation, an intelligent space is built firstly based on the Zigbee wireless sensor network system (WSNS) and the distributed vision system. After the environmental information for the mobile robots is achieved through the vision system in the built intelligent space, on the one hand, the static obstacles are segmented by using the Otsu threshold method and the mathematical morphology, and the environmental map of the mobile robot is built; on the other hand, the position and the direction of the mobile robot is achieved through the background difference, and the vision-based location for the mobile robot is realized. The experimental results of the environmental map building and the mobile robot location verify the validity of the proposed intelligent space and the image processing method for the map building and location.

Keywords: Environmental map building. Vision location, Mobile robot, intelligent space.

1 Introduction

With the development of the science and technology and the social economy, the application fields of the mobile robots are undergoing tremendous change. To improve the execution efficiency of the mobile robot system, the complete navigation information needs to be provided for the mobile robots. One of the navigation information is the environmental map consisted of the viable region and non-viable region; the other is the location information of the mobile robot, which is mainly related to the coordinates and the direction of the mobile robot. The environmental map building is mainly based on the detection of the environment information surrounding the mobile robot through the sensor technologies and the image segmentation, which is the premise of finishing safe autonomous navigation for the mobile robots. Wang *et al*. finished the map building in unknown environments by

combining particle filter and dot-line congruence [1]. The experimental results with real data demonstrated that the proposed approach was effective and robust for the indoor environment mapping. Li *et al*. finished the map building of mobile robots based on hybrid DSm (Dezert- Smarandache model) according to the detected sonar information in the dynamic environments [2]. All above methods can solve the map building; however, they are mainly for the dynamic environments, and the building is achieved on the basis of the local information. How to achieve accurate environmental map in the global environment needs further study.

The mobile robot location mainly solves the problem where the robot is, and is the basic link in the mobile robot navigation. Yu *et al*. presented a location method for service robots based on RFID [3]. Li *et al*. proposed an improved particle filter algorithm for the robot location [4], and the simulation results showed that the proposed algorithm was characterized by the high location precision. Like the environmental map building, the larger location error will directly affect the safe navigation of mobile robots. To solve the map building and location during the mobile robot navigation, the more accurate map building and location are realized in this paper based on the image processing technology in a built intelligent space.

This paper is arranged as follows. Section 2 provides the building of the intelligent space. Section 3 discusses the environmental map building based on the Otsu threshold method and the mathematical morphology. Section 4 presents the mobile robot location based on the background difference. Finally, the conclusions are stated in Section 5.

2   Building of the Intelligent Space

In order to help the mobile robot achieve more complete environmental information and its pose information, building the intelligent space is the main research direction in recent years. Tomomi *et al*. built a smart space consisted of PC, RFID tags and readers, and wireless sensor network [5]. In the smart space, the cooperation among different robots was realized through the u-objects finding (UOF). Baeg *et al*. also built a smart home environment for service robots based on the RFID and the sensor networks [6]. In order to provide the complete environmental information and the pose information of the mobile robot during the autonomous navigation, a new schematic intelligent space as shown in Fig.1 is built based on the distributed vision system and the Zigbee-based wireless sensor network (WSN) system [7]. The distributed vision system provides the whole visual environmental information for the mobile robot. The WSN system involves a host computer, a location dongle connected to the host computer through serial port (RS232), some reference nodes, a blind node installed on the mobile robot. The mobile robot is a two wheeled differential drive robot using the stepper motor. Through the built schematic intelligent space, on the one hand the robot communicates with the WSN system and receives the controlling instructions from the host computer through the blind node; on the other hand the robot realizes the self-location through the WSN system / DR, which can make up for the visual blind spot during the vision location. Based on the intelligent space, the environment map and vision location for mobile robot are finally realized.
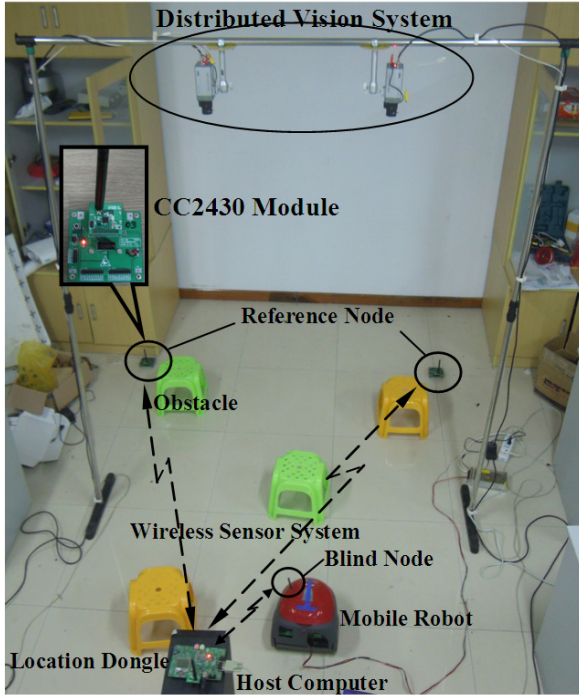
Fig. 1. Intelligent space

3 Environmental Map Building Based on the Otsu Threshold Method and the Mathematical Morphology

In order to efficiently and safely complete the autonomous navigation for the mobile robot, the environment information surrounding the mobile robot should be provided firstly; then the pose information of the mobile robot should be provided. The former mainly refers to the accurate environmental map including the viable region and non-viable region. The latter mainly refers to the mobile robot location. In this paper, the environment information and the location of the mobile robot are achieved based on the vision system in the intelligent space.

3.1 Environmental Map Building Algorithm

The segmentation of viable region and non-viable region in the robot moving space is realized based on the image segmentation technologies [8,9]. Before the segmentation, the image needs preprocessing. The concrete algorithm steps for the environmental map building are as follows:

Step 1. Convert the indoor images achieved from the vision system of the intelligent space into the grayscale image based on the color model.

The color images from the vision system are in the RGB space, namely the image is built up of three colors: Red, green and blue. In order to facilitate the image segmentation during the environmental map building of the mobile robot, the images are converted from the RGB space into the YUV space. The conversion between the RGB and the YUV can be described as follows:

$$\begin{bmatrix} Y \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.169 & -0.3316 & 0.500 \\ 0.500 & -0.4186 & -0.0813 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{1}$$

The conversion equation from the color pixel to the gray pixel can be described as follows:

$$Gray = 0.299R + 0.587G + 0.114B \tag{2}$$

Step 2. Calculate the optimal threshold value of the gray images according to the Otsu threshold value, and then the gray-scale images are subjected to binarization processing.

Otsu threshold value method is named after Nobuyuki Otsu. The best threshold of Otsu's method [8,9] is described as follows:

$$T = Arg \underset{0 \leq T \leq m-1}{Max} \left\{ \omega_0(T) \big[\mu_0(T) - \mu \big]^2 + \omega_1(T) \big[\mu_1(T) - \mu \big]^2 \right\} \tag{3}$$

Where, T is the threshold value. $\omega_0(T)$ and $\omega_1(T)$ are the proportion of the target portion and background portion, respectively. $\mu_0(T)$ and $\mu_1(T)$ are the mean gray value of the target portion and background portion, respectively. μ is the mean gray value of the whole image.

Step 3. Select a 3×3 structural element to execute an "Open" computation to the binarization images based on the mathematical morphology, and then select a 5×5 structural element to execute a "Close" computation.

Let $f(x)$ be the input image, and $g(x)$ be the structural element, the "Open" and the "Close" computation for $f(x)$ to $g(x)$ can be described respectively as follows:

$$(f(x) \circ g(x)) = [(f(x) \Theta g(x)) \oplus g(x)] \tag{4}$$

$$(f(x) \bullet g(x)) = [(f(x) \oplus g(x)) \Theta g(x)] \tag{5}$$

The "Open" computation can remove small isolated points, burr and small bridges, and the whole position shape won't change. The "Close" computation can fill the tiny holes in the target, connect the broken neighboring target, and smooth the boundaries, and it won't change the area.

Step 4. Finish the segmentation of the indoor viable region and non-viable region, and realize the achievement of the environmental map information for the mobile robots in the intelligent space.

3.2 Building Experiment of the Environmental Map

To verify the validity of the proposed map building algorithm in this paper, an experiment is performed as shown in Fig.2. Fig.2(a) is from the vision system of the built intelligent space. In the visual range of the camera, there are four obstacles, and one mobile robot. According to the algorithm flow based on the map building algorithm, the segmentation result can be achieved as shown in Fig.2(b). From the figure, it can be seen that the obstacles and mobile robot can basically be segmented, and the environmental map for mobile robot can be achieved, which verifies the validity of Otsu threshold method.

(a) Original video image   (b) Segment results
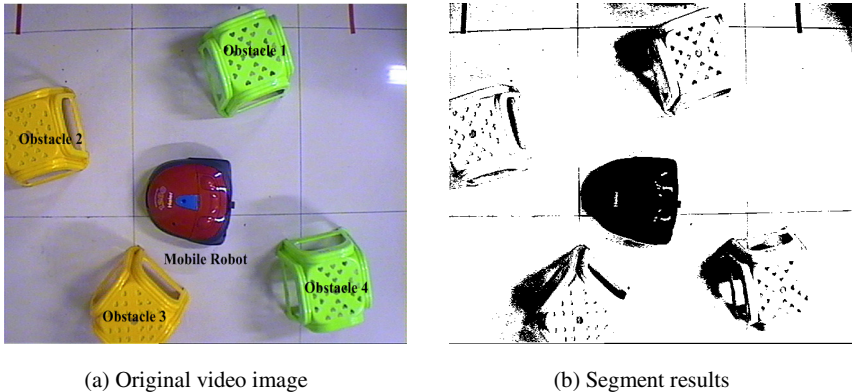
Fig. 2. Segmented image by using Otsu threshold method

4 Mobile Robot Location Based on the Background Difference

The achievement of the environmental map is the precondition in the mobile robot autonomous navigation; however, the robot can not be lost during the autonomous navigation. It is very necessary and important for the mobile robot to finish the location. The location of the mobile robots includes the determination of the position coordinates and the direction angle.

Video-based moving target detection is the basic research in the field of image processing and computer vision problems, and it is the research focus. The background difference method is very suitable for the detection of the moving target in the complex background, and it is generally able to provide the most complete characterization data. Considering that the background difference method is characterized by small calculation, and it is suitable for the occasions where need high real-time performance, the mobile robot location is finished based on the background difference method in this study.

As mentioned above, the mobile robot location includes the determination of position coordinate and direction angle.

4.1   Determination Algorithm for Position Coordinate

The algorithm includes the following processes:

Step 1. Get the color image containing the mobile robot by using camera of intelligent space.

Step 2. Achieve the moving target (namely mobile robot) based on the background difference method.

Step 3. Get the binary image of the moving target based on the Euclidean distance of the color pixel vector.

Let F be the RGB value of current image, and B be the RGB value of background image. The Euclidean distance S of two pixels can be defined as follows:

$$S = \|F - B\| = \sqrt{(F\_r - B\_r)^2 + (F\_g - B\_g)^2 + (F\_b - B\_b)^2} \qquad (6)$$

Where, F_r, F_g, F_b, B_r, B_g and B_b are the RGB component value of current image and background image.

If $S \geq$ optimal threshold value T, $S=1$; otherwise $S=0$. The optimal threshold value T can be got through experiments, and T is 65 in this study. The achieved binary image through above Euclidean distance is often affected by noise. To get more accurate moving target, we must make a deal with the binary image.

Step 4. Have a denoising treatment on the binary image using the "Open" computation in mathematical morphology.

Step 5. Scan the binary image line by line, and synthesize the pixel of mobile robot to be a connected region.

Step 6. Achieve the position coordinate of the mobile robot according to the connected region.

4.2   Determination Algorithm for Direction Angle

The algorithm includes the following processes:

Step 1. Paste a T-type blue ribbon on the center position of a mobile robot.

Step 2. Get the color image containing the mobile robot, and then get the moving target based on the background difference method too.

Step 3. Transform the color different robot image from RGB space to HIS space.

The transformation equation between RGB and HIS can be described as follows:

$$\begin{cases} I = \dfrac{R+G+B}{3} \\[2mm] S = 1 - \dfrac{3}{R+G+B} \min(R,G,B) \\[2mm] H = ar\cos(\dfrac{(R-G)+(R-B)}{2\sqrt{(R-G)^2 + (R-B)(R-G)}}) \end{cases} \qquad (7)$$

Step 4. Extract the T-type blue ribbon according to the setting values of H and S.

Step 5. Smooth the image of T-type blue ribbon by using the "Open" and "Close" computation.

Step 6. Achieve the final direction angle of the mobile robot through the linear fitting of the T-type ribbon and other assistant measurements.

4.3   Vision Location Experiments of the Mobile Robot

To verify the validity of the proposed algorithm for the mobile robot location in this paper, two experiments are finished.

Extraction Experiment of the T-type Ribbon. Fig. 3(a) is the image containing the mobile robot pasted T-type ribbon, which is got by using the camera in the intelligent space. Fig. 3(b) is the segmented image of the T-type ribbon based on the background difference method and the HIS color model. The final direction angle of the mobile robot is 192.215719°.
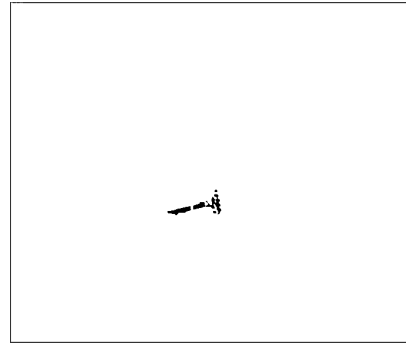
(a) mobile robot pasted T-type ribbon   (b) Segment results

Fig. 3. Vision location experiment

Real-Time Location Experiment of the Mobile Robot. To further verify the validity of the whole vision location algorithm, a real-time location experiment is carried out. Fig. 4(a)-(f) are the original sequence images of the moving robot in the built intelligent space. Fig. 5 (a)-(f) are the vision location results. From the location results, it can be seen that the moving target can be accurately segmented, and the final position coordinates of Fig.5 (a)-(f) are: (116,472), (186,482), (228,441), (265,377), (337,251) and (388,161), respectively.
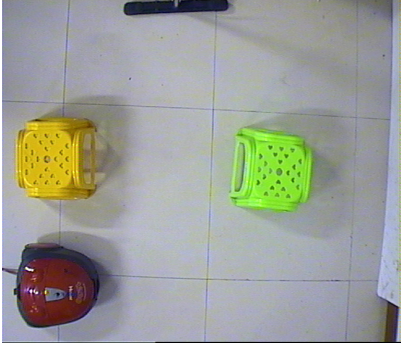
(a)
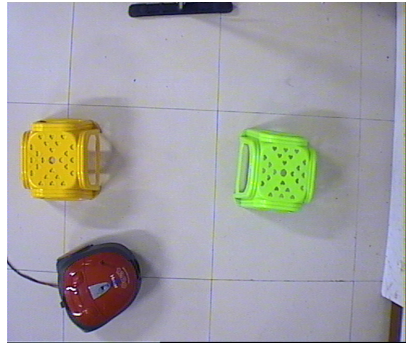
(b)
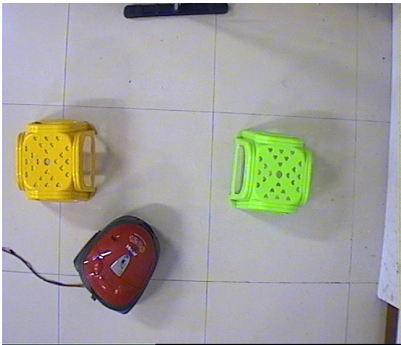
(c)
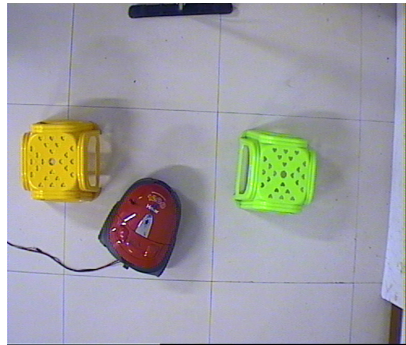
(d)

(e)
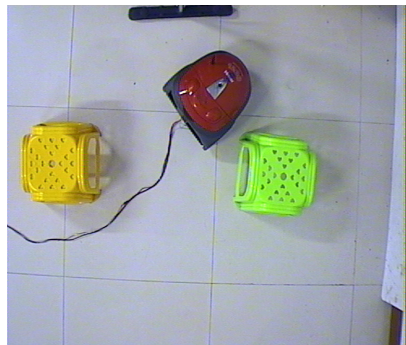
(f)

Fig. 4. Sequence images for vision location

(a)

(b)

(c)

(d)

(e)

(f)
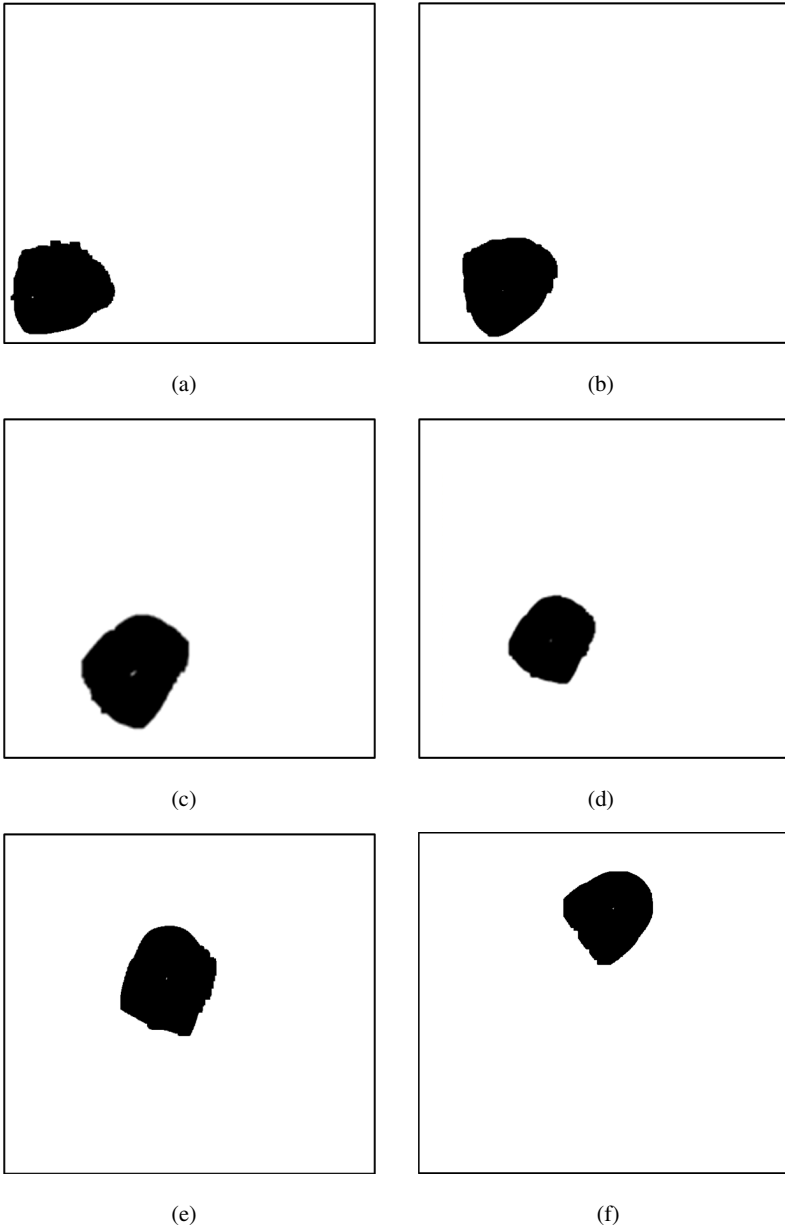
Fig. 5. Location results of mobile robot

5   Conclusions

To achieve the environmental map and solve the robot location during the mobile robot autonomous navigation, a new intelligent space is designed based on the

distributed vision system and wireless sensor network system. A segmentation algorithm based on the Otsu threshold method and the mathematical morphology is presented to finish the image segmentation for the environmental information of the mobile robot, and then the environmental map is built for the robot. By pasting the T-type blue ribbon on the center position of the mobile robot, the robot location is realized based on the background difference method and mathematical morphology.

References

1. Wang, W.F., Xiong, R., Chu, J.: A simultaneous localization and mapping approach by combining particle filter and dot-line congruence. Acta Automatica Sinica 35(9), 1185–1192 (2009)
2. Li, P., Huang, X.H., Wang, M.: Hybrid-dsm-model-based Mobile Robot Map Building in Dynamic Environment. ROBOTA 31(1), 40–52 (2009)
3. Yu, Q.X., Wu, K.K., Yuan, C., Fu, Z.: Application of RFID Technology in Restaurant Service Robot Localization. Journal of Donghua University (Natural Science) 38(4), 460–464 (2012)
4. Li, Y., Zhang, H., Qu, Y., Zhou, H.P.: Application of Improved Particle Filter in WSN-Aided Robot Localization. Computer Simulation 27(5), 166–169 (2010)
5. Tomomi, K.S., Ma, J.H., Bernady, A.: A System Prototype with Multiple Robots for Finding U-Objects in A Smart Space. In: International Conference on Embedded Software and Systems, pp. 229–236. IEEE Press, Chengdu (2008)
6. Baeg, S.H., Park, J.H., Koh, J.: Building A Smart Home Environment for Service Robots Based on RFID and Sensor Networks. In: IEEE International Conference on Control, Automation and Systems, pp. 1078–1082. IEEE Press, Seoul (2007)
7. Sun, X.X.: Fast Image Stitching and Application Software Development for Intelligent Space. Bachelor dissertation, School of Mechatronics and Automotive Engineering, Jiangsu University of Science and Technology, Zhangjiagang, China (2011)
8. Liu, P.: Image Mosaic and Obtaining Environment Map in the Intelligent Space. Bachelor dissertation, School of Mechatronics and Automotive Engineering, Jiangsu University of Science and Technology, Zhangjiagang, China (2011)
9. Zou, X.Y.: Intelligent Navigation Research of Autonomous Mobile Robot. Ph.D. dissertation, College of Electrical Engineering, Zhejiang University, Hangzhou, China (2004)

Topic-Centric Candidate Priors
for Expert Finding Models

Jian Liu[1,2,*], Baojuan Li[2], Baohong Liu[1], and Qi Li[1]

[1] College of Mechatronics and Automation, National University of Defense Technology,
Changsha, Hunan, 410073
[2] Fourth Military Medical University, Xi'an, Shaanxi, 710032, China
{supakito,libjuan}@163.com, lbh_nudt@126.com, liqi_nudt@sina.com

Abstract. To improve the performance of the expert finding system, a topic-centric candidate priors (TCCP) model was proposed. This model treated the documents as mixtures of topics. The topics of the candidates and terms were then used as latent variables to calculate the candidate priors. Latent Dirichlet Allocation (LDA) was utilized to find the topics. Finally, the parameters of the model were estimated using Gibbs sampling. The performance of the proposed model was evaluated using the CSIRO Enterprise Research Collection. The results showed that the performance of the expert finding system was dramatically improved by using the candidate priors estimated by the TCCP model. Compared to classical models, the proposed candidate priors model exploited the information in the whole data corpus and thus furnished more reasonable candidate priors.

Keywords: Expert finding, candidate priors, topic-centric model.

1   Introduction

With the rapid development of information retrieval (IR) technologies, the scope of IR has gone far beyond document retrieval. In many cases, the information needs are not confined to documents but extend to entities including people, places, organizations, etc. Entity retrieval which aims at finding entities given a query is now a new focus of IR. As one of the most important branches in entity retrieval, expert finding has gained increasing attention since the advent of the expert search task in the Text Retrieval Conference (TREC) 2005. Expert finding is also referred to as expert recommendation[1], expert search[2, 3], expert identification[4],etc. Given a topic, expert finding systems try to find persons who are knowledgeable about the topic and rank the retrieved persons according to their level of expertise[5].

Finding experts is a challenging task. Compared to traditional document retrieval systems, expert finding systems can only use indirect information in the documents to

---

infer the association between the query and a candidate expert. Moreover, it is usually difficult to describe and quantify the expertise. The majority of conventional expert finding models fall into two types: profile-based models and document-based models. As for profile-based models, a profile for each candidate is first built using the documents that are relevant to the candidate. Traditional ad-hoc information retrieval techniques are then applied to the candidates' profiles and candidates are finally ranked according to their profiles. In contrast, for the document-based models, traditional information retrieval method is first applied to find the documents which are relevant to the given topic. Candidates are then found and ranked using these documents.

Although there have been great advances in the field of expert finding in the past few years, the performance of current expert finding systems is still unsatisfactory. A major problem in established models is the lack of the candidate priors' model. Very few attempts have been made to model candidate priors and most of the studies utilized a uniformed distribution for candidate priors. The current study developed a topic-centric candidate priors (TCCP) model for expert finding. This model tried to make use of the information in the whole data corpus to furnish more reasonable candidate priors.

This paper is organized as follows. Section 2 describes related work. Section 3 introduces the baseline model and the TCCP model for the expert finding system. Parameter estimation scheme for the TCCP model is given in section 4. Experimental setup and results are provided in section 5. Finally, conclusions and future work are discussed in section 6.

2   Related Work

Generally speaking, there are three core parts in expert finding models: document priors that encode prior knowledge about the documents, candidate priors that represent the prior probabilities of candidates and the associations between candidates and queries given the documents. Most of the previous researches in the field of expert finding have focused on developing better models to calculate the three core parts[3, 6-8]. Deng et al. used the citation numbers of the documents to calculate document priors [8]. The length of the URL for the document[9] and the document type[5] were also used to obtain document priors. Recently, we have proposed a DocRank-based document priors model for expert finding[6]. It has been suggested that candidate priors can improve the performances of an expert finding system dramatically if appropriately defined. However, there are only very few researches that have addressed this issue and most of the studies still use uniform prior probability for candidate priors [10-15].

3   Topic-Centric Model for Expert Finding

3.1   Baseline Model

The basic idea behind models for expert finding is to evaluate the level of expertise of a candidate expert based on measures of the co-occurrence of the candidate and the query terms in the document collection [16-21]. Suppose $D=\{d_1,d_2,...,d_N\}$ is a collection of

supporting documents, where d_i denotes a document in the collection. Let $q=\{t_1,t_2,\ldots,t_K\}$ denote the user's query, and t_i denote a term in the query. If we represent the candidate as ca, the co-occurrence of the candidate ca and the query q can be written as follows according to [5]:

$$p(ca,q) = p(q \mid ca)p(ca) \tag{1}$$

Here, $p(q|ca)$ denotes the probability of the query q given the candidate ca, while $p(ca)$ is the probability that the candidate ca is an expert. If we assume that $p(ca)$ is uniform as most of the researches have done, the co-occurrence of the candidate ca and the query q is proportional to $p(q|ca)$:

$$p(ca,q) \propto p(q \mid ca) \tag{2}$$

If we extend Equation 2 by using the documents in the collection D to connect the candidate ca and the query q, we can obtain:

$$p(q \mid ca) = \sum_{d \in D_{ca}} p(q \mid d,ca)p(d \mid ca) \tag{3}$$

where D_{ca} is the collection of the documents associated with candidate ca. Assuming that the candidate ca and the query q are conditional independent given document d, we can estimate $p(q|ca)$ as follows:

$$p(q \mid ca) = \sum_{d \in D_{ca}} p(q \mid d)p(d \mid ca) \tag{4}$$

where $p(q|d)$ represents the probability of the query q given document d and can be estimated by inferring a document model for each document [5]. Assuming all the terms in the query are sampled identically and independently, $p(q|d)$ can be expressed as follows according to [22-27]:

$$
\begin{aligned}
p(q \mid d) &= \prod_{t_i \in q} p(t_i \mid d) \\
&= \prod_{t_i \in q} (1 - \lambda_d) \frac{n(t_i, d)}{|d|} + \lambda_d p(t_i)
\end{aligned}
\tag{5}
$$

where $n(t_i,d)$ represents the number of times term t_i is present in document d, and $|d|$ denotes the length of the document. The introduction of the smoothing parameter λ_d ensures that there are no non-zero probabilities due to data sparsity [28-32]. $p(t_i)$ is the prior probability of term t_i

On the other hand, $p(d|ca)$ in Equation 4 can be calculated as follows according to[5]

$$p(d \mid ca) = \frac{a(d,ca)}{\sum_{d' \in D} a(d',ca)} \tag{6}$$

where $a(d,ca)$ represents the association between the document d and the candidate ca. This association can be estimated using the following boolean model:

$$a(d,ca) = \begin{cases} 1, ca \text{ is present in document } d \\ 0, otherwise \end{cases} \tag{7}$$

3.2   Topic-Centric Candidate Priors Model for Expert Finding

We have assumed that the prior probability $p(ca)$ is uniform in the baseline model in section 3.1. Apparently, this assumption is unrealistic because some people such as an authority within the organization[5] may have larger probability to be an expert. In this section, we proposed a new model based on probabilistic topic models to obtain more reasonable candidate priors. Probabilistic topic models were first proposed to analyze the content of documents and the meaning of words [10-13]. The basic idea of topic models is to treat the documents as mixtures of topics, while each topic has a probability distribution over words. In this paper, a TCCP model based on Latent Dirichlet Allocation(LDA)[12] was developed to calculate candidate priors (Fig. 1). LDA is among the most commonly used topic models and is a springboard for many other topic models[14]. A document usually addresses multiple topics which are denoted by the hidden variable Z. For each topic z, different words representing candidates (ca) and terms (t) are used to describe the topic. Fig. 1 also illustrates the procedure by which documents in the corpus are generated:
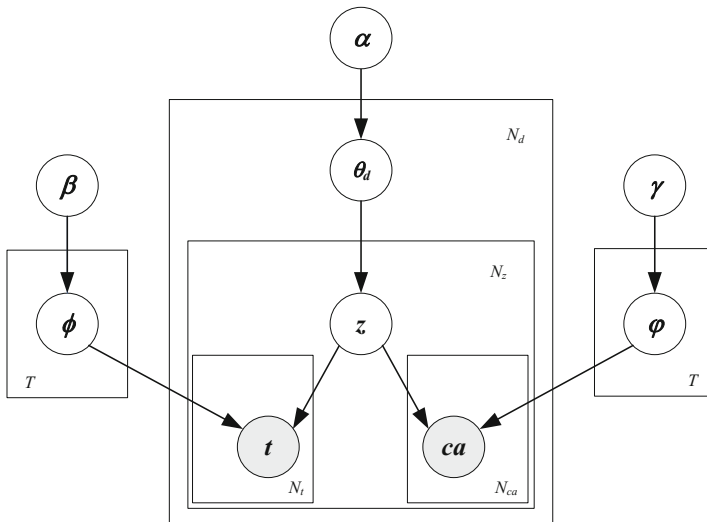
Fig. 1. Graphical model representation of the LDA model used for expert finding

For each document d in the corpus, choose $\theta_d \sim Dir(\alpha)$

For each word (ca or t) in document d, choose a topic $z \sim Mult(\theta_d)$

for each topic z, choose $\phi_z \sim Dir(\beta)$

for each topic z, choose $\varphi_z \sim Dir(\gamma)$

draw a candidate $ca \sim Multi(\phi_z)$

draw a term $t \sim Multi(\varphi_z)$

It can be seen from Fig. 1 that in the topic-centric model a candidate ca is generated by a topic z. Thus the prior probability of a candidate ca can be expressed as:

$$p(ca) = \sum_{z \in Z} p(ca \mid z) p(z) \tag{8}$$

where Z denotes the set of the topics. Similarly, we can also obtain the prior probability of the topic z:

$$p(z) = \sum_{d \in D} p(z \mid d) p(d) \tag{9}$$

Finally, by absorbing Equation 9 to Equation 8, the prior probability of the candidate can be written as :

$$p(ca) = \sum_{z \in Z} p(ca \mid z) \sum_{d \in D} p(z \mid d) p(d) \tag{10}$$

We call this model as the TCCP model.

4   Model Inversion and Parameter Estimation

The Gibbs sampling was used to invert the model and estimate the parameters. Gibbs sampling is a special case of Markov Chain Monte Carlo. According to Fig. 1, the probability of topic z given the candidates and terms $p(z|\mathbf{ca},\mathbf{t})$ can be written as:

$$
\begin{aligned}
&P(z_i = j \mid \mathbf{z}_{-i}, \mathbf{t}, \mathbf{ca}) \\
&\propto P(z_i = j \mid \mathbf{z}_{-i}) P(t_i \mid \mathbf{z}, \mathbf{t}_{-i}) P(ca_i \mid \mathbf{z}, \mathbf{ca}_{-i}) \\
&\propto \frac{H_{dj}^{N_d N_z} + \alpha}{\sum_{j'} H_{dj'}^{N_d N_z} + T\alpha} \frac{H_{mj}^{N_t N_z} + \beta}{\sum_{m'} H_{m'j}^{N_t N_z} + N_t \beta} \frac{H_{nj}^{N_{ca} N_z} + \gamma}{\sum_{n'} H_{n'j}^{N_{ca} N_z} + N_{ca} \gamma}
\end{aligned} \tag{11}
$$

Performing Gibbs sampling according to Equation 10, we can get a series samples of $p(z|\mathbf{ca},\mathbf{t})$. θ_d, ϕ_z and φ_z can then be estimated using in the following way:

$$\hat{\theta}_j^{(d)} = \frac{H_{dj}^{N_d N_z} + \alpha}{\sum_{j'} H_{dj'}^{N_d N_z} + N_z \alpha}$$

$$\hat{\phi}_j^{(t)} = \frac{H_{mj}^{N_t N_z} + \beta}{\sum_{m'} H_{m'j}^{N_d N_z} + N_t \beta} \tag{12}$$

$$\hat{\varphi}_j^{(ca)} = \frac{H_{nj}^{N_{ca} N_z} + \gamma}{\sum_{n'} H_{n'j}^{N_{ca} N_z} + N_{ca} \gamma}$$

Finally, the candidate priors $p(ca)$ can be calculated using estimates of θ_d, ϕ_z and φ_z:

$$p(ca) = \sum p(ca \mid z) p(z)$$
$$= \sum_z p(ca \mid z) \sum_d p(z \mid d) p(d) \tag{13}$$
$$= \sum_z \hat{\varphi}_j^{(ca)} \sum \hat{\theta}_j^{(d)}$$

5   Experimental Setup and Results

5.1   Experimental Setup

5.1.1   Test Corpus

The CERC corpus was used to evaluate the performance of the TCCP model and the baseline model. The CERC corpus comes from the expert search task in the TREC 2007 Enterprise track. The crawl contains 370,715 documents, with a total size of 4.2GB. This data set comprises many different types of documents including web pages, sources codes, pdf documents, etc.

5.1.2   Candidate List and Queries

Unlike the 2005 and 2006 edition of the TREC Enterprise track, a list of possible candidates is not available for the CERC corpus. In that case, our first task is to recognize candidates' occurrences within the documents. In this paper, we built our list of potential candidate experts by modifying the candidate list provided by Balog et al.

[5]. Some emails such as csiro.au@csiro.au which were obviously not candidates were excluded from our candidate list. Our final candidate list consists of 3559 candidates. The candidates' names in the documents were then replaced by their email addresses in the candidate list. After that, the data is parsed and indexed using Java and the Lucene open-search engine. For the queries, we used the 50 queries provided by the CERC corpus.

5.1.3   Evaluation Metrics

The evaluation metrics used to measure retrieval effectiveness included Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), and precision at rank n (P@5, P@10).

5.2   Results

The performance of the proposed model and the baseline model are shown in table 1 while the number of the topics was set to 25. It can be seen from table 1 that by appropriately defining the candidate priors, the performance of the expert finding system can improved dramatically. More specifically, the MAP was improved by 7% while the MRR was increased as much as 12%.

Table 1. Comparison of the performance of the baseline model and the TCCP model (T=25)

Model	MAP	MRR	P@5	P@10
Baseline	33.47	44.77	20.40	13.60
TCCP	35.84	50.24	19.60	14.00

The robustness of the TCCP model was also evaluated while the number of the topics varied from 1 to 40. The MAP and the MRR of the two models are shown in Fig. 2. Apparently, the performance of the TCCP model exceeds that of the baseline model under all the parameter settings. It can also be seen that when the number of the topics is set to 25, the MRR achieves its peak value while we can also get a high MAP value.
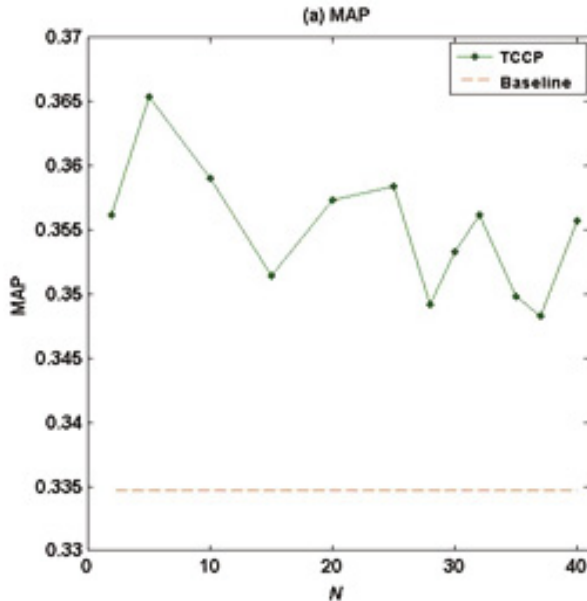
(a)
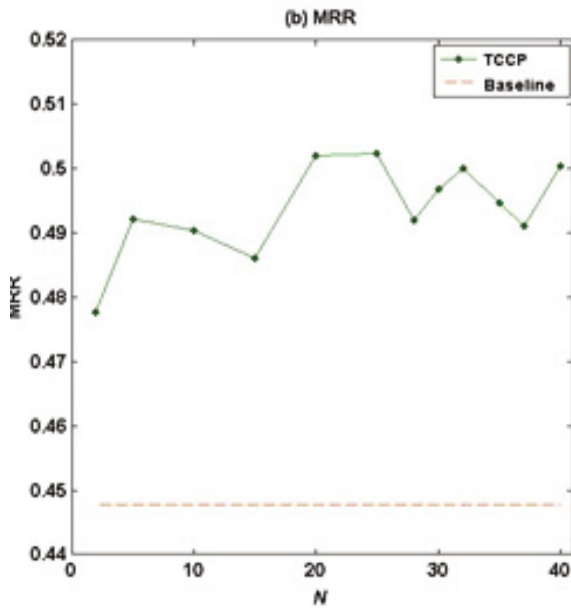
(b)

Fig. 2. Performance of the baseline model and the TCCP model under different parameter settings

6   Conclusion

In this paper, we established a novel model to calculate candidate priors for expert finding based on topic models. One advantage of the topic models is that the global information in the whole data collection can be used to obtain the prior knowledge of the candidates. The comparison between the TCCP model and the baseline model showed that the performance of the expert finding system was dramatically improved by using the candidate priors estimated by the TCCP model. Our future work will focus on develop a unified framework for expert finding by combing the TCCP model and the DocRank-based document priors model which we proposed previously.

References

1. Li, M., Liu, L., Li, C.-B.: An approach to expert recommendation based on fuzzy linguistic method and fuzzy text classification in knowledge management systems. Expert Syst. Appl. 38, 8586–8596 (2011)
2. Guan, Z., Miao, G., McLoughlin, R., Yan, X., Cai, D.: Co-Occurrence Based Diffusion for Expert Search On the Web. IEEE Transactions on Knowledge and Data Engineering, 1 (2012)
3. Fang, Y., Si, L., Mathur, A.: Discriminative probabilistic models for expert search in heterogeneous information sources. Information Retrieval 14, 158–177 (2011)
4. Yang, K.-W., Huh, S.-Y.: Automatic expert identification using a text categorization technique in knowledge management systems. Expert Systems with Applications 34, 1445–1455 (2008)
5. Balog, K.: People Search in the Enterprise. PhD. University of Amsterdam, Amsterdam (2008)
6. Liu, J., Zhang, Y., Liu, B., Zha, Y.: A DocRank-based document priors model for expert search. In: The 2012 International Conference on Intelligence Science and Information Engineering, pp. 62–65. Atlantis Press, Pairs (2012)
7. Macdonald, C., Ounis, I.: The influence of the document ranking in expert search. Information Processing & Management 47, 376–390 (2011)
8. Deng, H., King, I., Lyu, M.R.: Formal Models for Expert Finding on DBLP Bibliography Data. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 163–172 (2008)
9. Balog, K., De Rijke, M.: Combining Candidate and Document Models for Expert Search. In: The Seventeenth Text Retrieval Conference Proceedings (TREC 2008). NIST (2008)
10. Griffiths, T.L., Steyvers, M.: A probabilistic approach to semantic representation, pp. 381–386. Citeseer (2002)
11. Hofmann, T.: Probabilistic latent semantic analysis. In: The Fifteenth Conference on Uncertainty in Artifical Intelligence (1999)
12. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)
13. Steyvers, M., Griffiths, T.: Probabilistic topic models. In: Handbook of Latent Semantic Analysis, vol. 427, pp. 424–440 (2007)
14. Blei, D., Lafferty, J.: Topic models. In: Srivastava, A., Sahami, M. (eds.) Text Mining: Classification, Clustering, and Applications. Taylor & Francis, London (2009)
15. Andon, P.: Accounting-related research in PPPs/PFIs: present contributions and future opportunities. Accounting, Auditing & Accountability Journal 25(5), 876 (2012)

16. Lord, A., Tewdwr-Jones, M.: Is Planning "Under Attack"? Chronicling the Deregulation of Urban and Environmental Planning in England. European Planning Studies 1 (2012)
17. Hodges, R.: Joined-Up Government and the Challenges to Accounting and Accountability Researchers. Financial Accountability & Management 28(1) (2012)
18. Mahony, P., Hextall, I., Richardson, M.: 'Building Schools for the Future': reflections on a new social architecture. Journal of Education Policy 26(3), 341 (2011)
19. Demirag, I., Khadaroo, I.: Accountability and value for money: a theoretical framework for the relationship in public–private partnerships. Journal of Management & Governance 15(2), 271 (2011)
20. Toms, S., Beck, M., Asenova, D.: Accounting, regulation and profitability: The case of PFI hospital refinancing. Critical Perspectives on Accounting 22(7), 668 (2011)
21. Appuhami, R., Perera, S., Perera, H.: Coercive Policy Diffusion in a Developing Country: The Case of Public-Private Partnerships in Sri Lanka. Journal of Contemporary Asia 41(3), 431 (2011)
22. Hodkinson, S.: The Private Finance Initiative in English Council Housing Regeneration: A Privatisation too Far? Housing Studies 26(6), 911 (2011)
23. Arnaboldi, M., Lapsley, I.: Asset management in cities: polyphony in action? Accounting, Auditing & Accountability Journal 23(3), 392 (2010)
24. Rangel, T., Galende, J.: Innovation in public–private partnerships (PPPs): the Spanish case of highway concessions. Public Money & Management 30(1), 49 (2010)
25. English, L., Baxter, J.: The Changing Nature of Contracting and Trust in Public-Private Partnerships: The Case of Victorian PPP Prisons. Abacus 46(3) (2010)
26. Hyndman, N., McDonnell, P.: Governance and Charities: An Exploration of Key Themes and the Development of a Research Agenda. Financial Accountability & Management 25(1) (2009)
27. Hillier, J., Wezemael, J.V.: 'Empty, Swept and Garnished': the Public Finance Initiative case of Throckley Middle School. Space and Polity 12(2), 157 (2008)
28. Benito, B., Montesinos, V., Bastida, F.: An example of creative accounting in public sector: The private financing of infrastructures in Spain. Critical Perspectives on Accounting 19(7), 963 (2008)
29. Athiyo, R., Smith, N.J., Aritua, B.: Private finance for the delivery of school projects in England. Proceedings of the ICE - Management, Procurement and Law 161(4), 141 (2008)
30. Broadbent, J., Guthrie, J.: Public sector to public services: 20 years of "contextual" accounting research. Accounting, Auditing & Accountability Journal 21(2), 129 (2008)
31. Khadaroo, I.: The actual evaluation of school PFI bids for value for money in the UK public sector. Critical Perspectives on Accounting 19(8), 1321 (2008)
32. Jones, M.J., Mellett, H.J.: Determinants of changes in accounting practices: Accounting and the UK Health Service. Critical Perspectives on Accounting 18(1), 91 (2007)

Signer-Independent Sign Language Recognition Based on Manifold and Discriminative Training

Xunbo Ni[1,2], Gangyi Ding[1], Xunran Ni[3], Xunchao Ni[4], Qiankun Jing[1],
JianDong Ma[1,2], Peng Li[1], and Tianyu Huang[1]

[1] School of Software, Beijing Institute of Technology. 100081, Beijing, China
[2] School of Computer Science and Technology, Beijing Institute of Technology. 100081, Beijing, China
[3] Harbin Medical University, Antituberculosis Hospital. 150081, Harbin, China
[4] Southpac Construction pty, Add: 5 Belmore Street Burwood. 2134, NSW, Australia
nixunbo@bit.edu.com, nxb_6@163.com

Abstract. Signer-independent sign language recognition is an urgent problem for the practicability of sign language recognition system. Currently, there is still a huge gap between signer-independent sign language recognition and signer-dependent sign language recognition system owing to the variation of sample data and the insufficient of training samples. Discriminative training can well compensate the recognition shortages caused by insufficient training samples and the similarity of sign language models. This paper applied the HMM (hidden Markov model) training parameter model modified by DT (discriminative training) method to recognize signer-independent sign language. The modified HMM model reduced the effects of small training samples to signer-independent sign language recognition. Furthermore, this paper proposed a tangent vectors method of manifold (TV/HMM) concept to improve the statistical model of sign language recognition considering the learning and reasoning capability of manifold concept in sign language recognition field. The proposed model reduced the variation impact of sign language to signer-independent sign language recognition. Finally, a novel statistical training model (DT+TV/HMM) combining discriminative training and manifold methods was established to solve the data variation and small samples problems of sign language recognition system. Experiments show that the discrimination rate of the integrated DT+TV/HMM recognition system is 82.43% increasing by 15.07% compared with traditional MLE recognition system.

Keywords: HMM, Discriminative training, Manifold, Tangent Vector, DT/HMM (Discriminative training improved HMM model), TV/HMM (Tangent vectors method of manifold improved HMM), DT+TV/ HMM (Discriminative training and Tangent vectors method of manifold improved HMM).

1   Introduction

The research [1-8] on sign language recognition begins in the 90's of the last century. Improving the recognition rate of signer-independent sign language in large

vocabulary case is an essential problem in the practicability of sign language system. Currently, the data set of the signer-independent sign language is small. It can only contain the data for 20 people at most. The Kadous's recognition rate of non-registered set signer-independent sign language is 15% in [12]. The Assan's recognition rate is 51% in [13]. When the training people become more, the results will be better. In [14], the recognition rate of the data set containing 7 people is 85.3%. In [15-16] the recognition rates achieve 85%~91% based on the hand shape classifiers. Kong[17] classified the signer-independent by the 3D trajectories and obtained 91.2% recognition rate. In the previous design of SOFM/HMM[18], the data is turned into a lower dimensional representation. The feature is extracted implicitly. In this case, the extraction lacks the comprehension of sign language. It will cause losing some crucial feature. In the field of speech recognition, the speaker-adaptation algorithm can overcome the shortcomings in signer-dependent and signer-independent system. Ong[19] applied MAP in the estimate of component classifier based on Bayesian networks. The recognition rate achieves 88.5%.

In the current research on sign language recognition, HMM utilizes MLE criterion function and BW（Baum-Welch）algorithm to iteratively revaluate each parameter in the model. The revaluation omits the relevance between models. Compared the maximum mutual information criterion of MMIE[20] and MLE, MMIE not only considers the information of the current model in the training but also the information of other model. Normandin et al.[21] applied EBW (Expend- Baum-Welch) algorithm to solve the open problem of MMIE criterion in the continuous HMM model. Then the method of training distinction is widely introduced in speech recognition. In the field of sign language recognition, the method of training distinction is also applied by the researchers.

The late 70's of twentieth century, there exits manifold model in PAMI. In 2000, there are 3 papers[22-24] in Science. Wang[25] found the continuity rule in the similar things in the nature. People with strong cognitive abilities are based on this stable manifold of visual memory[22]. In recent years, manifold learning has become an attractive issue in the field of machine learning[26,27]. LLE[24] and Isomap[23] are two representations for the nonlinear dimensionality reduction methods. Donoho extended the algorithm of LLE and proposed HLLE. The other improvements are based on Isomap or LLE. For classification tasks, Discriminant Analysis has proved its effectiveness on various applications, i.e. Linear Discriminant Analysis (LDA) and Nonparametric Discriminant Analysis (NDA), Subclass Discriminant Analysis (SDA) and Kernel Fisher Discriminant (KFD).

However, the previous work does not involve how to improve the recognition rate, when the training sample is small or data of sign language varies. First, for the existence of variation in the data of sign language, this paper introduces the concept of Discriminative training improved HMM model (DT/HMM). Secondly, we proposed a tangent vectors method of manifold (TV/HMM) concept to improve the statistical model of sign language recognition considering the learning and reasoning capability of manifold concept in sign language recognition field. Finally, a novel statistical training model (DT+TV/HMM) combining discriminative training and manifold methods was established to solve the data variation and small samples problems of sign language recognition system.

2   Discriminative Training Improved Hidden Markov Model

$$\bar{a}_{ij} = \frac{\sum_{u \in A_v} \sum_{t=0}^{T^u} p_\theta(s_t = i, s_{t+1} = j \mid O^u, \upsilon) - h \sum_{u \in B_v} \sum_{t=0}^{T^u} p_\theta(s_t = i, s_{t+1} = j \mid O^u, \upsilon)}{\sum_{u \in A_u} \sum_{t=0}^{T^u} \psi_i^u(t) - h \sum_{u \in B_v} \sum_{t=0}^{T^u} \psi_i^u(t)} \tag{1}$$

$$\bar{c}_{ik} = \frac{\sum_{u \in A_u} \sum_{t=0}^{T^u} \psi_{ik}^u(t) - h \sum_{u \in B_v} \sum_{t=0}^{T^u} \psi_{ik}^u(t)}{\sum_{u \in A_u} \sum_{t=0}^{T^u} \psi_i^u(t) - h \sum_{u \in B_v} \sum_{t=0}^{T^u} \psi_i^u(t)} \tag{2}$$

$$\bar{\mu}_{ikj} = \frac{\sum_{u \in A_u} \sum_{t=0}^{T^u} \psi_{ik}^u(t)[O_t^u]_j - h \sum_{u \in B_v} \sum_{t=0}^{T^u} \psi_{ik}^u(t)[O_t^u]_j}{\sum_{u \in A_u} \sum_{t=0}^{T^u} \psi_{ik}^u(t) - h \sum_{u \in B_v} \sum_{t=0}^{T^u} \psi_{ik}^u(t)} \tag{3}$$

$$\bar{\sigma}_{jkj}^2 = \frac{\sum_{u \in A_u} \sum_{t=0}^{T^u} \psi_{ik}^u(t)([O_t^u]_j - \bar{\mu}_{ikj})^2 - h \sum_{u \in B_v} \sum_{t=0}^{T^u} \psi_{ik}^u(t)[O_t^u]_j([O_t^u]_j - \bar{\mu}_{ikj})^2}{\sum_{u \in A_u} \sum_{t=0}^{T^u} \psi_{ik}^u(t) - h \sum_{u \in B_v} \sum_{t=0}^{T^u} \psi_{ik}^u(t)} \tag{4}$$

In sign language recognition, HMM based on Gauss mixture probability density is commonly used. The traditional MLE criterion function and BW（Baum-Welch）algorithm is employed to iteratively revaluate each parameter. This method of revaluation only considers the samples in the current training model but omits the relevance between models. The maximum mutual information criterion of MMIE is most commonly applied in the training distinction. Compared with MLE, MMIE not only considers the information of the current model in the training but also the information of other competitive models. The properties can make up MLE criterion well. In the research of Wang], only expectations and variances in DT/HMM are deduced. It achieves good recognition results in the small vocabulary of signer-independent sign language. Ni[32] deduced all parameters in the HMM. And the results are consistent with the revaluation of BW (Baum-Welch) algorithm. DT/HMM acquires ideal effect in the large vocabulary of signer-independent sign language. The revaluation formulas of DT/HMM are as follows and the reduction can be found in.

3   Tangent Vectors Method of Manifold Improved HMM and New Reductions

The metric in the Euclidean space will be influenced by some small deformations which do not affect the category. But when introducing the tangent vector, tangent distance will compensate the influence of the deformations in a certain extent. So it is necessary to introduce the manifold tangent distance in the topological space to improve the original statistical model and construct new statistical model of TV/HMM.

To reduce the computational complexity, the maximum approximation method is applied to deal with the formula (7) in [25]. The maximization of this formula equals to the minimization of distance $d(x,\mu)$. Thus, the metric of distance is unchanged.

$$d(x,\mu) = -2\log p(x\mid\mu) \approx \min_{\alpha}\left\{\frac{1}{\gamma^2}\sum_i \alpha_i^2 + \left(\mu + \sum_i \alpha_i\mu_i - x\right)^T \Sigma^{-1}\left(\mu + \sum_i \alpha_i\mu_i - x\right)\right.$$
$$\left. + \left(\sum_i \alpha_i\mu_i\right)^T \Sigma^{-1}(\mu - x) + \left(\sum_i \alpha_i\mu_i\right)^T \Sigma^{-1}\left(\sum_i \alpha_i\mu_i\right)\right\}$$
(5)

From the formula (7) in [33], we have $\left(\sum_i \alpha_i\mu_i\right)^T \Sigma^{-1}\left(\sum_i \alpha_i\mu_i\right) = \sum_i \alpha_i^2 \mu_i^T \Sigma^{-1}\mu_i$

On the other hand, the second term is independent with α and equals to the third and forth terms. So the formula can be simplified to:

$$d(x,\mu) \approx (\mu - x)^T \Sigma^{-1}(\mu - x) + \min_{\alpha}\left\{\sum_i \alpha_i^2\left(\frac{1}{\gamma^2} + \mu_i^T\Sigma^{-1}\mu_i\right) + 2(\mu - x)^T\Sigma^{-1}\left(\sum_i \alpha_i\mu_i\right)\right\}$$

$$= (\mu - x)^T \Sigma^{-1}(\mu - x) - \sum_i \frac{\left((\mu - x)^T\Sigma^{-1}\mu_i\right)^2}{\frac{1}{\gamma^2} + \mu_i^T\Sigma^{-1}\mu_i} + \min_{\alpha}\left\{\sum_i\left(\frac{1}{\gamma^2} + \mu_i^T\Sigma^{-1}\mu_i\right)\left(\alpha_i + \frac{\left((\mu - x)^T\Sigma^{-1}\mu_i\right)^2}{1/\gamma^2 + \mu_i^T\Sigma^{-1}\mu_i}\right)\right\}$$

$$= (\mu - x)^T \Sigma^{-1}(\mu - x) - \sum_i \frac{\left((\mu - x)^T\Sigma^{-1}\mu_i\right)^2}{1/\gamma^2 + \mu_i^T\Sigma^{-1}\mu_i}$$
(6)

The result is from a weighted sum of squares of the minimized the value of zero. And apply the following expression:

$$x^T\left(A^{-1} + bb^T\right)x = x^TA^{-1}x + x^Tbb^Tx = x^TA^{-1}x + \left(b^Tx\right)^2$$
(7)

We have

$$d(x,\mu) \approx (\mu - x)^T\left(\Sigma^{-1} - \sum_i \frac{\left(\mu_i^T\Sigma^{-1}\right)^T \mu_i^T\Sigma^{-1}}{1/\gamma^2 + \mu_i^T\Sigma^{-1}\mu_i}\right)(\mu - x)$$
(8)

Get the inverse from the undetermined coefficient method:

$$\left(\Sigma^{-1} - \lambda\sum_i \frac{\left(\mu_i^T\Sigma^{-1}\right)^T \mu_i^T\Sigma^{-1}}{1/\gamma^2 + \mu_i^T\Sigma^{-1}\mu_i}\right)\left(\Sigma + \kappa\sum_i \frac{\mu_i\mu_i^T}{1/\gamma^2 + \mu_i^T\Sigma^{-1}\mu_i}\right) = I - (\lambda - \kappa + \lambda\kappa)\sum_i \frac{\Sigma^{-1}\mu_i\mu_i^T}{1/\gamma^2 + \mu_i^T\Sigma^{-1}\mu_i}$$
(9)

When $\lambda - \kappa + \lambda\kappa = 0$ or $\kappa = \dfrac{\lambda}{1 - \lambda}$, the product as a unit matrix. Since the tangent distance is $\lambda = 1$, then $\kappa \to \infty$. And it can be written as:

$$p(x\mid\mu) = N\left(x\mid\mu, \Sigma'\right) \text{ with } \Sigma' = \lim_{\kappa\to\infty}\left(\Sigma + \kappa\sum_i \frac{\mu_i\mu_i^T}{1/\gamma^2 + \mu_i^T\Sigma^{-1}\mu_i}\right)$$
(10)

which means the inverse does not exist in $R^{D\times D}$.

It can be seen as a degenerate normal distribution. More precisely, it is a normal distribution in reduction of feature space from the projection direction of μ_l. This model is called the linear model. It brings a few normalization problems when $\gamma \to \infty$. The spatial distribution is considered to be based on the projection of tangent vector subspace to solve this issue. The formula (7) in [25] can be expressed as follows:

$$p\left(x|\mu,\Sigma\right)=\left(2\pi\right)^{-\frac{d}{2}}\cdot\left|\Sigma+\gamma^{2}\sum_{l=1}^{L}\mu_{l}\mu_{l}^{T}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left[\left(x-\mu\right)^{T}\left(\Sigma^{-1}-\frac{1}{1/\gamma^{2}+1}\sum_{l=1}^{L}\left[\mu_{l}^{T}\Sigma^{-1}\right]^{T}\left[\mu_{l}^{T}\Sigma^{-1}\right]\right)\left(x-\mu\right)\right]\right\}\tag{11}$$

The introduction of the tangent vector only affects the covariance matrix. Denote a covariance matrix $\tilde{\Sigma}$. Its determinant $\left|\tilde{\Sigma}\right|=\left(1+\gamma^{2}\right)^{L}\cdot\left|\Sigma\right|$ is independent with the tangent vector. It can be ignored in the Maximum Likelihood Method. It can be noted the exponent in (11) is the Mahalanobis distance when $\gamma\rightarrow0$ and is tangent distance when $\gamma\rightarrow\infty$. That is, the introduction of tangent distance is adding a correction term in the Mahalanobis distance.

The discriminative training method with HMM only improve the original HMM training model. A statistical model still use mixed Gauss model. The improved model is focused on the parameter training. The purpose is to solve the small sample problem in the process of training.

Manifold HMM is also based on the original HMM training model. The improvement is based on the parameter training. Its purpose is to solve the problem of small sample in the process of training.

The parameters training model and the improvement of tangent vector in discriminative training can be seen in [24, 25]. Combining the above two improvements, we acquired the parameters revaluation of discriminative training with manifold:

① The parameters revaluation utilizes the discriminative training with manifold. Seen (1)~(4) above.

② A statistical model utilizes mixed Gauss model with tangent vector. That is, formula (11) we will deduce below. $p\left(x_{k,t}\mid s\right)=\sum_{m=1}^{M}c_{sm}N\left(x_{k,t}\mid\mu_{sm},\Sigma_{sm}\right)$, We have

$$G[\mu_{jm},\Sigma_{jm},O_{k}]=\frac{1}{\sqrt{2\pi}\,|\Sigma_{jm}+\gamma^{2}\sum_{l=1}^{L}\mu_{jml}\mu_{jml}^{T}|^{\frac{1}{2}}}\cdot\exp[-\frac{1}{2}(O_{k}-\mu_{jm})^{T}\left(\Sigma_{jm}^{-1}-\frac{1}{1/\gamma^{2}+1}\sum_{l=1}^{L}\left[\mu_{jml}^{T}\Sigma_{jm}^{-1}\right]^{T}\left[\mu_{jml}^{T}\Sigma_{jm}^{-1}\right]\right)(O_{k}-\mu_{jm})]\tag{12}$$

4   Experiments of Combining Manifold Model with Discriminative Training Method

The experiment is based on large vocabulary set. The database is composed by 61356 samples. 6 signers acquire 5113 sign language words twice, respectively. In the experiment to get the parameter h, 6 signers in a group as a registered test set and the rest data as the training set of DT/HMM. In the signer-independent sign language recognition experiments, cross validation method is applied. 5 personal data (10) is for training, leaving one data as non-registered training set (UnReg). The details of confusion set can be referred in [26].

4.1   Experimental Analysis of Parameter h in DT+TV /HMM

As Fig.1"*" represent the experimental results of combining manifold model with discriminative training method. That is, the results of discriminative training and Tangent vectors method of manifold improved HMM (DT+TV/HMM). The blue line

represents the result of discriminative training purely using HMM. Since in the discriminative training the probability density of tangent vector is introduced, it makes the parameter h bigger. It indicates that TV/HMM with the probability density of tangent vector raises the recognition rate and speeds up the convergence of parameters in the discriminative training compared with that of HMM. And it also states that sign language data has the manifold structure. In the registered set, the recognition rate is 93.6% and h is 1.9.
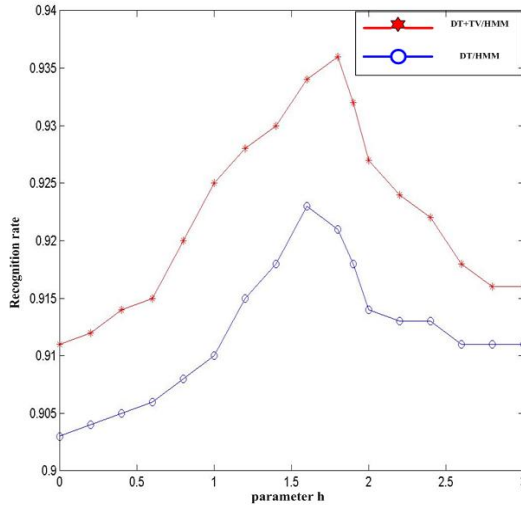
Fig. 1. The recognition rate changing with DT/HMM parameter h

4.2   Experimental Analysis of the Iteration Number in DT+TV /HMM

Through a lot of experiments on the parameter training iterations, the best convergence result of discriminative training is in the range of 4 to 6 times of iterations. The recognition rate reduces, when it is out of the range. The best convergence result of manifold is in the range of 10 to 12 times of iterations. The recognition rate reduces, when it is out of the range. The best convergence result of combining manifold with discriminative training is in the range of 8 to 10 times of iterations. The recognition rate reduces, when it is out of the range. It states that in the criterion H of discriminative training h speeds up the convergence of manifold. Manifold tangent vector also reduce the divergence degree of the discriminative training when the recognition rate is reduced. Fig.2 shows as follows.
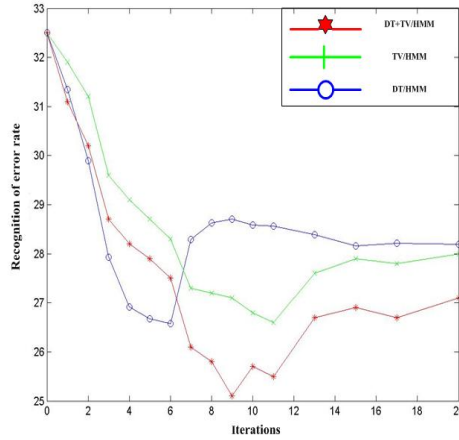
Fig. 2. The recognition of error rate changing with DT+TV/HMM iterations

4.3   Experimental Analysis of the Tangent Vectors and Parameter γ in DT+TV /HMM

In the criterion H of discriminative training, h highly speeds up the convergence of revaluation. In the case of combining manifold with discriminative training, it speeds up the convergence of parameters in the manifold and makes the parameter γ increasing with the error recognition rate. But it does not influence tangent vector much. Here, γ is set to 4 and the number of tangent vectors is 7, which remains unchanged. As shown in Fig.3 and Fig.4, the red line represents the experimental result of DT+TV /HMM and the green line represents the experimental result of TV /HMM.
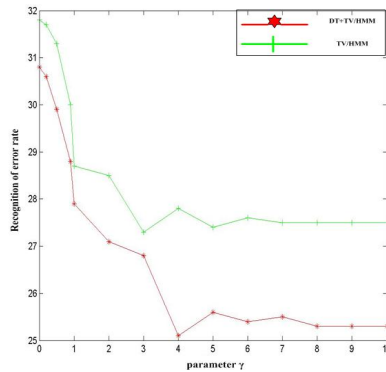
Fig. 3. The recognition of error rate changing with DT+TV/HMM parameter γ
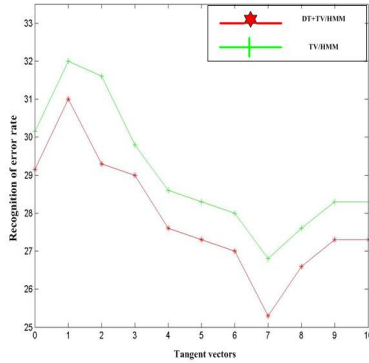
Fig. 4. The recognition of error rate changing with the number of DT+TV/HMM tangent vectors

From Table 1, introducing tangent vectors into the recognition system of non-registered set is much better than the traditional recognition system. The recognition rate of non-registered confusion set NEW, which does not contain subjective experience, increases by 11.32% and 10.22% respectively, comparing with MLE and EBW confusion sets. When adding subjective experience, the recognition rate increases by 15.07% and 13.97% respectively, comparing with MLE and EBW confusion sets.

Compared the results in Table 1 with that of [24, 25], the average recognition rate of non-registered confusion set NEW increases by 4.6% and 2.54%, respectively. The average recognition rate of non-registered confusion set EXP increases by 4.42% and 2.57%, respectively.To sum up, the statistical training model(DT+TV/HMM) combining discriminative training and manifold methods can solve the data variation and small samples problems of sign language recognition system

Table 1. Recognition results of DT+TV/HMM unregistered confusion sets on a vocabulary of 5113 signs

Signer	MLE(%)	EBW(%)	NEW (%)	EXP(%)
ljh	69.43	70.56	79.53	83.49
ygy	62.96	64.08	76.77	79.64
lwr	66.77	67.82	78.35	82.75
lz	66.95	68.05	77.27	82.32
my	70.71	71.81	81.53	83.95
Average	67.36	68.46	78.68	82.43

5   Conclusion

As the variation of sign language data and small training sample, the recognition rate of the signer-independent sign language is much lower than that of signer-dependent sign language. One reason is that the training method of traditional hidden Markov model (HMM) is based on the maximum likelihood criterion (MLE). On the other hand, HMM can not describe the variation of the sign language data. This paper applies discriminative training (DT) method to solve the problem of the insufficient training samples in the signer-independent sign language recognition. We also proposed a tangent vectors method of manifold (TV/HMM) concept to improve the statistical model of sign language recognition considering the learning and reasoning capability of manifold concept in sign language recognition field. The experimental results show that the DT+TV/HMM model based on the two properties has better recognition effects.

References

1. Charayaphan, C., Marble, A.: Image processing system for interpreting motion in American sign language. Journal of Biomedical Engineering 14(15), 419–425 (1992)
2. Fels, S.S., Hinton, G.E.: Glove-talk: a neural network interface between a data-glove and a speech synthesizer. IEEE Trans. Neural Networks 4(1), 2–8 (1993)
3. Kadous, M.W.: Machine recognition of Auslan signs using PowerGloves: towards large-lexicon recognition of sign language. In: Proceedings of the Workshop on the Integration of Gesture in Language and Speech, pp. 165–174. DE, Wilmington (1996)
4. Grobel, K., Assan, M.: Isolated sign language recognition using hidden markov models. In: Proceedings of the International Conference of System, Man and Cybernetics, Orlando, pp. 162–167. (1997)
5. Sagawa, H., Takeuchi, M., Ohki, M.: Description and recognition methods for sign language based on gesture components. In: Proceedings of the International Conference on Intelligent User Interfaces, Florida, USA, pp. 97–104 (1997)
6. Kim, J.S., Jang, W., Bien, Z.: Dynamic gesture recognition system for the Korean sign language (KSL). IEEE Trans. Systems, Man, and Cybernetics 26(2), 354–359 (1996)
7. Deng, J.W., Tsui, H.T.: A two-step approach based on PaHMM for the recognition of ASL. In: Proceedings of the Fifth Asian Conference on Computer Vision, Melbourne, Australia, pp. 126–131 (2002)
8. Shamaie, A., Sutherland, A.: Accurate recognition of large number of hand Gestures. In: Second Iranian Conference on Machine Vision and Image Processing, New York, USA, pp. 308–317 (2003)
9. Chen, F.S., Fu, C.M., Huang, C.L.: Hand gesture recognition using a real-time tracking method and hidden Markov models. Image and Vision Computing 21(8), 745–758 (2003)
10. Huang, C.L., Jeng, S.H.: A model-based hand gesture recognition system. Machine Vision and Application 12(5), 243–258 (2001)
11. Kobayashi, T., Haruyama, S.: Partly-Hidden Markov model and its application to gesture recognition. In: Taylor, F.J. (ed.) Proc. of the Int'l Conf. on Acoustics, Speech and Signal Processing, pp. 3081–3084. Academic Press, New York (1997)

12. Kadous, M.W.: Machine recognition of auslan signs using PowerGloves: Towards large-lexicon recognition of sign language. In: Messing, L. (ed.) Proc. of the Workshop Integration of Gestures in Language and Speech, pp. 165–174. IEEE Computer Society Press, Delaware (1996)

13. Assan, M., Grobel, K.: Video-Based sign language recognition using hidden Markov models. In: Wachsmuth, I., Fröhlich, M. (eds.) GW 1997. LNCS (LNAI), vol. 1371, pp. 97–109. Springer, Heidelberg (1998)

14. Vamplew, P., Adams, A.: Recognition of sign language gestures using neural networks. Australian Journal of Intelligent Information Processing Systems 5(2), 94–102 (1998)

15. Handouyahia, M., Ziou, D., Wang, S.: Sign Language recognition using moment-based size functions. In: Proc. of the Int'l Conf. on Vision Interface, pp. 210–216. CRC Press, Kerkyra (1999)

16. Su, M.C.: A fuzzy rule-based approach to spatio-temporal hand gesture recognition. IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews 30(2), 276–281 (2000)

17. Kong, W.W., Ranganath, S.: 3-D Hand trajectory recognition for signing exact English. In: Proc. of the Int'l Conf. on Automatic Face and Gesture Recognition, pp. 535–540. ACM Press, Seoul (2004)

18. Gao, W., Fang, G.L., Zhao, D.B., Chen, Y.Q.: A Chinese sign language recognition system based on SOFM/SRN/HMM. Pattern Recognition 37(12), 2389–2402 (2004)

19. Ong, S., Ranganath, S.: Deciphering gestures with layered meanings and signer adaptation. In: Proc. of the Int'l Conf. Automatic Face and Gesture Recognition, pp. 559–564. ACM Press, Seoul (2004)

20. Bahl, L.R., Brown, P.F., de Souza, P.V., Mercer, R.L.: Maximum mutual information estimation of hidden markov model parameters for Speech recognition. In: Proc. 1986 Int. Conf. on Acoustics, Speech and Signal Processing, Tokyo, Japan, pp. 49–52 (April 1986)

21. Normandin, Y.: An improved MMIE training algorithm for speaker independent, small vocabulary, continuous speech recognition. In: Proc. ICASSP 1991, Toronto, pp. 537–540 (1991)

22. Seung, H.S., Lee, D.D.: The manifold ways of perception. Science 290(5500), 2268–2269 (2000)

23. Tenenbaum, J., Silva, D.D., Langford, J.: A global geometric framework for nonlinear dimensionality reduction. Science 290(5500), 2319–2323 (2000)

24. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500), 2323–2326 (2000)

25. Wang, S.-J.: Bionic(Topological) pattern recognition—a new model of pattern recognition theory and its application. ACTA Electronica Sinica 30(10), 1417–1420 (2002)

26. Michigan State University, http://www.cse.msu.edu/~lawhiu/manifold/

27. Zhang, J.-P., Wang, Y.: Manifold Learning, pp. 172–207. Tsinghua Press, Peking (2004)

Integrated Pollutants Control Equipment Optimization with Fluidized Bed Absorber Based on Flow Field Simulation

Xiaowen Hao[1], Mingchen Qin[2], Jianguo Yang[1,*], and Yi'e Zhao[1]

[1] School of Automotive Engineering, Harbin Institute of Technology at Weihai, Weihai, 264209, China
[2] National Engineering Laboratory for Coal-Fired Pollutants Emission Reduction, Shandong University, Jinan, 250061, China
`yjglryz@sina.com`

Abstract. The existing flue gas pollutants control system is facing great challenge with the promulgation of Emission Standard of Air Pollutants for Thermal Power Plants (GB13223-2011) in China. So much money is needed to improve or reinstall the pollutants disposal system, which is hard for the plants. The simple and low cost equipment called as the integrated pollutants control equipment with the fluidized bed absorber was firstly put forward to absorb the additional pollutants between the old and the new Chinese emission standards. Designed of the equipment was proceeded based on the uniform bag filtration and the AC absorption. The absorber was optimized by the flow field simulation. Two guide plates' angle of 45°, the uniformly distributed bags and the opposed bend pipes can make the flow field in the IPCE uniform. The large AC inlet velocity is good for the uniform AC distribution.

Keywords: Flow Field, Uniformity, Fluidized Bed Absorber, Activated Carbon, Bags.

1 Introduction

At the end of the 11th Five-Year Plan of China, the fly ash and SO_2 from the thermal power plants had been scale controlled. Low-NOx combustion technologies and flue gas denitration technologies were quickly put into the engineering. The Emission Standard of Air Pollutants for Thermal Power Plants (GB13223-2011) [1] implemented from Jan. 1st, 2012. This new emission standard for SO_2, NOx and Hg was much stricter than the old one. So the existing pollutants control system is facing great challenge with the promulgation of this new standard. But it is hard for the plants to spend much money to improve the pollutants control system. In some plants, the system has even to be rebuilt. A good way is to develop new technology which can remove the additional pollutants between the old and the new standards without modifying the existing desulfurization system, the denitration system, and the mercury control system. The integrated pollutants control technologies are attracting more attention.

---

[*] Corresponding author.

One kind of the integrated pollutants control technologies is focus on developing the synergistic SO_2 and NOx removal technologies based on the existing Ca-based FGD process [2-4]. These integrated technologies need the additives or the assistance of the electric field or the light field. So much extra money is needed, and the technologies are not mature till now. The other kind of the integrated technologies uses the activated carbon (AC) as absorbent [5, 6]. The AC can be installed in the absorbers. The absorbers are fixed bed, mobile bed, and fluidized bed. The absorbers have been applied only in small-scale power plants because of the large bed volume and the AC price.

The application of the AC is flexible, and it can be applied to absorb the additional pollutants between the old and the new emission standards. But the difficulties are how to make the operational and the modified cost low, make the apparatus area and the pressure drop small when the AC is applied in all kinds of power plants.

Only the bag filter and the electrostatic bag filter can satisfy the GB13223-2011 in the long time running by capturing PM2.5 effectively among all kinds of the dust collectors. There is large space in the bag filters. And the bags' filtration velocity has a broad range. So the AC absorber can be installed in the bag filters by replacing one or several bags. Thus novel integrated equipment called as the integrated pollutants control equipments (IPCE) was formed.

The fluidized bed absorber has the advantage of high mass transfer, and of removing particulates compared to the fixed bed and the mobile bed [7, 8]. The fluidized bed absorber can be sustained operated, and the AC can be easily recycled or be replaced in the fluidized bed absorber. So we selected the fluidized bed absorber as the AC absorber. The IPCE with the fluidized bed absorber was designed firstly. Then the flow field of gas-solid phase in the equipments was simulated and optimized.

2   Equipment Setup

The IPCE with fluidized bed absorber was designed for the shaken filter bags. The ash cleaning technologies called as the reverse gas blowing and the pulse-jet cleaning cannot be applied. It can be seen from Fig. 1 that the bags firstly removed almost all of the particles, and then the fluidized bed absorber absorbed the gas pollutants. The flue gas flux Q_v of the IPCE was set as 500 Nm^3/h. Its temperature in the flue gas inlet was 70 ℃.

Fig. 2 shows the IPCE with the fluidized bed absorber. The IPCE was cylindrical. Its middle was the fluidized bed absorber. The flue gas flowed firstly from the top cabinet, through the bags, and then into the fluidized bed absorber. The cleaned flue gas flowed into the cyclone which could separate the AC from the flue gas.

Six cylindrical bags with the height of 3000 mm were set as outer-filtrate bag-type. The bags were uniformly distributed around the absorber. The bags were numbered from 1 to 6. The blown ash was collected by the ash bunker and discharged by the air lockers. The bend pipes were imperforated.
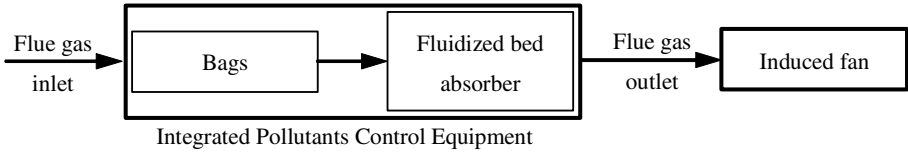
Fig. 1. Flow path of the flue gas in the IPCE
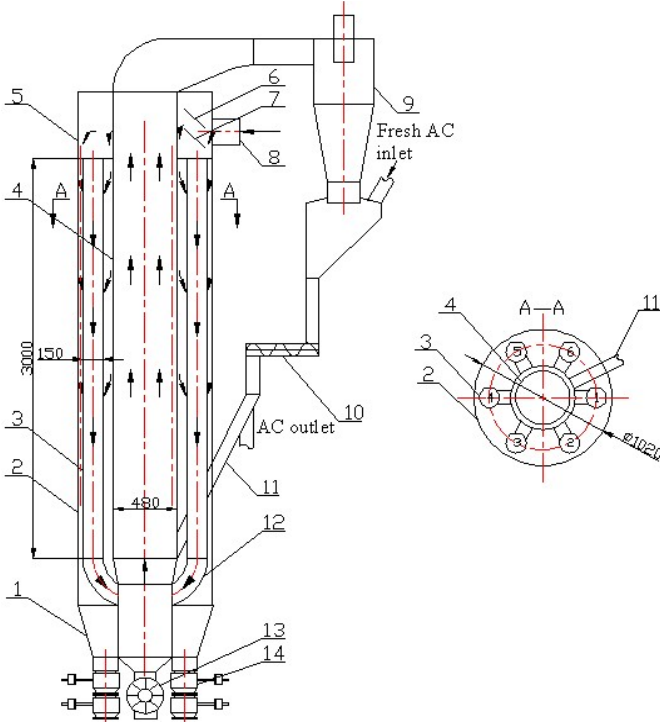
(1. ash bunker, 2. middle cabinet, 3. filter bags, 4. fluidized bed absorber, 5. top cabinet, 6. guide plates ①, 7. guide plates ②, 8.flue gas inlet, 9. cyclone, 10. screw feeder, 11. AC recycling pipe, 12. bend pipes, 13. star wheel type feeder, 14. air lockers.)

Fig. 2. IPCE with fluidized bed absorber

The fresh AC was added from the cyclone bottom. The AC could be recycled to the fluidized bed or be discharged after it was transported by the screw feeder. The AC could be regenerated for several times. So the operation cost was low. The cylindrical absorber could be installed by removing one or several bags from the bag filter. The modification was simple enough so that the transformation from the bag filters to the IPCE was very easy. The corresponding modification cost was low. The area of the IPCE mainly comprised the cyclone and the former bag filter. The added area compared to the bag filter was small enough that the IPCE can be easily accepted

by the power plant. The bags collection efficiency and the AC absorption efficiency are high if the flow field is uniform. And the flow field optimization was focus on the flow uniformity in the IPCE.

3   Simulation Setup

Air was used to substitute the flue gas in the simulation. The cyclone, the star wheel type feeder, the air lockers and the absorber in the ash bunker were ignored in the model.

The standard k-ε model was used to simulate the turbulent flow. Due to the dilute flow in the cyclone, the interaction between particles was neglected. The Lagrangian approach without considering the particle interactions, which was called discrete phase model (DPM), was used to track individual particles through the continuum fluid.

The flue gas inlet was set as velocity-inlet. The top of the fluidized bed absorber was set as pressure outlet. The uniform inlet velocity distributions and fully developed outlet were set. No slip condition was adopted at the wall. The gas turbulent transport was simulated by standard wall function. The unsteady flow field and the ten-second step time were set in the simulation.

The particle diameter in the flue gas fitted Rosin-Rammler diameter distributions. The mean fly ash diameter was 91μm. The boundary condition of the bags was set as porous jump, whose parameter was from Wang [9]. The operational fluidization velocity was two times higher than the critical fluidization velocity of the AC [7]. The AC was set as solid in the simulation. The critical fluidization velocity was identified as

$$u_{mf} = \sqrt{\frac{d_p(\rho_p - \rho_g)g}{24.5\rho_g}} \tag{1}$$

where u_{mf} is the critical fluidization velocity, d_p is the mean diameter of AC, ρ_p is true density of AC, ρ_g is the flue gas density.

The average diameter of the AC d_p was set as 0.28 mm. The true density of the AC was set as 2.13×10^3 kg/m^3.

4   Result and Discussion

Flow uniformity comprises the uniform flux in the bags to get uniform filtration, and the uniform flow in the fluidized bed absorber to get the uniform AC absorption. The guide plates are the usual way to get the flow uniform in the bags filters. Deng [10] considered that the guide plates angle from 30° to 60° were better. Table 1 shows the guide plates' angles between the horizontal plane and the guide plates in the flow field simulation. The angle between the adjacent bend pipes was 60°. And the exit of two bend pipes were opposed each other in order to make the flow in the fluidized bed absorber easy to be uniform.

4.1 Flux Uniformity of the Bags

The ideal flux in each bag is equal. So the impact to each bag by the flue gas is same, and then the lifetime of each bag is same. But the equal flux of the bags is hard to get because the gas flows into the bag filter from the horizontal pipe. The flux uniformity of the bags was optimized by the guide plates' angles. The gas mass flux ratio was defined as the simulated gas mass flux divided by the total flux. The flux nonuniformity coefficient was put forward to analyze the flux uniformity of the bags. Smaller nonuniformity coefficients meant more uniformity. The nonuniformity coefficient was defined as the root-mean-square of the flux in the bags.

$$\zeta = [\frac{1}{n} \sum_{i=1}^{n} (\frac{Q_i - \overline{Q}}{\overline{Q}})^2] \tag{2}$$

where ζ is the flux nonuniformity coefficient of bags in each model, Q_i is the flux of a bag, $\overline{Q}$ is the average flux of the bags, and n is the number of the bags, which is six.

Table 1. Flux ratio of the bags and flux nonuniformity coefficient of the bags

Bags number	Flux ratio, %					
		Guide plate angle				
	Without guide plates	30°	① 30° ② 45°	45°	① 45° ② 30°	60°
1	-15.2	22.3	19.4	17.5	28.1	29.4
2	20.6	27.0	15.5	19.2	23.1	33.8
3	23.4	28.5	17.1	11.1	8.6	18.7
4	25.3	19.4	15.6	19.9	-1.4	20.5
5	23.1	22.6	14.3	14.6	16.1	28.2
6	22.8	-19.8	18.1	17.7	25.5	-30.6
Flux nonuniformity coefficients, ζ						
	0.739	0.992	0.011	0.033	0.385	1.706

Table 1 shows the flux ratio of the bags and flux nonuniformity coefficient of the bags. It can be seen from Table 1 that the uniformity of the bags in the IPCE without the guide plates was bad. Flux in a bag was even negative. The guide plates can be used to make the flux in the bags uniform. But the unreasonable designed plates (such as 60°) made the flux in the bags even worse uniform than the IPCE without guide plates. The flux uniformity of IPCE with two guide plates' angle of 45°, guide plates' angles of 30° and 45° were better.

(a)without guide plates   (b) 30°   (c) ①30° ②45°

(d) 45°   (e) ①45°②30°   (f) 60°
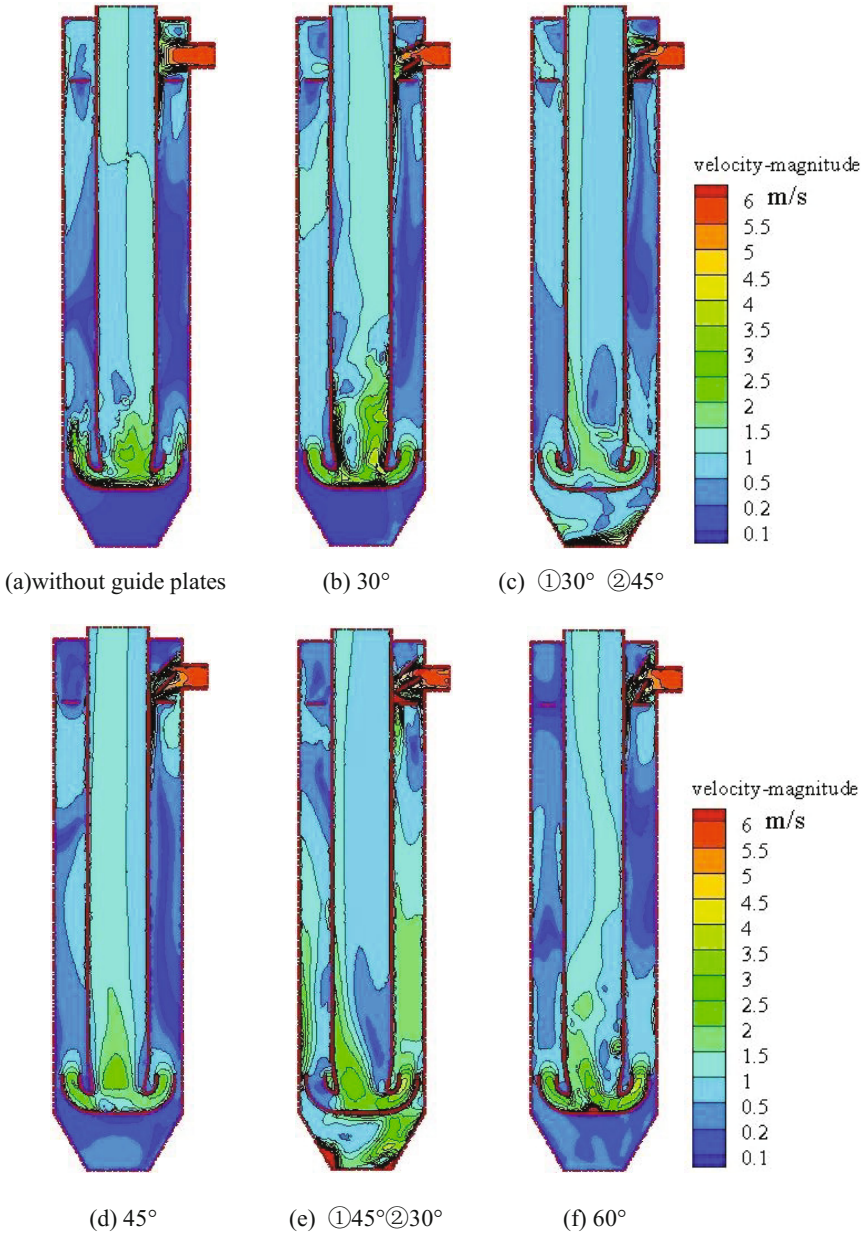
Fig. 3. Velocity magnitude profile in IPCE for different plate angles
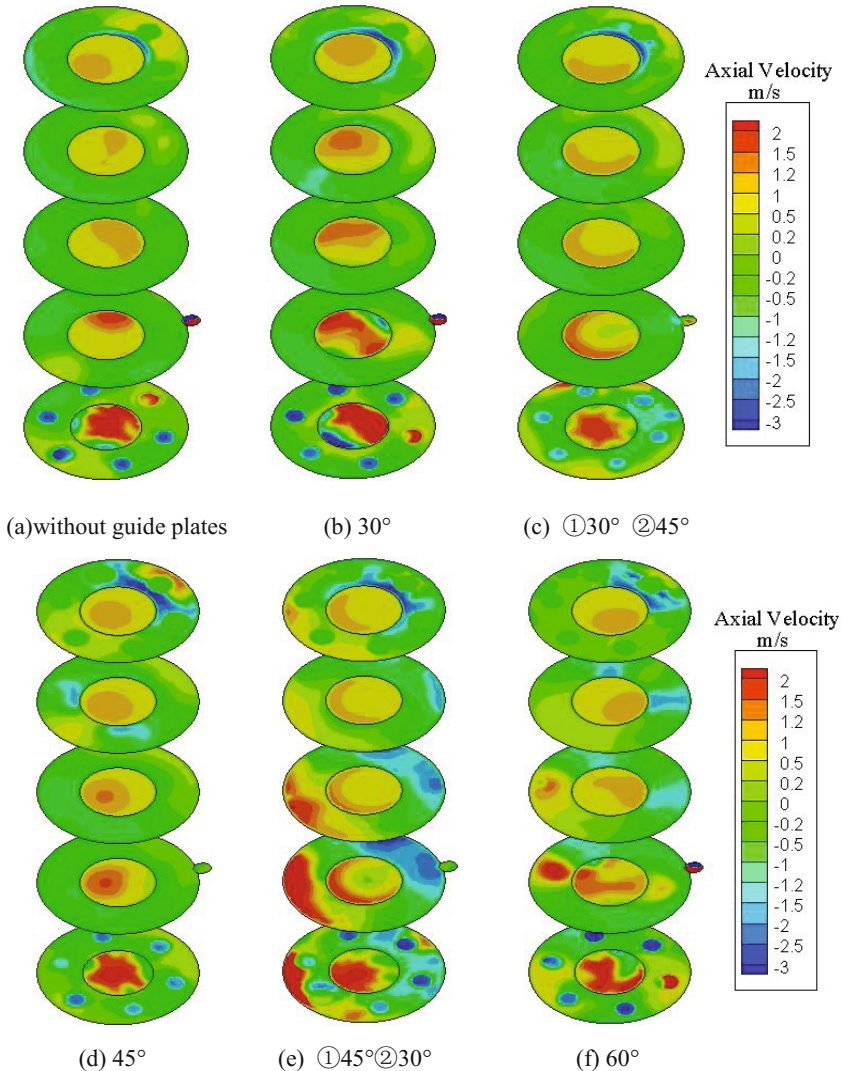
Fig. 4. Axial velocity profile of the plane surfaces in the middle cabinet for different plate angles

4.2 Flow Field in the Middle Cabinet and the Fluidized Bed Absorber

The guide plates affect the flux of the bags and the flow field in the middle cabinet and the fluidized bed absorber. Fig. 3 shows the velocity magnitude profile in the IPCE. Flow field in the IPCE was of great difference with the guide plates' angles. Flow field in the middle cabinet was hard to be uniform after the simulation of many IPCE models. But the flow field in the IPCE with the reasonable designed guide plates was better than that without guide plates. The flow field uniformity of the

plates' angle of 45° was much better than the other angles, especially in the fluidized bed absorber. The velocity magnitude in the fluidized bed absorber flowed to bed wall when the plates' angle was 45°. The reason was the transportation of AC from the recycling pipe.

Five plane surfaces were used to analysis the flow field in the middle cabinet. The top surface was the top of the middle cabinet. The distance of the two adjacent surfaces was 750 mm.

Fig. 4 shows the axial velocity profile of the plane surfaces in the middle cabinet. The AC from the recycling pipe affected the axial velocity in the fluidized bed absorber. But the influence of the guide plates' angles to the flow field was higher than that of the transportation of the AC. Similar regulation can be seen in Fig. 4 as in Fig. 3 that the flow field of the IPCE with the two guide plates' angles of 45° was best.

From the analysis of the flow field in the middle cabinet and the fluidized bed absorber, we found that the flow field could be more uniform with the guide plates and with the arrangement of the bags and the bend pipes.

4.3   Pressure Drop of the IPCE

Table 2 shows the pressure drop of the IPCE with the guide plates higher than it without the guide plates. The guide plates can decrease the pressure drop of the IPCE between the flue gas inlet and the fluidized bed absorber outlet. The lower of the flux nonuniformity coefficient, the lower of the pressure drop was from Table 1 and Table 2. And the flow field uniformity with the two guide plates' angles of 45°, the guide plates' angle of 30° and 45° were much better than the others. So the flow uniformity can decreased the pressure drop of the IPCE.

Table 2. Pressure drop of the IPCE with the guide plates higher than it without the guide plates

Guide plate angles	30°	(1) 30° (2) 45°	45°	(1) 45° (2) 30°	60°
pressure drop,%	50.9	-13.0	-8.2	47.7	25.8

4.4   Trajectories of the AC

The uniform distribution of the AC is hoped in the fluidized bed absorber, which is good for uniform absorption. The AC velocity is a key parameter when the IPCE is optimized. Fig. 5 shows trajectories of the AC in the fluidized bed absorber for the different AC inlet velocity. 4840 AC particles were traced. Only a few AC flowed into the fluidized bed absorber when the AC in the pipe inlet was 0.1m/s. More AC flowed into the fluidized bed absorber when the AC inlet velocity was higher. A lot of the AC particles flowed to the absorber bottom, and were not fluidized at 0.5 m/s. Most AC can be fluidized and the distribution of the AC particles was the most uniform at 1m/s than the other simulated AC inlet velocity.

(a) 0.1m/s   (b) 0.5m/s   (c) 1m/s
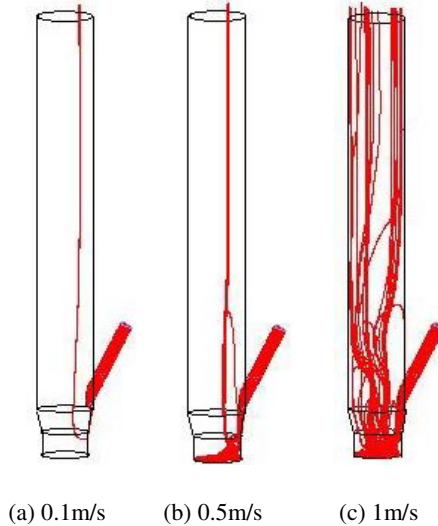
Fig. 5. Trajectories of the AC in the fluidized bed absorber for the different AC inlet velocity

5   Conclusion

The integrated pollutants control equipments (IPCE) can be applied to absorb the additional pollutants between the old and the new emission standards. The IPCE could be easily formed by installing the fluidized bed absorber into the bag filter by replacing one or several bags. The AC was transported into the absorber for several times by the cyclone to absorb gas pollutants. The corresponding modification cost and the operation cost are low. The area of the IPCE is a bit larger than the revised bag filter.

The IPCE was optimized by the flow field simulation. The IPCE with the fluidized bed absorber was designed for the shaken filter bags. The reasonable designed angles of the guide plates in the top cabinet of the IPCE made the bags flux uniform, the flow field uniform and the pressure drop low. Two guide plates' angle of 45°, the uniformly distributed bags and the opposed bend pipes could make the flow field in the IPCE uniform. More AC could flow into the absorber, and these AC particles were more uniformly distributed when the AC inlet velocity was larger. The AC inlet velocity being 1m/s was good from the simulation.

References

1. Emission Standard of Air Pollutants for Thermal Power Plants, GB 13223-2011, National Standard of P. R. China, China Environmental Science Press, Bei Jing (2011)
2. Gao, J.H., Chen, G.Q., Liu, J.X., et al.: Simultaneous Removal of SO_2 and NOx by Calcium Hydroxide at Low Temperature: Evolution of the Absorbent Surface Structure. Energy & Fuels 24, 5454–5463 (2010)
3. Hutson, N.D., Krzyzynska, R., Srivastava, R.K.: Simultaneous Removal of SO_2, NOx, and Hg from Coal Flue Gas Using a $NaClO_2$-Enhanced Wet Scrubber. Ind. Eng. Chem. Res. 47, 5825–5831 (2008)
4. Xu, F., Luo, Z.Y., Cao, W., Wang, P., et al.: Simultaneous Oxidation of NO, SO_2 and Hg^0 from Flue Gas by Pulsed Corona Discharge. J. Environ. Sci. 21, 328–332 (2009)
5. Olson, D.G., Tsuji, K., Shiraishi, L.: The reduction of gas phase air toxics from combustion and incineration sources using the MET-Mitsui-BF activated coke process. Fuel Process Technology 65-66, 393–405 (2000)
6. Ma, S., Yao, J., Gao, L., Ma, X., Zhao, Y.: Experimental study on removals of SO2 and NOx using adsorption of activated carbon/microwave desorption. Journal of the Air & Waste Management Association 62(9), 1012–1021 (2012)
7. Chiang, B.C., Wey, M.Y., Yeh, C.L.: Control of acid gases using a fluidized bed adsorber. Journal of Hazardous Materials 101(3), 259–272 (2003)
8. Zhang, L.Q., Jiang, H.T., Li, B., Cheng, L., Ma, C.Y.: Kinetics of SO_2 adsorption by powder activated carbon using fluidized bed reactor. Journal of China Coal Society 37(6), 1046–1050 (2012)
9. Wang, Y.: Analysis and Application of Large-scale Bag Filter Simulation. M.S. thesis. Donghua University, Shanghai, China (2010)
10. Deng, Y., Guo, L.Y.: Research on the Diversion of the Movable Guide Plates. Yellow River 28(5), 28–29 (2008)

PID Neural Network Adaptive Predictive Control for Long Time Delay System

Yang Zhi-gang and Qian Jun-lei

College of Electrical Engineering, Hebei United University,
Tangshan city, Hebei Province, China. 063009
`yzg-lzy@163.com`

Abstract. System with long time delay is difficult to control, and according to this fact an adaptive predictive controller based on PID neural network is proposed in this paper. This method identifies object model by PID neural network, and overcome the long time delay of the controlled value by recursive prediction. PID neural network based controller realizes the coordination control of overshoot and settling time of the system. A simulation of this control method is implemented on the electric furnace with large time delay, and the results show that the method has a faster system response, stronger adaptability and robustness.

Keywords: PID Neural Network, Neural Network Identification, Recursive Prediction, Neural Network Control，Long Time Delay System.

1   Introduction

In industrial production process, large time delay often occurs in the system, so it's difficult for the conventional PID controller to achieve the ideal control effect. With the development of intelligent control theory and technology, they are introduced to the process control of systems with large time delay, and intelligent PID predictive controller based on neural network is proposed [1]. Literature [2] proposes single neuron prediction PID control; literature [3] proposes PID controller based on three layers BP neural network. But it is difficult to determine the structure of BP network and the initial weights, at the same time BP network has the shortage of local minimum, slow training speed and low efficiency. Literature [4] overcomes the shortage of local minimum in the certain extent by using RBF radial basis networks instead of BP neural network.

PIDNN, Proportional Integral Derivative Neural Network, 1997 first proposed by Shu Huailin, a large number of theoretical and simulation study have been taken in a single variable system and multivariable system respectively [5-8]. PIDNN is a neural network fusion PID rules, the realization of fusion is by giving proportion, integral, differential function to the neurons in the hidden layer. By using error back propagation algorithm through the online training and self learning, PID neural network can modify the connection weight value, so that the optimal performance of

the control system can be achieved. Literature [9] realizes the nonlinear system identification and control by PID neural network. This paper, mainly based on the identification and control principle of literature [9], in view of the pure time delay control object, joins PID neural network multi-step prediction link, and realize the PID neural network predictive control of system with large delay.

2   PIDNN Structure and Its Control System

2.1   PIDNN Structure

PIDNN is 3-layer feedforward dynamic network including input layer, implicit layer and output layer. The structure is shown in Fig.1.
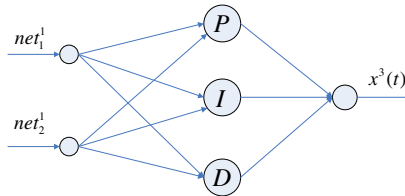
Fig. 1. PIDNN Structure

The forward algorithm of this network is used to realize the PIDNN control law, and the reverse algorithm is used to realize the adaptive adjustment of the PID2NN parameter. The nature of PIDNN is that it defines neurons with the function of proportion, integral and differential, which combines PID control law into the neural network. The number of neurons of each layer, the manner of connection and connection weight initial value of PIDNN are determined according to the basic principle of PID control law. The basic form of single variable PID neural network (SPIDNN) is the structure of $2 \times 3 \times 1$, a total of 6 neurons.

2.2   The Structure of PIDNN Predictive Control System

PIDNN Adaptive prediction controller is based on PIDNN. It's a combination of PIDNN identification (PIDNNI), PIDNN recursive multi-step prediction algorithm (PIDNNP), and PIDNN controller (PIDNNC). The structure of the system is shown in Fig.2.

The principle of PIDNN prediction control system is:

Firstly, identify the controlled object model by using the approximation and learning ability of nonlinear function of PIDNN neural network. Make $u(k-d-1)$ and $y(k-1)$ as the input of PIDNNI, and $y_I(k)$ the output. Use $e_I(k) = y(k) - y_I(k)$ to correct the weights of the network, which makes it adapt the feature of the identification object.
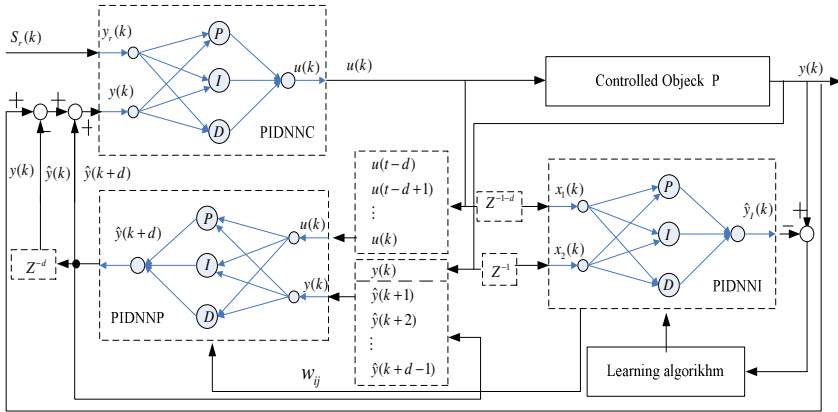
Fig. 2. The structure of PIDNN prediction control system

Secondly, when the identifier and the controlled object are basically the same, transfer the weights to PIDNNP, using recursion algorithm to predict the d-shot leading output value of the system to eliminate the influence of time delay to the output response of the controlled system.

At last, the predicted values and the set value as inputs, PIDNNC adjusts its own weights in real-time through the back propagation algorithm, and achieve the purpose of control. For time-varying system, PIDNNI can learn continuously to correct the weight value, follow the changes of controlled object, transfer weight value to PIDNNP and improve the accuracy of prediction output so as to eliminate the influence of the time-varying to the output response of the controlled system.

3   PIDNN Learning Algorithm

3.1   PIDNNI Identifying Algorithm

PIDNNI is constructed by PID neural network and two time delay links($z^{-1-\tau}$, z^{-1}), which is shown in Fig.2. The steady state input and output characteristics of proportion, integral and differential neurons in PIDNN are all generalized Sigmoid function, so it has the ability of arbitrary continuous function approximation. Seen from the outward structure of PIDNN, the number of neurons in implicit layer is too small, but in fact, one of the integral neurons equivalent k general neurons, the input of these general neurons are the sum of the output of the input layer neuron at time $1, 2 \dots k$ [7].

The input and output of PIDNN input layer is:

$$x_1^1(k) = p_1^1(k) = u(k-1-\tau)$$
$$x_2^1(k) = p_2^1(k) = y(k-1)$$
(1)

The input of the ^{i}th neuron in implicit layer is:

$$p_i^2(k) = \sum_{j=1}^{2} w_{ij}^1 x_j^1(k), i = 1, 2, 3 \tag{2}$$

Where, w_{ij}^1 is the weight of the j th joint of input layer to the i th neuron of implicit layer. The output functions of implicit layer are not the same; they are proportion, integral and differential respectively. Considering the actual control system, the controller outputs are limited value, so the upper and lower amplitude of the output are limited. Their definitions are:

The proportion neuron:

$$x_1^2(k) = \begin{cases} p_1^2(k) & \left| x_1^2(k) \right| \le 1 \\ sign(x_1^2(k)) & \left| x_1^2(k) \right| > 1 \end{cases} \tag{3}$$

The integral neuron:

$$x_2^2(k) = \begin{cases} p_2^2(k) + x_2^2(k-1) & \left| x_2^2(k) \right| \le 1 \\ sign(x_2^2(k)) & \left| x_2^2(k) \right| > 1 \end{cases} \tag{4}$$

The differential neuron:

$$x_3^2(k) = \begin{cases} p_3^2(k) + p_3^2(k-1) & \left| x_3^2(k) \right| \le 1 \\ sign(x_3^2(k)) & \left| x_3^2(k) \right| > 1 \end{cases} \tag{5}$$

The input, output of output layer neuron of PIDNN is:

$$x^3(k) = p^3(k) = \sum_{i=1}^{3} w_i^2 x_i^2(k) \tag{6}$$

The output of identifier of PIDNNI is:

$$y_I(k) = x^3(k) \tag{7}$$

The performance index function is:

$$E_I(k) = \frac{1}{2}[y(k) - y_I(k)]^2 = \frac{1}{2}e_I^2(k) \tag{8}$$

In accordance with the gradient descent method correction network weights, iteration equations of each layer weights are:

$$W(k+1) = W(k) - \eta \frac{\partial E}{\partial W} \tag{9}$$

After training k steps, adjust the algorithm from implicit layer to output layer:

$$\frac{\partial E}{\partial w_i^2} = \frac{\partial E}{\partial x^3} \frac{\partial x^3}{\partial p_i^2} \frac{\partial p_i^2}{\partial w_i^2} = -(y(k) - y(k-1))x_i^2 \tag{10}$$

Make $\delta^2 = (y(k) - y(k-1))$

$$w_i^2(t+1) = w_i^2(t) + \eta_2 \delta^2 x_i^2 \tag{11}$$

η_2 is the learning step length of W^2.

After learning k steps, adjust the algorithm from input layer to implicit layer:

$$\frac{\partial E}{\partial w_{ij}^1} = \frac{\partial E}{\partial x^3}\frac{\partial x^3}{\partial p^3}\frac{\partial p^3}{\partial x_i^2}\frac{\partial x_i^2}{\partial p_i^2}\frac{\partial p_i^2}{\partial w_{ij}^1} = -\delta^2 w_i^2 \, \mathrm{sgn}\frac{x_i^2(k)-x_i^2(k-1)}{p_i^2(k)-p_i^2(k-1)}x_j^1 \tag{12}$$

Make $\delta_i^1 = \delta^2 w_i^2 \, \mathrm{sgn}\dfrac{x_i^2(k)-x_i^2(k-1)}{p_i^2(k)-p_i^2(k-1)}$

$$w_{ij}^1(k+1) = w_{ij}^1(k) + \eta_1\delta_i^1(k)x_j^1 \tag{13}$$

η_1 is the learning step length of W^1.

After multiple step identification learning, the output of the network will approximates the controlled object output $y(k)$, and obtains the network weights. The weights will be passed to the PIDNNP prediction model.

3.2   PIDNNP Predictive Algorithm

For the time delay system, considering partial future controlled value in the multi step predictive model does not affect the part of future output value, based on the single step model, construct the recursive algorithm of multi step prediction model [10,11], which is:

$$\begin{aligned}
y_{pre}(k+1) &= f(u(k-d), y(k)) \\
y_{pre}(k+2) &= f(u(k+1-d), y_{pre}(k+1)) \\
&\cdots\cdots \\
y_{pre}(k+d) &= f(u(k-1), y_{pre}(k-1+d))
\end{aligned} \tag{14}$$

Where, $y_{pre}(k+i), (i = 1,2\cdots delay)$ is the output value of the multi step prediction model, and $f(\bullet)$ is forward calculation nonlinear function of PIDNN. The structure of recursive multi step model is shown in Fig.3.

The above prediction model is established on the basis of PIDNNI off-line identification and on-line learning. If the single step predictive model mismatches and if there's disturb in the system, the output of the multi step prediction model may be a cumulative error, which will lead to that the predictive output $y_{pre}(k+d)$ deviates from actual output $y(k+d)$. So it is necessary to carry out on-line correction to improve the accuracy of prediction. Direct correction is used here, after correction, the predictive value of the system is:

$$y_{Mpre}(k+d) = y_{pre}(k+d) + [y(k) - y_{pre}(k)] \tag{15}$$

Where: $y_{Mpre}(k+d)$ is the multi step predictive value of the system after correcting, $y_{pre}(k)$, $y(k)$ is the output value of the process object.
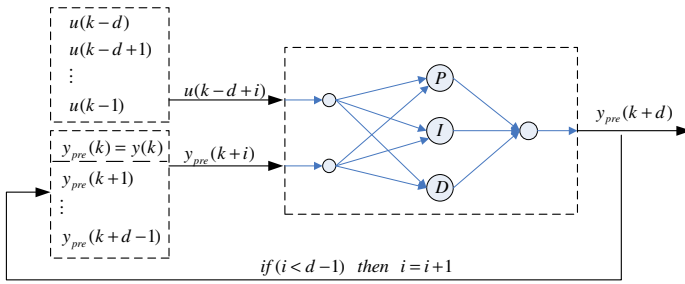
Fig. 3. The structure of recursive multi step model based on PIDNNP

3.3   PIDNNC Controller

The calculating process of PIDNNC is basically same with PIDNNI. The algorithm consists of the forward algorithm and the error back propagation algorithm. The difference is:

The inputs, outputs of the input layer neuron are:

$$x_1^1(k) = p_1^1(k) = r(k)$$
$$x_2^1(k) = p_2^1(k) = y_{Mpre}(k+\text{d}) \tag{16}$$

The input, output of the output layer neuron is:

$$x^3(k) = p^3(k) = \sum_{i=1}^{3} w_i^2 x_i^2(k) \tag{17}$$

PIDNNC outputs:

$$u(k) = x^3(k) \tag{18}$$

The performance index function is:

$$E_s(k) = \frac{1}{2}[r(k) - y_{Mpre}(k+\text{d})]^2 = \frac{1}{2}e_s^2(k) \tag{19}$$

Weights change from implicit layer to the output layer is:

$$\frac{\partial E_s}{\partial w_i^2} = \frac{\partial E_s}{\partial y_{Mpre}}\frac{\partial y_{Mpre}}{\partial u}\frac{\partial u}{\partial x^3}\frac{\partial x^3}{\partial p_i^2}\frac{\partial p_i^2}{\partial w_i^2} = -e_s \times \frac{\partial y_{Mpre}}{\partial u} \times x_i^2 \tag{20}$$

$$\frac{\partial y_{Mpre}}{\partial u} = \begin{cases} \dfrac{y_{Mpre}(k+\text{d}) - y_{Mpre}(k+\text{d}-1)}{u(k) - u(k-1)}, & \left|\dfrac{\partial y_{Mpre}}{\partial u}\right| \le 1 \\ sign(\dfrac{\partial y_{Mpre}}{\partial u}) & \left|\dfrac{\partial y_{Mpre}}{\partial u}\right| > 1 \end{cases} \tag{21}$$

4   Simulation Research

Choose the electric heating furnace which is commonly used in industrial process as the simulation controlled object, and the electric heating furnace is a nonlinear, time delay, big inertia, time variation and of unidirectional temperature rising control object. In order to compare the performance difference between PIDNN predictive controller and the conventional PID controller, the paper chooses the electric heating furnace model as one order pure delay object, as shown:

$$G_p(s) = \frac{K}{Ts+1}e^{-\tau s} \tag{22}$$

Here, T is the inertia time constant, K is the gain coefficient, and τ is the delay time constant.

To identify the controlled object model, let $K = 10$, $T = 60$, $\tau = 70$, $T_s = 5$, using sine exciting signal $U(K) = Sin(k/100)$.
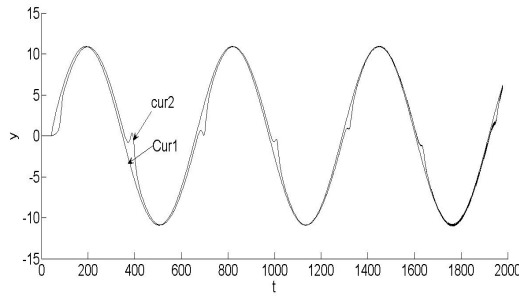
Fig. 4. PIDNNI The output identification curve of the system
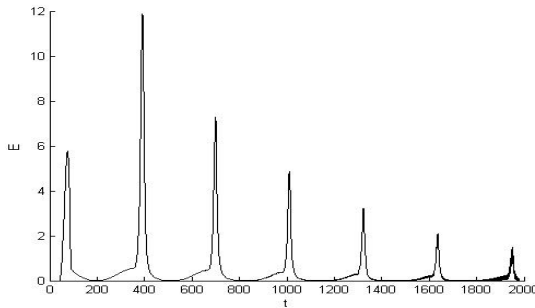
Fig. 5. Variance curve of the identification of PIDNNI system

The identification curve of the system is shown in Fig.4. curve 1 is the output curve of the object. Curve 2 is the identification curve. In Fig.5, E is the variance of the identification error.

The matching model indicates that Smith prediction model and PIDNNII off-line identification model are uniform with the controlled object model ($k = 10$, $T = 60$, $\tau = 70$), the simulating curve is shown in Fig. 6.
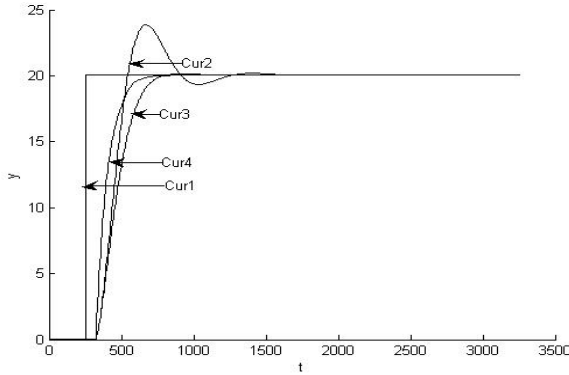
Fig. 6. The step response curve when model match

Otherwise, the mismatching model indicates that Smith prediction model and PIDNNII off-line identification model ($k = 11$, $T = 70$, $\tau = 90$) have certain deviation with the controlled object model ($k = 10$, $T = 60$, $\tau = 70$), the simulating curve is shown in Fig .7.
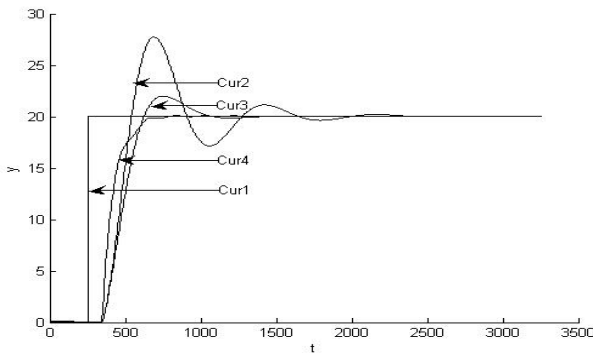
Fig. 7. The step response curve when model mismatch

In Fig.6 and Fig.7, curve 1 is the system setting value, curve 2 is the step response of the system with conventional PID controller, curve 3 with PID-Smith controller, and curve 4 with the PIDNN predictive controller (PIDNN-PC). The curves show that systems with PID-Smith and PIDNN-PC both get smooth system output response. Table1 indicates the overshoot, rise and setting time when system adopts these three control methods respectively. And we can see controlled system with PIDNN-PC controller has a smaller overshoot and a shorter setting time.

Table 1. Performance indices of step response of the system

Control Model		Max overshoot (%)	Rise time(s) (5～95%)	Setting time(s) (98～102%)
Model match	PID	19	185	910
	PID-Smith	0	320	475
	PIDNN-PC	0	230	370
Model not match	PID	38	170	1570
	PID-Smith	10	235	730
	PIDNN-PC	0	255	390

5   Conclusion

Fuse PID control law into neural network, PIDNN realizes the essentially combination of neural network and PID control. Having introduced proportion, integral and differential 3 units into implicit layer, it has become both a static and dynamic characteristics of multilayer feed forward neural network. The identifier constructed by PIDNNI is simple in structure and has fast identification speed. The controller based on PIDNNC can achieve good control effect based on the parameters of conventional PID controller as the network initial weights. The control algorithm is simple and easy to realize. Predicting the leading output of the system by recursive multi-step prediction algorithm makes the controller act early, which improves the large time delay system control quality significantly, and has good adaptability and robustness. Simulation results show that, this algorithm has a good control effect for time delay system.

References

1. Wen-zhan, D., Hai-chuan, L., Ai-ping, Y.: An overview of neural network predictive control for nonlinear systems. Control Theory & Applications 5, 521–529 (2009)
2. Xu, Y., Xuyongmao: Single-neuron predictive control for time-variable large delay systems. Journal of Tsinghua University 5, 383–385 (2002)
3. Shu, H., Youguo: PID neural networks for time delay systems. Computers and Chemical Engineering 28, 859–862 (2000)
4. Wen, D.: Research on the RBF neural network based intelligent control for improving a control system with time delay. Industrial Instrumentation & Automation 2, 31–34 (2008)
5. Shu, H.: PID neural network for decoupleing control of strong coupling multivariable time-delay systems. Control Theory and Applications 15, 920–924 (1998)
6. Shen, Y., Gu, X.: Identification and Control in Nonlinear System Based on PID Neural Network. Journal of East China University of Science and Technology 32, 860–863 (2006)
7. Shu, H.: PID neural network and control system. National Defence Industry Press, Beijing (2006)

8. Dong, W., Liu, C., Song, H.: Human perceive system of simulative aircraft based on PIDNN. Journal of Beijing University of Aeronautics and Astronautics 34(2), 153–157 (2008)
9. Shen, Y., Gu, X.: Identification and Control in Nonlinear System Based on PID Neural Network. Journal of East China University of Science and Technology 32(7), 860–863 (2006)
10. Lu, C.-H., Tsai, C.-C.: Predictive Control Using Recurrent Neural Networks for Industrial Processes. Journal of the Chinese Institute of Engineers 32(2), 1–9 (2009)
11. Li, H.-J., Xiao, B.: Multistep recurrent neural network model predictive controller without constraints. Control Theory & Applications 29(5), 642–648 (2012)

Design and Implementation of the Food Shelf-Life Monitoring System

Han-Fei Zhang and Ping Zhou

Huaiyin Normal University,
Chang Jiang Road, 111, Huaian, JiangSu, China
zhanghanfei2006@163.com,zhouping@hytc.edu.cn

Abstract. Due to the disadvantages of the traditional food shelf-life indicator and bar-code which cannot record the real time in the food cold-chain logistics, paper designs the food shelf-life monitoring system with the technology of embedded system、RFID and GPRS. This system contains the front monitor device, monitor terminal and monitor center. Having been debugged preliminary,the design of the system was tested successfully．the front monitor device can detect the food quality information; monitor terminal is able to send these data to the monitor center. This system could achieve real-time monitoring to food.

Keywords: Embedded system, RFID, GPRS, monitoring system.

1   Introduction

With the rapid development of our economy, people's living standard has been improved gradually and there are also higher requirements for food quality. Food quality is not only required to be safe, but its characteristics are also demanded to be almost perceptual. [1] But for plant food, animal food and artificial food, during raw materials removal, processing, logistics, warehousing, sales and other sectors, their water activity (AW), Total acidity, nutrients, natural microorganisms, enzymes, biochemical substrates and preservatives and other factors, will be affected by outside temperature, humidity, light and environmental microorganisms and the composition of the packaging gas. And these important factors go through physics, chemistry and microbiology change, suffer gradual loss of its original quality at a certain speed, in some manner. So many links such as food production and sales need to be monitored to make sure of food quality. However, the traditional methods simply set food production date and make use of bar code technology for food security and distribution management, which can't satisfy the more intensive and highly effective purpose of food safety management. [2-4]

   Therefore, the research on the monitoring of food safety will become a key research direction. Shanghai Ocean University 's professor Chen Ming and professor Feng Guofu have developed the aquatic shelf life indicator with the K-means clustering algorithm[5-8], which could monitor aquatic products in real time. Through

biochemical experiments contrasting, the data of the shelf life indicator has 98% fit with the biochemical data. The device has achieved the level of monitoring aquatic products in real time, but the monitoring data could not be sent to the managers, which could be managed to food in time. To solve this problem, based on the study of Professor Chen Ming' studying, paper designs a food-shelf-life monitoring system using RFID technology and GPRS technology [9-13].

2   Structure and Principle of System

Food-shelf-life monitoring system contains the front monitor device, monitor terminal and monitor center [14-19]. The front monitor device is responsible for collecting the information of food in the circulation, and computing the shelf-life through the K-means clustering algorithm; Monitor terminal is responsible for communicating with the front monitor device and monitor center; monitor center is responsible for receiving the data from the monitor terminal. The front monitor device uses a star network topology to communicate with the monitor terminal. They could be bidirectional communications, but it cannot communicate between the front monitor devices. The structure of food-shelf-life monitoring system is shown in Figure 1.
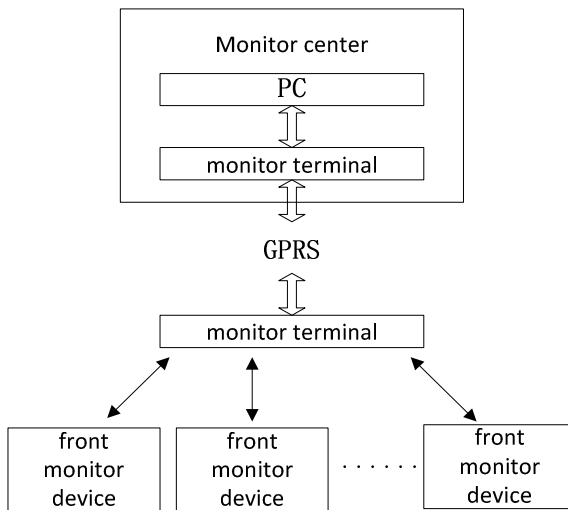
Fig. 1. Food shelf life monitoring system structure diagram

Firstly, the front monitor device detects the food quality information in the refrigerated, and then the information is sent to the monitor terminal, at last the information is sent to the monitor terminal of monitor center which communicates with PC by the RS232 interface.

3   Designing of System's Hardware

3.1   Designing of the Front Monitor Device's Hardware

The front monitor device is primarily responsible for collecting information of food quality and calculating the shelf-life,receiving monitor terminal' s instructions, and sending the data to monitor terminal via the nRF905 wireless RF module. The front monitor device contains the control module, RFID module (antenna, RF chip nRF905), temperature sensor module, user interface module (key module, clock module,debug interface JTAG), and the power management module [20-24]. RFID module is responsible for receiving commands from the monitor terminal and sending the date packet, temperature sensor module is responsible for collecting temperature of food. The control module manages the normal running of the device, the user interface module can be convenient for users, the power supply management module is used for the control module and RFID module. The front monitor device's structure diagram is shown in figure 2, The MSP430F149 pin distribution diagram of the front monitor device is shown in figure 3.
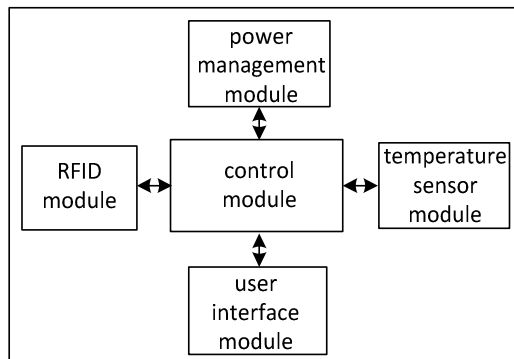
Fig. 2. The front monitor device's structure diagram

The microprocessor of the front monitor device is using MSP430F149, which comes with 60KB+256B Flash Memory. The code of system is stored in the 1100H-FFDFH Flash Memory address. The front monitor device's signal classifies into the analog portion and digital portion.the analog portion is responsible for communicating with monitor terminal, and the digital portion is responsible for processing data, controlling device for running and communicating with monitor terminal. The analog portion contains RF transceiver chip nRF905, PCB onboard antenna, crystal and related circuit. The digital portion contains an embedded microprocessor MSP430F149, temperature sensor DS18B20, calendar clock chip DS1302, debug interface JTAG, red and green lights, switch key, reset key and other peripheral circuit. RF chip nRF905 has four work modes, the normal working state is at the receiving state (ShockBurst RX model).

The microprocessor sends information to the NRF905 RF module through serial peripheral interface (SPI), and then sends to the monitor terminal.
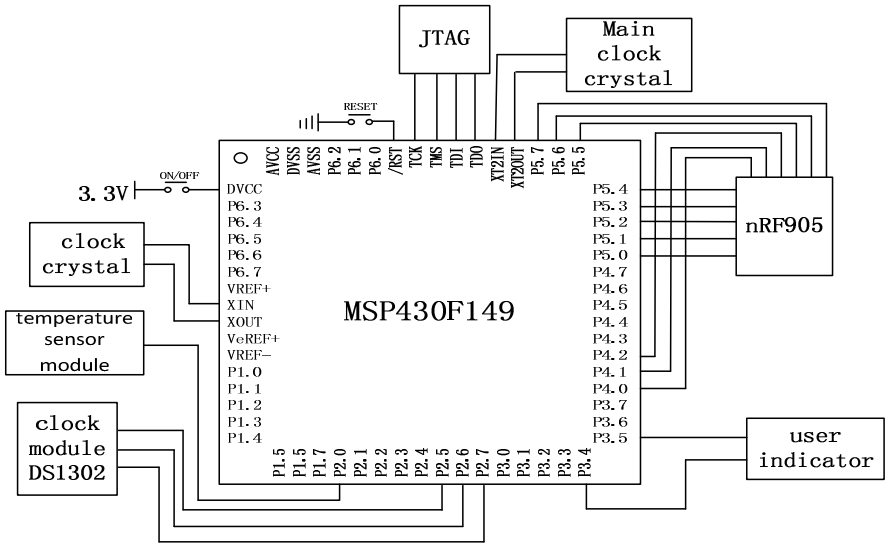
Fig. 3. The MSP430F149 pin distribution diagram of the front monitor device
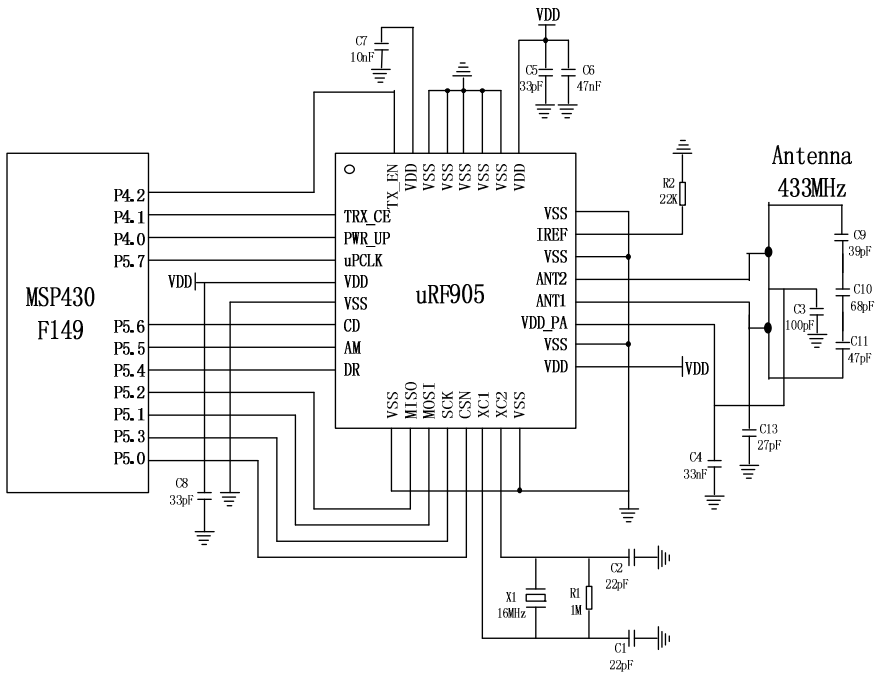
Fig. 4. The RF circuit of the front monitor device

RF circuit is mainly composed of three parts: connected with the single-chip interface circuits, nRF905 application circuit and antenna circuit. The power supply voltage of VDD pin range of 1.9-3.6V. Figure 4 is the RF circuit of the front monitor device.

3.2   Designing of the Monitor Terminal's Hardware

The task of the monitor terminal is to send information to the monitor center through GPRS, its structure diagram is shown in figure 5.
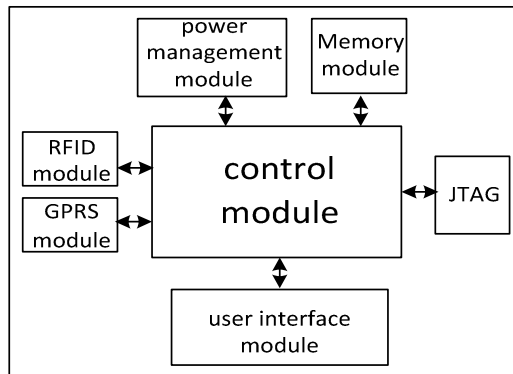
Fig. 5. The monitor terminal's structure diagram

The microprocessor of monitor terminal nRF905 RF module through the SPI, and communicates with GPRS module through the serial port.

4   Designing of System's Software

4.1   Designing of the Front Monitor Device's Software

In the cold-chain logistics, the front monitor device is required to real-time monitoring food quality. When collecting the data, the front monitor device will enter and be awakened up to monitor whether there is the monitor terminal's RF signal at regular intervals. If the front monitor device monitored signal of the monitor terminal, the front monitor device communicates with the monitor terminal. However, the front monitor device enters a state of dormancy again. The temperature sensor module is awakened up to collection the data every 30 minutes, and then goes into a state of dormancy. Figure 6 is the working flow of the front monitor device.
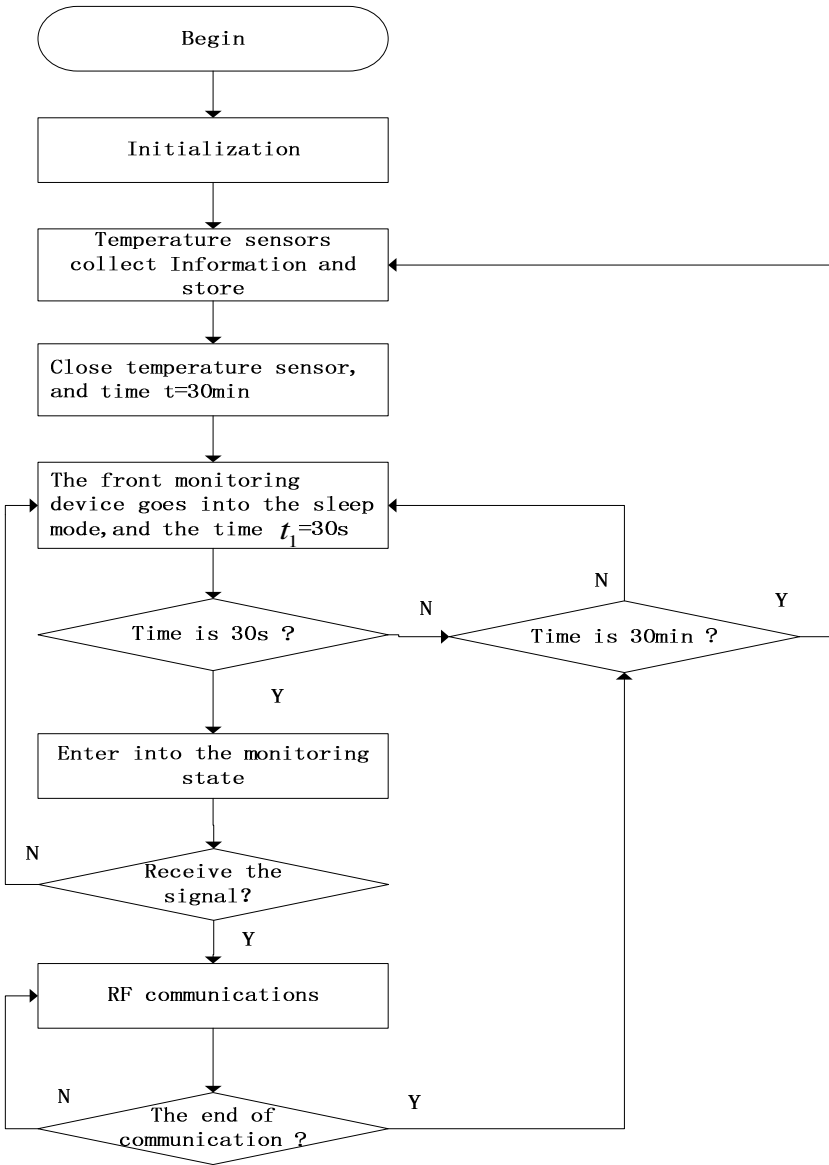
Fig. 6. The working flow of the front monitor device

4.2   Designing of the Monitor Terminal's Software

The monitor terminal needs to read the information of the front monitor device every 5 minutes. The monitor terminal transmits RF signals, and then the front monitor device is awakened up. They establish data communication. Figure 7 is working flow of the monitor terminal.
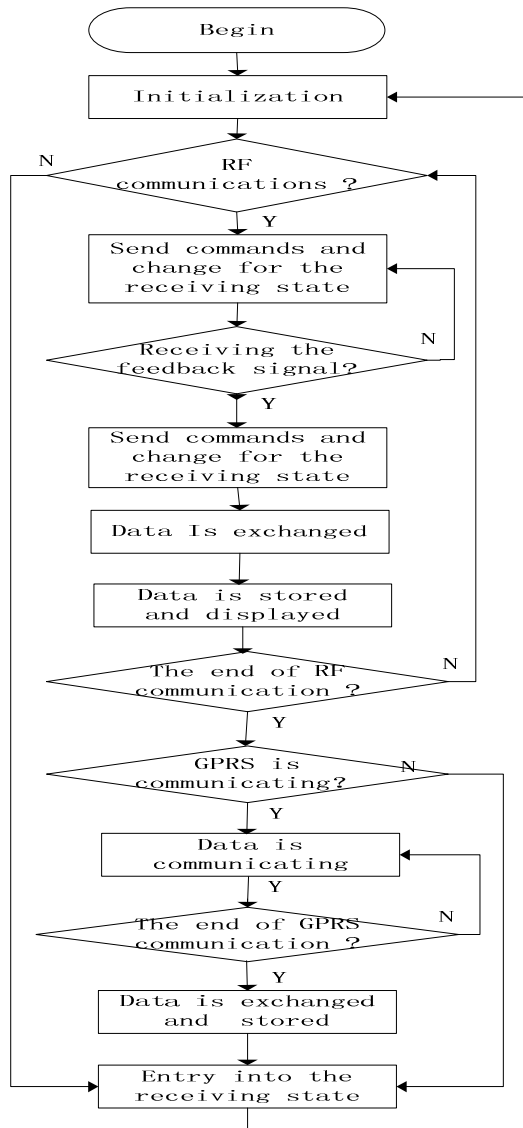
Fig. 7. Working flow of the monitor terminal

5   Test of System

5.1   Test of the Communication Distance

According to the above research, the paper designs the front monitor device and the monitor terminal, figure 8 is the front monitor device, figure 9 is the monitor terminal.

Fig. 8. The front monitor device

Fig. 9. The monitor terminal

Experiments were carried out in indoor, outdoor environment, indoor test is carried out in the office building, outdoor test is carried out in front of the office building. Table 1 is the results of the front monitor device and the monitor terminal's indoor testing,table 2 is the results of the front monitor device and the monitor terminal's outdoor testing.

Table 1. The test of recognition distance

Distance	The recognition results
10 meter	communication
20 meter	communication
30 meter	communication
35 meter	communication
36 meter	No communication

In the distance of 30 meters, the monitor terminal can receive signals of the front monitor device; in the distance of 35 meters, the monitor terminal receives signals unstabitily, more than this distance,the monitor terminal cannot receive the front monitor device's signals.

Table 2. The test of recognition distance

Distance	The recognition results
50 meter	communication
100 meter	communication
150 meter	communication
200 meter	communication
250 meter	communication
260 meter	No communication

in the distance of 150 meters, the monitor terminal can stably communicate with the front monitor device ; between the 200 meters to 250 meters range, the monitor terminal cannot stable receive the front monitor device's signals, more than 260 meters, the monitor terminal completely can not receive the signals sent by the front monitor device.The communication distance of the the front monitor device and the monitor terminal in the outdoor is more than in the door.

5.2   Test of Monitoring Data

In the monitor center, the monitor terminal is connected to the computer through the RS232 serial port, so that the data can be transmitted directly to the PC, table 3 is information that is sent by the front monitor device

Table 3. Monitoring data

Number	Date
A1	10.13℃
A2	11.15℃
A3	11.21℃
A4	10.64℃
A5	10.54℃
A6	10.09℃
A7	10.87℃

In the monitoring data, monitor terminal receives the front monitor device's ID number, and then temperature. Food-shelf-life monitoring system could monitor the food information in the circulation through experimental test.

The front monitor devices send the information to the monitor terminal through RF In the dioxide refrigerated lorry, and then the information is sent to the monitor terminal of the monitor center through the GPRS, and displayed to the user. This information has not missed and lost in the process of transmission, and the monitoring system could achieve to monitor the food in real time in the cold-chain logistics.

6   Conclusion

Food quality is not only required to be safe, but its characteristics are also demanded to be almost perceptual. But for plant food, animal food and artificial food, during raw materials removal, processing, logistics, warehousing, sales and other sectors, their water activity (AW), Total acidity, nutrients, natural microorganisms, enzymes, biochemical substrates and preservatives and other factors, will be affected by outside temperature, humidity, light and environmental microorganisms and the composition of the packaging gas. And these important factors go through physics, chemistry and microbiology change, suffer gradual loss of its original quality at a certain speed, in some manner. So many links such as food production and sales need to be monitored to make sure of food quality. However, the traditional methods simply set food production date and make use of bar code technology for food security and distribution management, which can't satisfy the more intensive and highly effective purpose of food safety management. Therefore, the research on the monitoring of food safety will become a key research direction. Food-shelf-life monitoring system could achieve the real-time to monitor food quality by the technology of RFID and GPRS. Monitoring system has the advantages of high reliability and adaptability, and can reduce perishables because of the harsh environment in the transportation. So it has the great valuable of application.

References

1. Ma, L., Nan, Q., Dai, R.: The influence to physicochemical and sensory characteristics of chilled pork in cold storage on the different color packaging. Chinese Society of Agricultural Engineering 19(3), 156–160 (2003)
2. Yuan, W.: The application of barcode technology in the logistics management. Degree Thesis of Huazhong University of Science and Technology (2005)
3. Wang, H., Xu, X.: Reserch peogress of traceability technology of livestock and poultry and its applications. Science and Technology of Food Industry 31(8), 413–416 (2010)
4. Ma, C., Zhao, D., Liu, Y., et al.: Overall monitoring of pork industrialization production and development of traceability system. Transaction of the CSAE 24(9), 121–125 (2008)
5. Liu, H., Chen, M., Xie, J.: Intelligent Shelf-life Prediction Device of Peneaus Vannamei. Computer Engineering 36(15), 277–279 (2010)

6. Ren, S., Xu, H., Li, A., et al.: Meat-productions tracking and traceability system based on internet of things with RFID and GIS. Transactions of the CSAE 26(10), 229–235 (2010)
7. Stanford, K., Stitt, J., Kellar, J.A., et al.: Traceability in earle and smallruminants in Canada. Rev. Sci. Tech. of Int. Epiz. 20(2), 510–522 (2001)
8. Xue, L., Wei, L., Zhu, S., Yu, C.: Design of door access control system based on GPRS and RFID. Computer Technology and Its Applications 38(6), 145–148 (2012)
9. Wang, Z., Hao, X., Wei, D.: Remote water quality monitoring system based on WSN and GPRS. Instrument Technique and Sensor 1(1), 48–50 (2010)
10. Chen, T., Tang, H.: Remote monitoring and forecasting system of soil moisture based on ARM and GPRS. Transactions of the CSAE 28(3), 162–166 (2012)
11. Li, Y.-X., Xu, J.-Z., Liu, A.-B.: Application of GPRS technology in automatic meter reading system. Electric Power Automation Equipment 23(12), 52–54 (2003)
12. Sun, Z., Cao, H., Li, H.: GPRS and WEB based data acquisition system for greenhouse environment. Transactions of the CSAE 22(6), 131–134 (2006)
13. Deng, X., Li, M., Wu, J., Che, Y., Zheng, L.: Development of mobile soil moisture monitoring system integrated with
14. Lipton, A., Fujiyoshi, H., Patil, R.: Moving target classification and tracking from real-time video. In: Proc. 4th IEEE Workshop on Applications of Computer Vision (WACV 1998), pp. 8–14 (1998)
15. Masaki, I.: Machine-vision systems for intelligent transportation systems. IEEE Intell. Syst. 13(6), 24–31 (1998)
16. Michalopoulos, P.: Vehicle detection video through image processing: The Autoscope system. IEEE Trans. Veh. Technol. 40, 21–29 (1991)
17. Murphey, Y.L., Lu, H., Lakshmanan, S., Karlsen, R., Gerhart, G., Meitzler, T.: Dyta: An intelligent system for moving target detection. In: Proc. 10th Int. Conf. Image Analysis and Processing, pp. 1116–1121 (1999)
18. Nakanishi, T., Ishii, K.: Automatic vehicle image extraction based on spatio-temporal image analysis. In: Proc. 11th Int. Conf. Pattern Recognition, pp. 500–504 (1992)
19. VideoTrak®,-905 datasheets, report (1999)
20. MOTION–Traffic-actuated signal control for traffic computer system (Special Edition of Blickpunkt ANL), report (1997)
21. Smith, C.E., Richards, C.A., Brandt, S.A., Papanikolopoulos, N.P.: Visual tracking for intelligent vehicle-highway systems. IEEE Trans. Veh. Technol. 45, 744–759 (1996)
22. Taktak, R., Gresson, R., Dufaut, M., Husson, R.: Vehicle detection for automatic traffic control during the whole day. In: Proc. 1st IFAC-Workshop on Advances in Automotive Control, pp. 215–220 (1995)
23. Tomasi, C., Kanade, T.: Detection and tracking of point features, Carnegie Mellon Univ. (1991)
24. Yoshinari, K., Michihito, M.: A human motion estimation method using 3-successive video frames. In: Proc. Int. Conf. Virtual Systems and Multimedia, pp. 135–140 (1996)

Theoretical Framework of Technical Kinematics Evaluation Software

Xu Xiaofeng, Di Jianyong, Jing Lixian, and Gao Yansong

Sports Department, Hebei United University
xxf05106@sina.com

Abstract. The system, which takes hi-tech Olympics and improving national sports science and technology as well as training levels of athletes as the background, and considers current three-dimensional video analysis system for diagnosis of sports technology as the platform, constructs data and knowledge base based on three-dimensional information and parameters by mainly utilizing modern computer technology. It analyzes human movement behaviors with systematic conceptions, and studies technical movements of athletes through input-output relationship of human movement behaviors. Through fuzzy mathematics, mathematical statistics and artificial intelligence, integrating and drawing reference from sport biomechanics, graphics, human anatomy and expert system etc., the system solves a series of key problems both in theory and application from information extraction, inference to multi-objective parameter fusion and decision-making, and then designs and develops biomechanical evaluation system for shot sport technology. Considering the multiple events that the sport contains, the research primarily proceeds from shot, and establishes "coach-athlete" expert evaluation system.

Keywords: Shot, Evaluation System, Computer.

1 Introduction

In the field of sports science and technology, emulation technique application is rarely seen due to the difficulty in establishing behavior model of human beings. With the development of modern mathematics, it becomes possible to calculate emulation of human behaviors in models [1-4]. To name a few of the algorithms, fuzzy-neural network and genetic algorithm etc., which are well verified in the area of artificial intelligence; on the other hand, the object-oriented visual programming technology provides favorable tools in the aspect of developing friendly human-computer interfaces, all of which have offered theoretical and technical guarantee to solving problems in sports area with the system simulation technology [5].

The research, which takes hi-tech Olympics and improving national sports science and technology as well as training levels of athletes as the background, and considers current three-dimensional video analysis system for diagnosis of sports technology as the platform, proceeding from shot sports, constructs data and knowledge base based on three-dimensional information and parameters [6-8]. It analyzes human movement

behaviors with systematic conceptions, and studies technical movements of athletes through input-output relationship of human movement behaviors [9]. Through fuzzy mathematics, mathematical statistics and artificial intelligence, integrating and drawing reference from sport biomechanics, graphics, human anatomy and expert system etc., the study solves the problem of coaches determining technology of athletes just by self-experience, researches and develops the biomechanical evaluation system for shot sport technology [10-13].

2   Research Methods

2.1   Main Thought

In order to realize the goal mentioned above, the method of computer software engineering is applied to design on the computer a biomechanical evaluation system for shot sport technology featuring strong practicability, easy operation, clear picture, accurate and reliable information processing, and high speed of service, which also marks main thought of the topic.

2.2   Research Methods

2.2.1   The Literature Material Law

Through retrieving the website: www.cnki.net and data base of outstanding theses and dissertations, a large number of relevant literatures, theses and books are checked and read regarding sports information science, shot, and computer evaluation system etc. [14-15]

2.2.2   Experimentation

Collect sport technique images of human, and then conduct three-dimensional analyses so as to obtain sport parameters such as angle of the moment of throwing in the early phase, initial velocity and vertical height between the shot and the ground when throwing occurs etc.

Measure indicators of human body morphology, function and quality of athletes with isokinetic testing system, three-dimensional force measurement system and back force equipments; to name a few, absolute force of legs and arms of athletes.

2.2.3   Expert Interview

Predominant objective of the interview is determining indicators influencing technical factors of shot sports, indexes reflecting current situation of athletes as well as those used to evaluate professional ability and technical ability of the athletes.

2.2.4   Method of Software Engineering

System thinking of software engineering runs through design and development of the system, which follows theory of life circle of soft wares, as well as the thought of

problem definition, feasibility study, demand analysis, overall designing, detailed designing and coding etc.

The software system is designed and developed with tools of object-oriented visual programming language: Visual Basic 6.0, C++, image manipulation software: OpenGL and statistical software: Matlab etc.

2.2.5   Mathematical Modeling

According to primary factors influencing sport performance gained through expert interview, decision model can be established with genetic algorithm, kinematics indicators obtained by image analyses and measured by equipments.

3   System Analyses

3.1   Demand Analyses

The system, which obtains kinematics parameters by gathering technical images and conducting three-dimensional analyses, evaluates sports techniques through finding out movement laws in a rapid and accurate manner by drawing reference from analyses and processing of parameters, providing theoretical foundation for improving technology of athletes [16-19].

3.2   System Construction

Figure 1 demonstrates construction process and application of the system according to hierarchy principles.
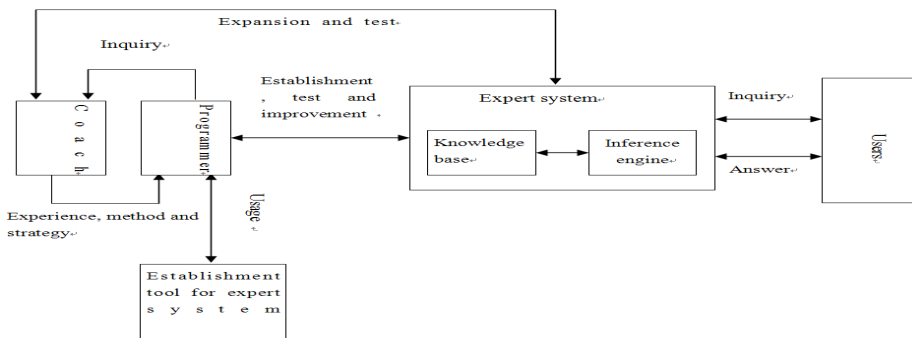
Fig. 1. Construction process and application of the system

Expert system marks the most active and effective research area in artificial intelligence. Expert system, a computer knowledge system based on knowledge, obtains knowledge from experts in the human sphere, and applies it to solve problems that can be tackled by experts. Therefore, expert system can be defined as a program system that possesses abundant knowledge and experience in certain areas. With

artificial intelligence technology, it conducts inference and judgment according to knowledge and experience offered by one or multiple experts, simulates thinking process of solving problems of the experts as so to solve diverse problems of the area.

3.3 Module Features of the System

Figure 2 demonstrates module features of the system
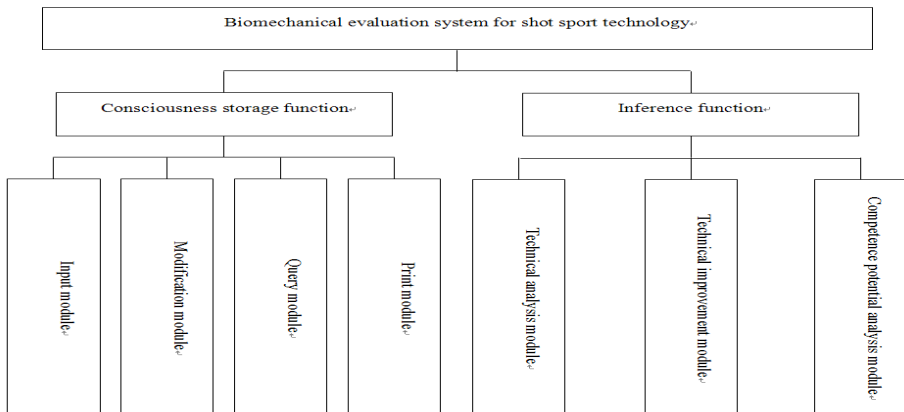
Fig. 2. Diagram of basic property structure of the system

3.3.1 Consciousness Storage Function

Input module
Main function of the module is that coaches can directly input basic information of the athletes and some kinematics parameters.

Modification module
Primary function of the module is to modify basic information of the athletes and some kinematics parameters, including adding and deleting characters.

Query module
Main function of the module is enabling coaches to search for information.

Print module
Predominant function of the module is printing materials needed by coaches.

3.3.2 Inference Function
The function primarily encompasses technical analysis module, technical improvement module and competence potential analysis module.
    At the higher level, athlete training requires high pertinence, accurate quantification, reasonable training rhythm and heavy load. As the blueprint in engineering projects, the

scientific training plan, which serves as pre-theoretical design of future training content for athletes, expresses interrelation as well as variation laws for abundant information in the training process, and counts scientific basis for athletes to shift from actual state to objective status. By means of analytical formation and management of training plan with computer and IPT, scientific analytical calculation can be performed on actual state and competence potential of athletes, their training status can be compared in an accurate manner, thus providing scientific predictions.

Technical analysis module
Function of the module is analyzing technical movement of athletes. Refer to Table 1 for details.

Table 1. Assessment criterions for shot technology

	Technical specifications	Requirements on the correct shot put technology
Trajectory of shot	(Run up)Linearity of sliding step Height between shot and point of support prior to sliding step Strength exertion distance of the shot	Linearity on the vertical plane Trajectory on the horizontal plane 80—100 cm above the point of support 160~180 cm
Track of center of body weight for athletes	Center of body weight constantly shifts in both horizontal and vertical directions Parabolic path for center of body weight and the shot	No pause for center of body weight prior to pushing the shot Consistent angle between center of body weight, shot track direction and the ground
Dynamics characteristics of articular movement	Evenly straighten all joints	15~20° from this phase to another stage
Shot speed variation during shot put	Uniformly speed up, exert high initial velocity for final force	An acceleration of 3.5-3.7 m/second is maintained before pushing the shot; At the moment of throwing, stretch the left foot to the ground, straighten left leg and arm for throwing, and the shot falls behind in a long distance at a lower location.

Table 1. (*Continued.*)

Two-leg movement with final force	Explosive exertion in final force	Fully stretch the legs, and be consistent with the time of throwing the shot
Action of the right leg with final force	Impulse of high force opposite to the direction of shot put	Rotate the legs—Forward and initiative action
Action of the left leg with final force	Impulse of high force opposite to the direction of pushing the shot out	Prompt and positive braking action of the left leg
Body action with final force	Body perform whipped action Tightening angle between pelvic axis and upper limb	The body rotates and stretches forwards by a wide margin, and exceeds tightening as much as possible
Action of the right arm with final force	Fully stretch the right arm when pushing the shot out	Ultimately, push the shot from the finger, the right arm turns to the pushing direction, and then fully straighten elbow joint

Technical improvement module
Through complying with the aforementioned development trend and characteristics of shot put event, and combining kinematic indexes obtained by image analysis as well as the measured dynamics indictors, MATLAB, the statistical software is taken advantage of to process data, and then genetic algorithm is applied to perform modeling on athletes, thus improving their techniques.

Competence potential analysis module
Analysis on competence potential of athletes is conducted through technical improvement.

4   Software Designing and Development

Content of the biomechanical evaluation system for shot sport technology can be divided into the seven parts below according to its work process and function of each part: image collection, playing of the sport technique images, joint extraction, processing of image data, result output, analysis and evaluation of sport biomechanics and function of assistance.

4.1   Image Collection

The camera is connected to video in of built-in image acquisition card of the computer (CPE-3000). Useful part needs to be collected to internal storage and hard wares of the computer.

4.1.1   Image Card

Image card employed in the system, which is researched and developed by Institute of Automation, Chinese Academy of Sciences, features high performance, high fidelity, and the two systems of PCI bus true color as well as black and white (CPE—3000) .

4.1.2   Preprocessing of the Images

Under VC development environment, OpenGL technology as well as editing, modifying and video compositing software that can virtualize human movement in sport analysis system is utilized to transfer by frames the sport video input into computer (In the format of .avi) into sport image sequence.

4.2   Play the Sport Technique Images

Just like image collection, playing of sport technique pictures is realized through calling DLL as well as Windows API function in the development base offered by image acquisition card driver. The speed of playing can be controlled, for instance, normal playing, backward playing, speed up, slow down, play by frames, and pause; meanwhile, the following functions can be added: ① Local movement of certain links can be amplified; ② Two windows can be set to perform contrast playing of the two images.

4.3   Identification of Joints

Computer automatic recognition of identification point of framework in the images or human joints and artificial cognition are combined by making use of the motion estimation algorithm based on block matching against the VC++ development environment.

4.4   Process Image Data

This part mainly includes DLT calculation, kinematics parameters as well as smooth processing of data.

Point coordinate data of the image are obtained through identifying pixel on computer screen. Generally speaking, top left corner is taken as the coordinate system's origin. If the display mode is 800×600, that indicates that pixel in the horizontal direction is 800 and that in the vertical direction is 600. Select panel point for frame image, and joint for movement images.

After pre-processing the data, call the procedures of drawing curves, which are relatively easy. Display the curves on the left or right half through controlling the variables. Function multiple segments on the form surface with Line To and Move To, the two methods provided by Tcanvas so as to fit curves. However, parameters offered by the two methods are LfMargin+(i-1)*Setp_H and Round(DataArray[i])respectively, among which LfMargin and Step_H are the determined left margin and transverse step, I marks the variable of frame number. DataArray is the data need to be displayed after pre-processing.

Human parameters of movement are obtained by applying model of Chinese people, collecting technical images and conducting three-dimensional analysis on the images.

4.5   Result Output

Result display and output: including graphic, images and data tables.

Results can be displayed in four forms:

① Local movement of certain links can be enlarged;

② Control speed of playing, for example, speed up, slow down, play by frames and frame pause;

③ Set up two windows and conduct comparative play of the two images;

④ Synchronous contrast play of different images.

In addition, the method of directly input initial parameters of athletes can be applied to display on the computer the process of putting the shot and throwing range.

4.6   Analysis and Evaluation of Sport Biomechanics

Through analyzing and processing the parameters, the system can find out technical laws of movements in a rapid and accurate manner, evaluate the movements, and provide theoretical basis for improving technology and technique levels of the athletes.

4.7   Function of Assistance

The system, which marks a comprehensive biomechanical evaluation system for shot sport technology, enjoys comprehensive and strong function of customization. Therefore, it documents in an overall and detailed manner every function.

5   Expected Accomplishments

5.1 Due to some objective factors as well as the restrictions of the develop system, the system enjoys space of improvement in the aspect of high intelligence, and cannot realize full-automation and complete the function of technical improvement. Efforts need to be made in future high intelligence. In particular, further studies need to be conducted on the algorithm that can solve the problem of auto match of models so as to enhance relevant efforts and launch better software system as soon as possible.

5.2 Modeling of biomechanical evaluation system for shot sport technology enjoys quite complicated content. Diverse physical models need to be established for different people in different levels. The system mainly performs modeling on athletes of the same level under the most ideal conditions.

References

1. Wang, X., Zhu, J., Yan, H.: Discussion on how to appreciate sport competition. Sichuan Sports Science 3(4), 17–18 (2006)
2. Wang, S.: Athletic basketball sports competition process structure research. Journal of Wuhan Sport College 45(2), 98–100 (2011)
3. Yang, Y., Cheng, L.: Review of Monitoring Research on Tennis Specialized Training (3), 24–29 (2012)
4. Xie, Y., Qiao, P.: Research on the Regulation and Control of Training Process of Elite Athletes in the Year of Olympic Games. Journal of Beijing Sport University 34(7), 121–124, 127 (2011)
5. Wang, J.: Extraction and tracking of human joints in sport images, pp. 29–31, 41, 43. Hebei University of Technology, China (2004)
6. Behrad, L., Shahrokni, A., Motamedi, S.A., Madani, K.: A robust vision-based moving target detection and tracking system. In: Proceeding of Image and Vision Computing Conference, vol. 52(6), pp. 627–630 (2001)
7. Chen, C.J., Zhang, B.J.: An Analysis on Modern Sprint's Technique Characteristic. Zhejiang Sports Science 29(3), 101–103 (2007)
8. Sun, Y., Wang, Y., Sun, Z.G.: Detection of Lower Limb Joints Based on Image Information. Chinese Journal of Biomedical Engineering 26(1), 157–160 (2007)
9. Liu, N., Yang, T.W.: A Comparative Study on the Length and Frequency of Step between 100m Top Athletes in China and Their Counterparts Abroad. Sports Science Research 8(4), 53–56 (2004)
10. Badler, N.I., Allbeck, J., Lee, S.-J., Rabbitz, R.J., Broderick, T.T., Mulkern, K.M.: New behavioral paradigms for virtual human models. Technical Paper 2005-01-2689. SAE International, Warrendale, PA (2005)
11. Badler, N.I., Phillips, C.B., Webber, B.L.: Simulating Humans: Computer Graphics Animation and Control. Oxford University Press, New York (1993)
12. Badler, N.I., Palmer, M.S., Bindiganale, R.: Animation control for real-time virtual humans. Communications of the ACM 42(8), 64–73 (1999)
13. Chaffin, D.B. (ed.): Digital Human Modeling for Vehicle and Workplace Design. SAE International, Warrendale (2001)
14. Choe, S.B., Faraway, J.: Modeling head and hand orientation during motion using quaternions. Technical Paper 2004-01-2179. SAE International, Warrendale (2004)
15. Danker, J.S., Reed, M.P.: A behavior-based model of clavicle motion for simulating seated reaches. Technical Paper 2006-01-0699. SAE International (2006)
16. Warrendale, P.A., Dickerson, C.R., Chaffin, D.B.: Dynamic loading and effort perception during one-handed loaded reaches. In: Proceedings of the 27th Annual Meeting of the American Society of Biomechanics, Toledo, Ohio (2003)

17. Marler, T., Rahmatalla, S., Shanahan, M., AbdelMalek, K.: A new discomfort function for optimization-based posture prediction. SAE Technical Paper 2005-01-2680. SAE International, Warrendale, PA (2005)
18. Martin, B.J., Kim, K.H.: Effort perception of workers with spinal cord injury or low-back pain in manual transfer tasks. In: Proceedings of the International Ergonomics Association Conference, Seoul, Korea (2003)
19. Monnier, G., Wang, X., Verriest, J.-P., Goujon, S.: Simulation of complex and specific task orientated movements — application to seat belt reaching. Technical Paper 2003-01-2225. SAE International, Warrendale, PA (2003)

Real-Time Monitoring System
for Transformer Based on GSM

Li Zou

Anhui Vocational College of Defense Technology, Luan, Anhui, China

Abstract. For real-time monitoring and data acquisition applications of distribution transformer, the paper put forward to wireless sensor network embedded system based on GSM, which can real-time collect and transmit data of monitoring and a variety of on-site information. Practice shows that the GSM networking solutions achieve a unified data management and user remote access based on S3C2440 chip as the core of the embedded gateway monitoring data for processing. This paper gives the system software implementation method, constructed monitoring and management information platform. Through the distribution transformer factory system operating status of the actual monitoring, reliability and timeliness to meet the requirements.

Keywords: GSM, transformer, embedded systems, ARM, Real-time monitoring, data collection.

1   Introductions

Building process of intelligent network progresses, the transformer monitoring and data acquisition technology has developed rapidly [1-3]. All kinds of transformer monitoring products have been applied in many power companies, which ensure grid security and reliable, cost-effective operation played an important role. But there are also some disadvantages [4-8]: System monitoring points less, monitoring frequency is low, real bad, can not be fully reflected immediately grid operation status and parameters circumstances; mostly single-function monitoring and management, such as oil dissolved gas line monitoring system, power quality monitoring systems, core grounding monitoring system, it is difficult to become more highly integrated fusion integrated monitoring platform; various monitoring subsystem development different manufacturers, system scalability and the interface between the poor interoperability, data can not be collected in real time integration and real-time analysis of data correlation; conventional wired data collection methods cabling project large, high installation and maintenance costs. Therefore, the development of more advanced performance, integrated applications and strong transformer data acquisition and monitoring system is necessary.

It is in this situation, according to the power company on the grid monitoring system requirements, using the most advanced GSM / GPRS networks and public telecommunications network data transmission technology, power monitoring

technology and software design technology, successfully developed a real-time distribution transformers monitoring system [8, 10]. The system enables managers timely and accurate understanding of the distribution transformer load change conditions and the operation of equipment, to improve the quality of power supply, improve the reliability of power supply that can play a positive role, to truly achieve a distribution automation, it can fully meet the stage, as well as future needs of the power distribution system automation [11-15].

2   GSM / GPRS Communication Descriptions

GSM is acronym of the Global System for Mobile and the second-generation cellular systems (Digital Cellular System) standards [16-20].

GSM services can be divided telecommunications and data services according to the principles of ISDN, telecommunications services include standard mobile phone business, the mobile station or base station initiated operations as well as emergency calls and faxes. Data services are limited to the Open System Interconnection Reference Model (OSI RM) on the first layers 1, 2, supporting packet-switching protocol, the transparent and non-transparent transmission in two ways. There is a digital service ISDN supplementary service, including call forwarding, caller identification and closed user groups [21, 22]. Widely used in the domestic short message service (SMS) also belongs in one supplementary ISDN services. The service allows GSM mobile phones and base stations in the normal voice service; you can pass a certain length (140 8bit ASCII characters) alphanumeric messages. "Distribution Monitoring Automation Systems' use of GSM to networking, which utilizes the GSM short message function to pass data and control commands.

GPRS and GSM voice existing system is the most fundamental difference is: GSM is a circuit switching system, GPRS is a packet switching system [23]. Therefore, GPRS particularly suitable for intermittent, sudden or frequent, small amounts of data transmission, but also for the occasional large amounts of data [24-26]. This feature is suitable for wide range of large public, special transformer monitoring. GPRS packet switching technology is relatively original GSM dial-up circuit-switched data transmission, a "real-time online", "by the amount of billing," "Quick Login", "high-speed transmission" advantages.

3   System Architecture

Transformer time monitoring and data acquisition system shown the overall structure in Figure 1, the use of a distributed architecture, including four parts [27, 28]: Embedded system by means of various sensors disposed in the transformer (or integrated monitoring subsystem) and GSM module is primarily responsible for collecting real-time operating status of the transformer, which will include voltage, current, temperature, vibration, noise, and other parameters,that can be monitored and sent S3C2440 module. Data were aggregated to the embedded multi-hop routing gateway

system, the core of the system as a whole, using the ARM9-based S3C2440 core processing control module to handle a variety of data, and communicate with the various networks, providing a variety of types of interactive features; network transmission part, including the Internet and GSM/CDMA/GPRS/3G other wired or wireless communication network, to provide users with remote access and control, will monitor the data transmitted to the remote monitoring center and a variety of mobile devices; user through the monitoring center, you can get off-site monitoring data, data analysis and early warning processed and graphically displayed, enabling remote access and control.
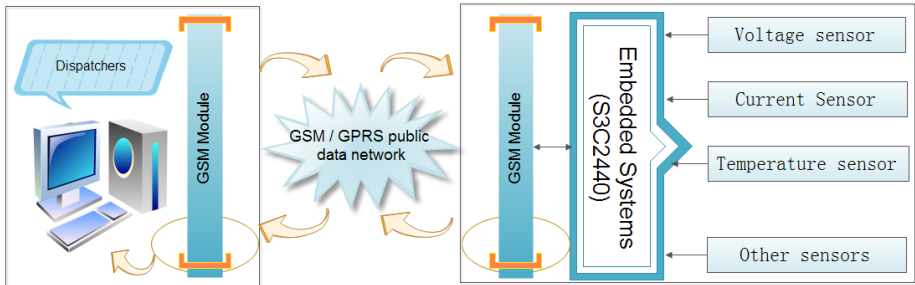
Fig. 1. System architecture of real-time monitoring data acquisition for transformer

4   Hardware Design

As there are many different types of sensors or monitoring subsystem, taking into account compatibility and scalability engineering requirements, the article uses an embedded real-time operating system, system hardware connections is shown in Figure 2. The circuit is composed of sensors and control circuits, GSM module and embedded
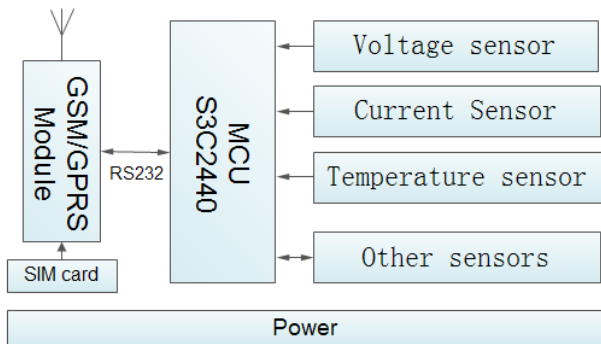
Fig. 2. Hardware connection of real-time monitoring data acquisition system for transformer

processors and other components. Core of circuit is Samsung S3C2440 processor, the function is responsible for the organization and resource allocation system, mainly including the power supply module, low-voltage protection circuit, a memory unit, communication module, analog signal conditioning modules and other components, and deployment in the transformer voltage sensors and current sensors through various sensors such as I / O interface, IIC interfaces, RS232 interface, JTAG interface to connect and have program download and debugging functions. Sensor data acquisition function, the analog signal through power conditioning circuit into a digital signal, and then by the micro-controller on the data obtained after processing package packing after GSM / GPRS module to send.

4.1   Main Processor

S3C2440 processor is 16/32 RSIC microprocessor architecture based on the ARM920T core , not only has powerful data processing capabilities, compatible with most current general-purpose peripherals. Its main properties are: clocked at 400MHz, up to 533 MHz; There are three road UART, SPI and IIS and other types of serial interfaces, 130 general-purpose I / O, 24 external interrupt sources; 4-channel DMA, 4 个 PWM function 16-bit timer / counter and a 16-bit internal timer; 2 external 32 M onboard SDRAM; supports Nor Flash and Nand Flash (SLC) start mode; integrated LCD, I2C, Camera, etc. controller; an SD interface, audio codec interface, 3-way USB control, 16 watchdog timer, clock and power management module, transformer monitoring system able to fully meet the data processing and network transmission requirements.

4.2   Other Circuits

Low-voltage protection circuit prevent overvoltage, overcurrent to burned module; signal conditioning circuit monitoring system is an essential part.Under the control of the master module, the sensor group acquisition phase transformer voltage, current, power, power factor, harmonic harmonic distortion (THD) of the basic power parameters and environmental data. As in the industrial field ,the signal through the field after the reunification of instruments to measure the output of a standard 4 ~ 20 mA standard current signal, as long as the standard output of various types of sensors current input to the processor in the analog input module, the differential inputs converted into 1 ~ 5 V standard voltage signal, the signal conditioning circuit after conditioning input A / D converter, and the upper layer software by the control program and the database can achieve multi-channel data acquisition. Also, each monitoring unit via RS232 or RS485 and MODBUS standard communication protocols and monitoring center for two-way data communication.

5   System Software Design

Large power transformer core current transformer directly reflects changes in the fault condition - the existence of multi-core grounded. Conventional monitoring methods substation duty by regularly using handheld devices to detect and record the biggest drawback of the method is not immediately react to the fault, to further expand the presence of severe accident risks. With the power of raising the level of automation and unmanned substations increases, the urgent need for a reliable, high precision, strong function of the large power transformer core current online monitoring devices. The device described in this article is designed for this purpose, with a high application value. Currently, on-site personnel to use more grip core grounding wire clamp meter to monitor its current, but the interference of strong magnetic field of the transformer, the measured value is very precise, and even several times the same measurement point the case is very different from the measured value, so that measurement the reference value is questionable; while some domestic units developed automatic monitoring devices, been able to discover more ground fault, but the lack of real-time fault monitoring. For this reason, the design of a GSM-based communications line monitoring device, a better solution of the transformer core grounding current online monitoring issue. In order to solve the normal state of the ground current is small, but the interference by the site more serious problem.

5.1   System Program Running and Gsm Initialization

First, GSM hardware initialization, and then open the global interrupt, the program has begun to enter the application. If the node joint to network successfully, the system enters the sleep state; then the system sample data by the control command, the main processing module complete the initialization and the corresponding interrupt, if the monitored parameters do not exceed the system reservation threshold value, the one operation completed, the next cycle. Either argument exceeds the threshold, the alarm monitoring center. To achieve low power consumption, a data acquisition nodes collect and transmit every success goes to sleep and then wake up again for a second collection work and the data sent to the remote host. GSM initialization function:

```
void gsm_init(void)
  {
   LOOP:
   delay_ms(3000);
   sendstring(AT);
   delay_ms(1000);
   sendstring(ATE);
   delay_ms(1000);
   sendstring(AT_CNMI);
   delay_ms(1000);
   sendstring(AT_CSMP);
   delay_ms(1000);
```

```
•• clr_inbuf();
•• sendstring(AT_CMGF);
•• delay_ms(1000);
•• if((in_buf[2]=='O')&&(in_buf[3]=='K'))
•• {
••   clr_inbuf(); //clear buffer
••   LED_1=0;
••   delay_ms(1000);
••   LED_1=1;
•• }
•• else
•• {
••   clr_inbuf();
••   goto LOOP;
•• }
••}
```

Where "if"：Determine whether the module initialization is successful, successful module would reply "OK" to single chip, if the SCM does not receive OK, just continue to send initialization directives

5.2   Software Design for Monitoring Center

According to 2011 release of "Q/GDW534-2010 substation equipment monitoring system Technical Guide" and other series of standards, monitoring system of transformer configuration is composed of the DGA (multicomponent), partial discharge, power quality monitoring, environmental monitoring, alarm linkage, and many other monitoring systems. Oriented to meet the power enterprise needs of complex business application integration technology architecture, both run independently of each subsystem, but interrelated, not only have basic supervisory control and data acquisition (SCADA) functions, including information collection and analysis of management functions, and should have and power management information system (MIS) expansion interconnect to form a more large-scale information systems.

Monitoring center software is design in modular, common business processes and work-related needs to centralized, unified planning and consolidation, the purpose is to build intelligent management requirements in line transformer for distributed, open platform. Which provides real-time monitoring data, information transfer, database systems, human-machine interface, applications and other services, and the application of intelligent transformer monitoring subsystem in open system architecture (OSA) based on the integration. Program achieved a major system management, system monitoring, parameter setting, communication management, data processing and statistical analysis functions. We can be reserved multiplexed data communications and other automated systems interface to enable information sharing, to facilitate a smooth upgrade the system in the future, expand and maintain.

6   Conclusion

This paper presents a real-time transformer monitoring system based on ARM core that can be used for real-time information to achieve remote monitoring of transformers and a variety of on-site data collection and transmission. The operational status of the system is performance and stability through the transformer testing.The system monitor data compared with Fluck43 harmonic analyzer, the values is very close to an accuracy of 0.001 dimensionless level,the system can be more accurately complete transformer monitoring data acquisition and data transmission via GPRS. But for how to integrate existing monitoring component transformer function, to better achieve the multi-sensor data fusion systems, scheduling, distribution and other aspects of coordination needs further study.

With China's rapid economic development, urban electricity consumption increasing urban distribution network security and reliability are also increasingly important. Safety and reliability of distribution transformers distribution network determines the safe and reliable operation, therefore, need for distribution transformers through real-time monitoring of running status. Distribution transformers located in various enterprises, residential and urban streets, because of communication difficulties with conventional automated detection means difficult to achieve real-time monitoring of a large number of distribution transformers. It is in this situation, according to the power company for the grid monitoring system requirements, using the most advanced GSM / GPBS network and the public telephone network data transmission technology, power monitoring technology and software design technology, successfully developed the EU-2500 type with real-time monitoring system electrical transformer. The system enables managers with timely and accurate understanding of the varying load conditions and changes in the operation of equipment, to improve the quality and reliability of power supply can play a positive role, to truly achieve a distribution automation, it can meet the stage and future demand for power distribution system automation. According to the power companies on the grid monitoring system requirements, using the most advanced GSM / GPRS network and the public telephone network data transmission technology, power monitoring technology and software design techniques proposed EU2500 type distribution transformers real-time monitoring system. The system enables managers with timely and accurate understanding of the varying load conditions and changes in the operation of equipment, to improve the quality of power supply, improve the reliability of power supply can play a positive role, to truly achieve a distribution automation, which can fully meet the current stage and the future needs of the power distribution system automation, optimization systems from different aspects of the operation and management level of automation.

Now the state is the development of intelligent grid, while the transformer core line monitoring system is a form of intelligent grid more important part of the foreign in the 1980s have begun to realize the transformer core monitoring system to study and achieve unmanned substation, can save a lot of manpower and resources, and the ability to timely and effective discovery transformer core ground situation and the data can be analyzed through a series of multi-point grounding transformer exists the danger, and

better ensure the normal operation of the transformer equipment. Current national safety regulations have been proposed in the transformer core online monitoring requirements, within the next five years to achieve full-line monitoring of transformer core deployment. Foreign scientists and authorities for such regular maintenance and preventive testing has long been recognized inadequacies have have committed to electrical equipment on-line monitoring and condition-based maintenance study. Its advantage is that it can detect early signs of failure, so that operation and maintenance personnel when a failure in its infancy through early detection means to eliminate hidden dangers, so as to avoid the occurrence of fatal accidents, improve service quality and efficiency. With the sensor technology, signal processing technology and the development and application of computer technology as a basis for state maintenance of the electrical equipment on-line monitoring technology has been rapid development, has become insulated detection is an important part, it will depend in many ways make up only Regular preventive tests brought inadequacies. Through the core ground current on-line monitoring and accurately determine the core of the working conditions, which targeted the core before failure timely maintenance, not only effectively improve the reliability of power supply, but also reduces the running costs of the power system, for guarantee the safe operation of power transformers has very important significance.

Winding of the transformer and the core is passed, changing the main components of the electromagnetic energy, to ensure their safety and reliable operation of the key of the transformer. Statistics show that the core problems caused due to failure, accounting for the total transformer accident in the third. During normal operation, the transformer core need to have some ground to avoid floating potential core due to discharge its core grounding current is very small, about a few to tens of milliamps, when multi-core transformer ground fault occurs when generates eddy currents will increase the core grounding or even tens of amperes to several amperes, which will cause local overheating of the core, causing localized overheating core insulating oil decomposition, also make the ground sheet blown or burned core, lead core potential suspension, resulting in discharge, causing light to heavy gas gas action and even tripping, or even damaged transformers, main transformer caused a major accident. By measuring the ground current transformer core transformer directly reflects the fault status - whether there is core multi-point grounding. In the "electrical equipment preventive testing procedures" (Q/CSG10007-2004) in 5.1 "oil-immersed power transformers" on "the core and clamp insulation resistance" requirement: "Running core grounding current is generally not greater than 0.1A ". Therefore, accurate and timely diagnosis and treatment of multi-core transformer ground fault, to ensure the safe operation of the transformer has an important significance.

References

1. Tan, Y.: GSM/GPRS communication in the distribution transformer automatic meter reading application. Northeast Dianli University 24(4), 61–63 (2004)
2. Dan, C., Xu, S.: GSM/GPRS communication in Distribution Automation System. Electronic Design & Application 3, 66–70 (2004)

3. Xu, Y.: Based on GPRS technology distribution transformer monitoring system terminal design. South China University of Technology (Natural Science Edition) 34(9), 35–38 (2006)
4. Long, Z., Jun, L., Li, Y.: Based on GSM network distribution transformer monitoring system. Electric Power Automation Equipment 23(6), 57–59 (2003)
5. Peng, L., Wu, J., Yang: Political Civilization based on Bluetooth technology and GSM communication centralized meter reading system. Demand Side Management, 05 (2005)
6. Zhao, Jun, L., Che, R.F.: GSM distribution transformer monitoring applied research. Electric Power System and Automation, 04 (2003)
7. Tan, B., Wang, P.: SMS-based communications repeater control system design. Microcontrollers & Embedded Systems, 10 (2002)
8. Zhang, S., Lie, C., Pai.: Distribution Transformer Based on GPRS remote monitoring alarm management information system design. Electrical Switch, 02 (2005)
9. Fan, C.-Y., Wen-Bin, Wang, B.: Based on the GSM short message communications Air Traffic Management Monitoring System. Electronic Measurement and Instrument, 04 (2005)
10. Xu, S., Wang, R.: distribution network based on GSM network fault auto Processing System. North China Electric Power, 02 (2005)
11. Fei, C.: Wireless mobile digital ECG machine design. Beijing Biomedical Engineering, 03 (2005)
12. Lei, Liao, J., Chen, J.: Mobile intelligent data service network functional model of research and design. Telecommunication Engineering, 03 (2004)
13. Cheng, H.X.: Based on GSM transmission mode electric tube is a modern management system. Electrical Measurement & Instrumentation, 07 (2003)
14. Zhao, Jun, L., Che, R.F.: GSM used in distribution transformer monitoring study. Electrotechnical Journal, 07 (2003)
15. Shao, L., Wang, H., Fang, P., Li, N.: Intelligent Supervisory Control System of Urban Street Lamp Based on GSM Short Message. In: Seventh International Symposium Test Technology (2007)
16. Wang, Z., Fan, C.-Y.: GSM protocol converter design and implementation. In: Chinese Association of Automation, China Instrument Society 2004 Southwest Provinces and One City Automation and Instrumentation Annual Conference Proceedings (2004)
17. Zhang, W., Lei, Wei, X.: RS485 and LAN-based power plant auxiliary power automatic meter reading system. In: Eleventh National Electrical Math Conference Proceedings (2007)
18. Zhang, H., Wang, A., Xu, J.-Y.: Ground state based on GSM technology locking system monitoring and error prevention. In: First Session of the Electrical Equipment and Intelligent Conference Proceedings (2007)
19. Yang, C., Wang, F., Dehua: GSM data services and carrying capacity. Telecommunication Engineering, 01 (2002)
20. Lin, L.: Distribution automation systems with mixed communication scheme. Power Systems, 023 (2001)
21. Wang, Y., Jia, Hu, X.: Based on wireless network distribution transformer monitoring system application. Power Systems, 16 (2002)
22. Chen, Y.-W., Yu, L., Wang, Y.: Based embedded Ethernet or GSM network electric energy telemetry system. Power Systems, 15 (2003)

23. Yu, L., Luo, X., Yu, S.-L.: GPRS-based distribution automation system solutions. Power System, 24 (2003)
24. Yong, L.: PDU analysis and cell phone SMS Control Development. Electric Power Systems, 12 (2004)
25. Wang, H.P., Zeng, R., He, J.-L.: Distribution network using the GSM network remote data acquisition feasibility analysis. High Voltage Apparatus, 05 (2002)
26. Zhe, W., Zhou, Y.: Based on GSM wireless network remote meter reading system design. Hebei University of Technology, 04 (2003)
27. Cao W.Q., Han B.: Use of GSM short message to achieve remote monitoring. Radio Engineering, 10 (2002)
28. Li, Y., Niu, Z.-X., Lin, D.: Novel based on short message service remote smart meter reading system. Information Engineering University, 01 (2003)

Sliding-Mode Controllers for Ship's Track-Keeping Control Systems Based on Input-Output Linearization

Chen Shao-chang[1], Fanyue Yao[1], and Li-ping[2]

[1] College of Electronics Engineering, Naval University of Engineering 430033, Wuhan, China
[2] Northwest Regional Air Traffic Management Bureau of Civil Aviation 710082, Shanxi Xi'an, China

Abstract. For the ships tracking nonlinear system with steering link, the pole assignment and sliding mode controllers based on input-output linearization are designed under the assumption that the system parameters are constants. The simulation results prove: in the case of nominal parameters, the designed sliding mode controller can track the setting path effectively, which has strong robustness under the condition of disturbance, but produces the chattering.

Keywords: nonlinear, ship's track-keeping control system, input-output linearization, Sliding-mode control.

1   Introduction

This paper presents an algorithm for a ship course control system. The algorithm converses the nonlinear tracking control system into linear mathematical model by linearizing the input-output parameters. A sliding mode variable structure controller is then introduced in order to take advantage of its strong robustness, which also serves as the controller for the nonlinear tracking control system. Lastly, Simulation is performed to verify the efficiency of the controller designed [1-3].

With the continuous development of the shipping industry. Increasing the density of ship navigation. Waterways and ports become quite narrow. And thus the safety of navigation and economic requirements are also constantly improving, on track and heading control accuracy increasingly high requirements [4-7]; currently ship control system for heading control or track control system mainly indirectly. Its control performance and structure have yet to be further improved. Network Control System (Networked Control Systems, NCS) is the last ten years, with control technology, network technology and the rapid development of computer technology and the emergence of a new control system theory. Ethernet as a successful network technology in the commercial and office technology are widely used because of its high-speed communications. Technical resource-rich [8-11], low cost. So that this technology is increasingly being applied to industrial control. The ship automation systems is toward digital, network -oriented direction. For this paper, an Ethernet-based ship tracking control method is used. Firstly, the use of MMG (Maneuvering Mathematical Model

Group) modeling idea to establish a mathematical model of the ship three degrees of freedom. And the establishment of wind, waves and currents and other disturbing force mathematical model. Then analyze the advantages and disadvantages of the traditional PID. And on this basis, using a nonlinear tracking differentiator (TD Ⅰ) on the input signal " set" reasonable transition process and to obtain its differential signal. But with another nonlinear tracking differentiator (TD Ⅱ) recover and get their feedback signal output differential signal. And then to deal with the nonlinear combinations error signal. The error signal and the error integral differential signals solves the contradiction between fast and overshoot. And then for the closed-loop control for delay systems, non-minimum phase system, the shortcomings of slow response. Adding a feedforward control, ultimately generating a control signal track. The article is part of the combination of the four non-linear PID controller. Steering the ship stage set point steering and steering the ship from the ship to reach this distance. And then transferred to the next track using the track segment controller. Secondly. This paper raises the network control system based on the concept for the Ethernet introduced its random delay characteristics. Then elaborated in this paper the design of butyric Ethernet ship control system simulation design principles. Introduces TCP / IP protocol. And network communication using client / server (C / S) communication model and communication theory in detail analysis, but raised the average network latency test methods. Finally, the network control system model is the lack of network environment. The actual completion of the LAN Ethernet-based ship track control system simulation. Simulation results show that the use of this article nonlinear PID controller. Enabling the ship during the voyage track fast track to achieve the control requirements.

In recent years, increased maritime transport. Ship towards large-scale and high speed. Resulting in greatly increased density of shipping. Ship during the voyage will be a variety of unpredictable external environment interference. Making the ship maneuvering problem becomes particularly complex. The ship in the distance between the two turning points sailed commonplace. To save time. Shorten the distance and cost savings. The ship straight track control is very targeted and practical ; Most ships are less driven. Using relatively small amount of control to achieve the tracks and other operation control. Thereby reducing the weight of the ship and operating costs, improved operational efficiency. Economy of operation to improve the performance of a ship. To sum up. The less drive straight track movement of the ship has very great significance. In this paper, the ship straight track movement as the research objects. The sliding mode variable structure control, adaptive control, backstepping. Neural network function approximation methods such as the introduction of straight track underactuated ship tracking control.

2   Mathematical Model for the Ships Tracking Control System

Figure 1 shows the schematic drawing for the coordinate system of the tracking system. X, y are the position of the ship's Center Of Gravity (COG) relative to the global coordinate XOY.
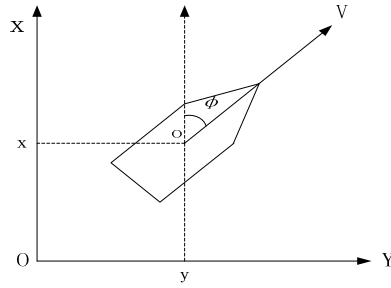
Fig. 1. Coordinates for the ship tracking control system

Assuming the ship is heading north and the mathematical model for the system can be shown as the following nonlinear state equation [5]

$$\begin{cases} \dot{y} = V \sin \phi \\ \dot{\phi} = r \\ \dot{r} = -\frac{1}{T}r - \frac{\alpha}{T}r^3 + \frac{K}{T}\delta \end{cases} \tag{1}$$

Y is the ship's horizontal displacement, ϕ is the heading angle, V is the ship's velocity. R is the heading angle acceleration, δ is the ship's control angle. K,T,α are determined by the ship's block coefficient. capacity weight, speed which are assumed to be constant.

The steer can be modeled as first order inertia link, and is described as the following

equation $T_E \dot{\delta} + \delta = K_E \delta_E$ (2)

T_E is the time constant of the steer, K_E is the Gain of the steering control, δ_E is the order from tracking controller.

The goal of the system is to set δe so that the tracking is stable nonlinearly at the same time, y, ϕ, r, δ converge to zero in the steady state.

3   Controller Design

3.1   Input-Output Linearization of Tracking Control

Let the state variable is $x = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix}^T = \begin{bmatrix} y & \phi & r & \delta \end{bmatrix}^T$,the input is $u = \delta_E$ and the output is $z = \phi + \eta(y)$.

In order to satisfy: a) the control strategy based on input-output linearization leads to a stable system; b) when z=0, y, ɸ, r converge to zero, the output z needs to meet the following requirements [4]:

1) z gradually converges to zero globally.

2) When $y \neq 0$, $\sin(\eta(y))y > 0$;

3) When $|y| \to \infty$, $\int_0^y \sin(\eta(y))dy \to \infty$.

Define the output z as the following $z = \phi + \eta(y) = x_2 + \arctan(kx_1)$ (3)

Where k is a real number larger than zero, chosen by the designer. It is obvious that the equation above satisfy the requirement.

Therefore, the system's nonlinear state space equations are as following:

$$\begin{cases} \dot{x} = f(x) + g(x)u \\ z = h(x) \end{cases}$$

(4)

Let $m_1 = \eta(y) = \eta(x_1)$. Take derivative of m_1 with respect to time, yield:

$$m_2 = \dot{m}_1 = \dot{\eta}(x_1)V \sin x_2$$

(5)

Take   derivative   of m_2 with   respect   to   time,   yield:

$$m_3 = \dot{m}_2 = \ddot{\eta}(x_1)(V \sin x_2)^2 + \dot{\eta}(x_1)Vx_3 \cos x_2$$

(6)

Take derivative of m_3 with respect to time, yield:

$$m_4 = \dot{m}_3 = \dddot{\eta}(x_1)(V \sin x_2)^3 + \ddot{\eta}(x_1)V^2 x_3 \sin(2x_2) + \ddot{\eta}(x_1)V \sin x_2 + \dot{\eta}(x_1)(Vx_3 \cos x_2)^{'}$$ (7)

Combine equation (4) and take derivative of z with respect to time, yield: $\dot{z} = x_3 + m_2$

(8)

Continue to take derivative of $\dot{z}$. With respect to time, yield

$$\ddot{z} = -\frac{1}{T}x_3 - \frac{\alpha}{T}x_3^3 + \frac{K}{T}x_4 + m_3$$

(9)

In order to calculate input u, take derivative of $\ddot{z}$, yield

$$\dddot{z} = (-\frac{1}{T} - \frac{3\alpha}{T}x_3^2)(-\frac{1}{T}x_3 - \frac{\alpha}{T}x_3^3 + \frac{K}{T}x_4) - \frac{K}{TT_E}x_4 + \frac{KK_E}{TT_E}u + m_4$$ (10)

From the definition of relative order, it is found that the relative order of the system is 3, combined the relative knowledge of the input-output linearization, the input control u can be calculated as

$$u = \frac{-L_f^3 h + v}{L_g L_f^{3-1} h} = \frac{TT_E}{KK_E}[(\frac{1}{T} + \frac{3\alpha}{T} x_3^2)(-\frac{1}{T} x_3 \frac{\alpha}{T} x_3^3 + \frac{K}{T} x_4) + \frac{K}{TT_E} x_4 - m_4 + v \quad (11)$$

Substitute (12) in to (10), yield

$$\dddot{z} = v \qquad (12)$$

$$\begin{cases} \dot{z}_1 = z_2 \\ \dot{z}_2 = z_3 \\ \dot{z}_3 = v \end{cases}$$

The system's state equations are：   (13)

Where output $z = z_1$, therefore, the design problem of nonlinear system is converted into the design problem of linear system.

3.2   Sliding Mode Variable Structure Controller Design

In order to mitigate the influence of disturbance on the ship control, automatic steering system needs to be robust. Sliding mode control has the property of stay unchanged subject to either system parameter variances or outside disturbance, and can be used to design an robust controller. Figure 2 shows the flow chart of the sliding mode variable structure controller based on input-output linearization：

The sliding mode variable structure control process consists of two steps, 1) approaching stage, it is required for all the phase   portraits outside the transforming surface s=0 to arrive at the transforming surface， i.e the achievability. 2) the transforming surface is within sliding mode space and the sliding movement achieves stable gradually, having good dynamic property[8].
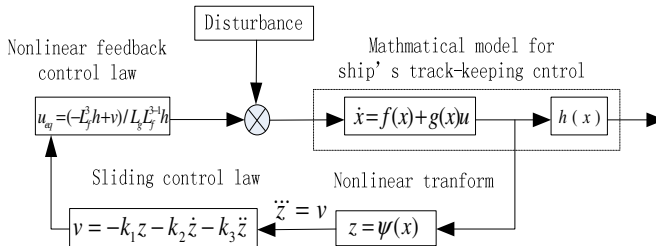

Fig. 2. The block diagram of sliding control

3.3   Sliding Mode Surface Design

After linearization, the tracking control system state space equation can be expressed as formula (10). Let ξ_d is required output. And define $e = \xi_1 - \xi_d$ (14)

Set the sliding mode surface as $s = c_1 e + c_2 \dot{e} + \ddot{e}$ (15)

Where are constants being larger than zero, they are the gains of the system chosen by designer. Choose proper value of c_1 and c_2 so that the eigenvalue of the linear system defined by (16) have negative real part, therefore, the sliding mode function approaches to zero, which guarantee the stability of sliding mode.

3.4   Control Strategy Design

The condition for the movement to arrive at transforming surface is

$$s\dot{s} < 0 \tag{16}$$

The condition however cannot reflect how the movement approaches the switch surface.nor can it guarantee the quality of approaching process. An approaching rule is used to design the control.

Take the exponential approaching rule:

$$\dot{s} = -\varepsilon sign(s) - k's \tag{17}$$

Where ε, k' are constants larger than zero, decrease ε and increase k' will expedite the approaching speed, and mitigate charttering[9]。

Set Lyapunov function as：

$$V = \frac{1}{2}s^2 \tag{18}$$

Yield：

$$\dot{V} = s\dot{s} = -s(\varepsilon sign(s) + k_1 s) = \frac{-\varepsilon s^2}{|s|} - k_1 s^2 < 0 \tag{19}$$

Therefore the overall control system is graduate (asymptotical) stable under the Lyapunov's meaning.

Combine (14),(16) and (18), yield：

$$v = -\varepsilon sign(s) - k's - c_1 z_2 - c_2 z_3 \tag{20}$$

Substitute (17) into (18) to get final feedback control rule, yield

$$u = \frac{-L_f^3 h + v}{L_g L_f^{3-1} h} = \frac{TT_E}{KK_E}[(\frac{1}{T} + \frac{3\alpha}{T}x_3^2)(-\frac{1}{T}x_3 - \frac{\alpha}{T}x_3^3 + \frac{K}{T}x_4) + \frac{K}{TT_E}x_4 - m_4 - $$
$$\varepsilon sign(s) - k's - c_1 z_2 - c_2 z_3] \tag{21}$$

ε、k'、c_1、c_2 are larger than zero, and can be obtained through adjustment.

4   Simulation Study

Using the example of the practice vessel in Reference [10], the simulation data are as follow: ship length 126 m, ship width 20.8 m, max draught 8.0 m, block coefficient

0.681, speed 7.7 m/s. K and T are calculated to be 0.478 and 216 respectively. Assume α=30. Using sliding mode control rule based on input-output linearization equation (18), choose the horizontal displacement of the ship corresponding to the output variable to be 300 m, heading angle to be 10 degrees.

Set : $k^{'} = 0.2, \varepsilon = 0.00024, k = 0.0017\ c_1 = 0.03, c_2 = 0.15$.

With the ship's initial horizontal displacement 300 m, initial heading angle 10 degrees, and the tracking direction to be north, simulations are performed using Simulink, the results are shown in figure 3, figure 4 and figure 5. Setting the distance of wind and wave to the ship tracking as δe=0.1+0.2sin(0.3t), the simulation results are shown in figure 6, figure 7 and figure 8.

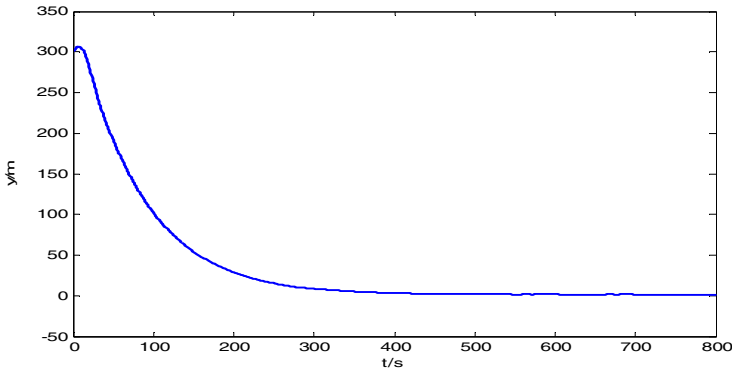$$\delta_e = 0.1 + 0.2\sin(0.3t)$$

Fig. 3. Horizontal displacement time history without disturbances
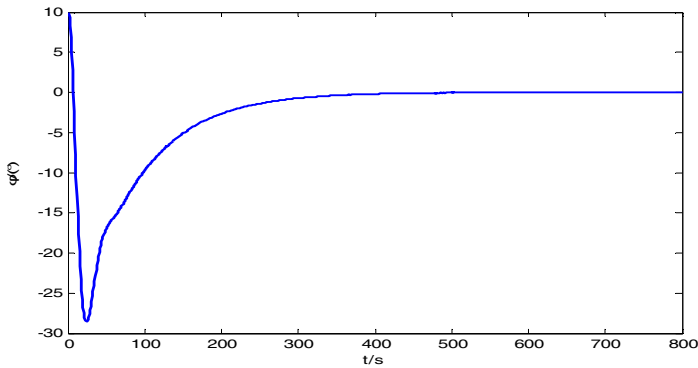

Fig. 4. Heading angle time history without disturbances
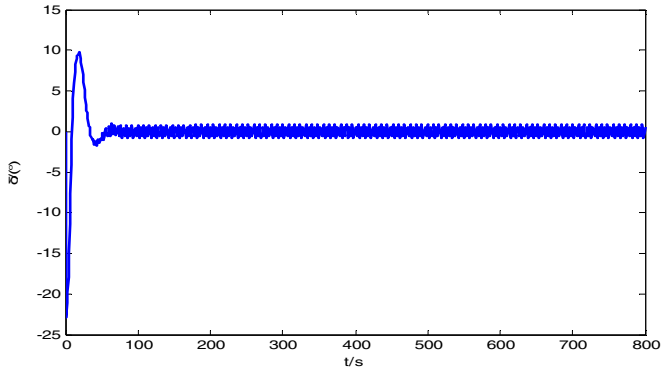
Fig. 5. Rudder Angle Time History without Disturbances
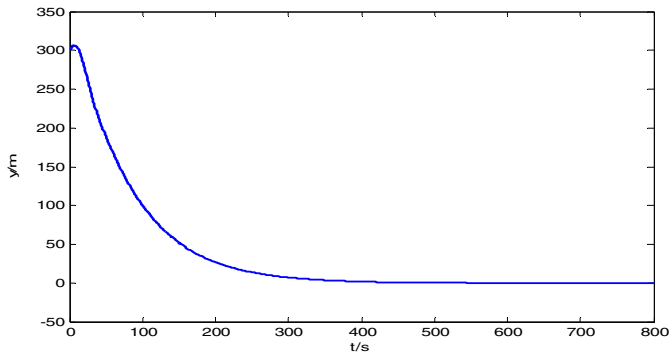
Fig. 6. Horizontal disturbances includingdisplacement time history
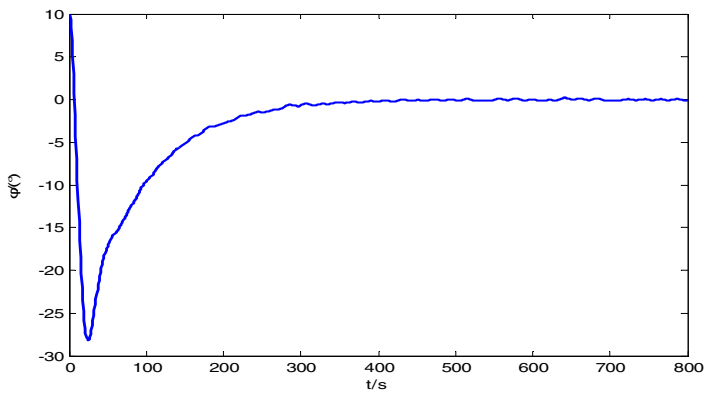
Fig. 7. Heading Angle Time History including Disturbances
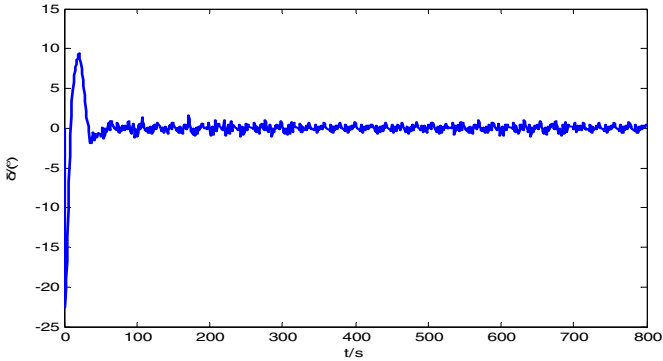
Fig. 8. Rudder angle time history curve including disturbances

It is observed from figure 3 through figure 5, for the nonlinear ship tracking model with steering model, the robust controller designed through input-output linearization and sliding mode control algorithm can keep the ship on the predetermined track relatively fast and accurately with symmetrical parameters. The drawback is that the control steering angle shakes, as shown in figure 3-16. Although the shaking magnitudes and frequencies are within the steering design load, it will wear and tear the steer. Under the disturbances from outside, the controller can overcome the equivalent distances to the input and output, set the ship's tracking efficiently, and converge the horizontal displacement to zero. In addition, the heading angle shaking (variance) is trivia, which demonstrate the robustness of the design.

5   Conclusion

This paper presents a design of sliding mode controller based on input-output linearization for ship tracking control system nonlinear mathematical model that includes steer.   Simulations are performed to verify the efficiency of the algorithm on tracking and its robustness subject to disturbances. However, the introduction of distances does not consider parameter perturbation. Also, the simulation results indicate that the controller exhibit some shaking behavior. The above two limitations are not permitted in real application, and should therefore be further studied.

References

1. Mei, K., Li, H.-G.: Sliding mode control for a class of improvement of reaching law. Anshan University of Science and Technology 04 (2007)
2. Ma, D., Zhigang, Yue: Expensive high; rocket with integral term Optimal Sliding Mode Servo Control. Ordnance Technology 10 (2007)
3. Chen, F.H., Ma, D., Yangbi, W., Zhu, Z.-C.: Rocket Global Sliding Mode Control AC servo system. Ordnance Technology 06 (2012)

4. White country long, Ye, Q.X., Wang, Z.-L.: Has multiplied nonlinear robust control motor speed. Journal of Beijing University of Aeronautics and Astronautics 07 (2008)
5. Zhao, X., Jiang, Y., Wu, Y., Zhou, Y.Q.: Based on multi-modal non-singular sliding fast terminal sliding mode control. Journal of Beijing University of Aeronautics and Astronautics 01 (2011)
6. Mei, Y., He, Y., Li, Z.-X.: Based on sliding mode variable structure indirect matrix converter - Induction Motor Control System. World of Inverters 10 (2011)
7. Xiaogang, Liu, J.: Palletizing robot inverse design based Neural Network Adaptive Control. Packaging Engineering 01 (2012)
8. Changxi, L., Jing, F., Li, Y.: Backstepping-based ship straight track not entirely driven controller design. Ship Engineering 04 (2008)
9. Jitong, B., Wang, H., Yang, X.: SMC-based platform for large-scale optical tracking DC speed control system algorithm. Journal of Jilin University (Information Science Edition) 01 (2009)
10. Gao, H., Shaw, K.-Y., Li, Y.: Uncertain Systems Adaptive Dynamic Neural Sliding upper bound control. Journal of Jilin University (Information Science Edition) 03 (2010)
11. Dong, L., Liu, X., Sun, Y.: Bearingless induction motor radial suspension force differential geometry Variable Structure Decoupling Control. In: Twenty-sixth Chinese Control Conference (2007)

Bus-Mounted Intelligent Control System
Based on GE Controller

Yanfei Chen

Nanjing Communications Institute of Technology, Nanjing, Jiangsu 211188 China

Abstract. By analyzing the existing bus-mounted system and combining with the present control technology, a new bus-mounted intelligent control system cored at GE controller is proposed in this paper, the function demands and design ideas of the bus-mounted intelligent control system are introduced, and finally the hardware and software design of the system is provided. A new control system is designed from lighting, energy conservation, escape, and intelligent scheduling, so as to provide a new idea for the existing bus-mounted system.

Keywords: Bus-mounted System, Intelligent Bus, GE Control System.

1   Introduction

According to the *Suggestions on Giving Top Priority to the Development of Urban Public Traffic* issued by the General Office of the State Council in 2005, it is known clearly that giving top priority to the development of urban public traffic and promoting the development of urban bus service are an important part of the public traffic. As one of the most frequently used tools for domestic residents, how to improve safety, implement green travel, increase the riding comfort of passengers, and reduce the workload of drivers is being increasingly focused by people. In the meantime, as the concept of green traffic is deeply rooted in the minds of people, higher requirements are proposed by people on the existing bus performances [1-5]. People's expressions on the current situation of the existing bus operation system and how to improve service requirements are shown in table 1.

Table 1. How to improve the quality of existing bus services

The current situation of existing bus services	How to improve the quality of bus services
Capacity is not effectively controlled	How many passengers can be effectively calculated?
Comfortable environment is absent	Can the temperature in bus be measured according to the actual temperature?
Safety is not ensured	Can the door of bus be effectively opened once a fire or other abnormal situation occurs? Can the passengers escape?

Table 1. (*continued*)

The control panel of driver is very complex	How is the driver's job complexity reduced for the sake of the driving safety?
The real-time condition of buses is not timely controlled	Can the frequency of dispatching a bus be adjusted in real time according to the actual situation?
Environmental pollution and energy waste	How to implement energy conversation and emission reduction

According to the above focuses, a bus-mounted intelligent control system is designed in combination with GE control technology. It is an intelligent, humane bus-mounted control system integrating the functions of automatically counting passengers, safely escaping from fire accident, and intelligent dispatching management system. The main functions of the system are as shown in figure 1.
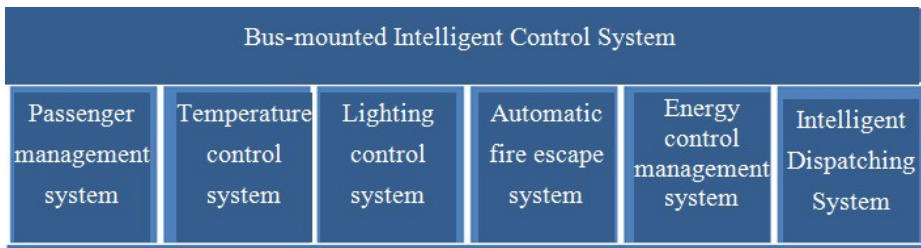
Fig. 1. The function structure of the bus-mounted intelligent control system

The bus-mounted intelligent control system is able to implement the automatic control of the buses according to testing environment variables and in combination with the setting of the bus drivers, so that an intelligent, safe, energy saving, environmental protection, and humanized environment is provided for passengers to take bus [6-9].

2   The Hardware Design of the System

In this design, the PAC control product of GE is used and the following modules are mainly applied [10-14]: (1) IC695CPU310 CPU module; (2) IC695ETM001 Ethernet communication module; (3) IC694MDL645 digital input module; (4) IC694MDL754 digital output module; (5) quick panel touch screen, etc. When it is used, the original parameters should be set, so that all parameters of a bus are set in the driver room through the touch screen, and also these setting values (SV) are written into GE PAC System RX3i controller via Ethernet. When buses run normally, PAC System RX3i controller enters a state of running; the operation result is read into the register and displayed on the driving room to be intelligently controlled through after the measured value (PV) obtained by reading the on-site sensor of bus's I/O device is operated. At the same time, these parameters are sent to the bus control room, through the wireless network system, and then intelligent dispatching is implemented by the control room according to the related data. The overall hardware design is as shown in figure 2.
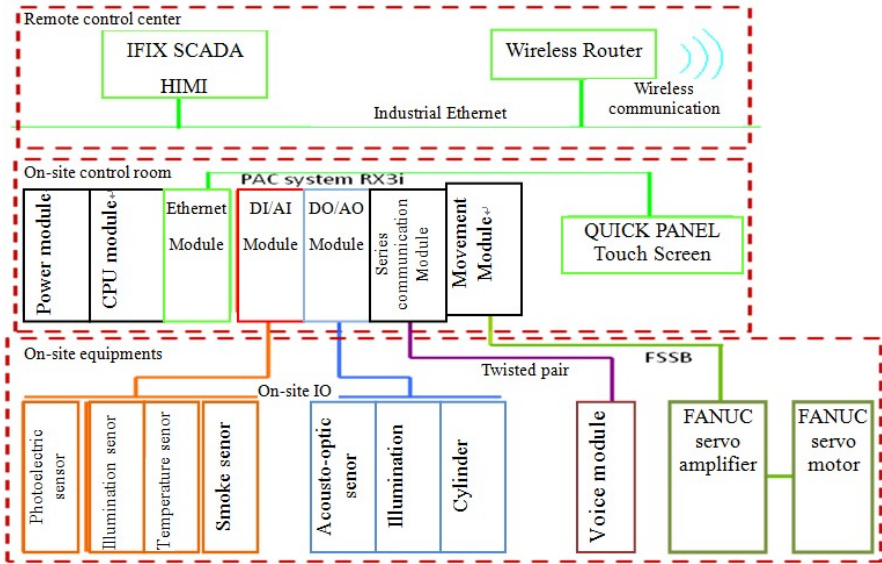
Fig. 2. The hardware design of the bus-mounted intelligent control system

3   The Software and Function Design of the System

Proficy Machine Edition is used as software platform in the system. Proficy Machine Edition provides users with a uniform user interface, supports the fast, powerful, and object-oriented programming, and can implement the functions such as detection, monitoring, execution, and automatic refresh. Also, it includes the web-based functions. Combined with the system functions, the program of the bus-mounted intelligent control system is described as follows:

3.1   Passenger Management System

Passenger management system is mainly composed of the infrared correlation sensors on two sides of front and back doors, the PAC for controlling host, and cockpit touch-screen display, etc. The system should be initialized. The photoelectric sensor is used for detecting if passengers get on or get off the bus and then calculate the number of the current passengers in the bus. The function flow of the passenger management system is as shown in figure 3. The program of the passenger management system is as shown in figure 4.
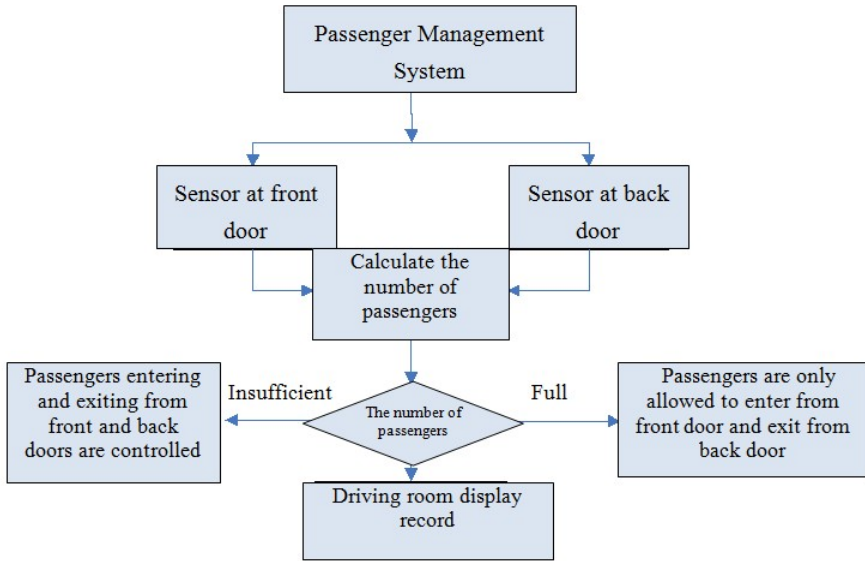
Fig. 3. The function flow of the passenger management system
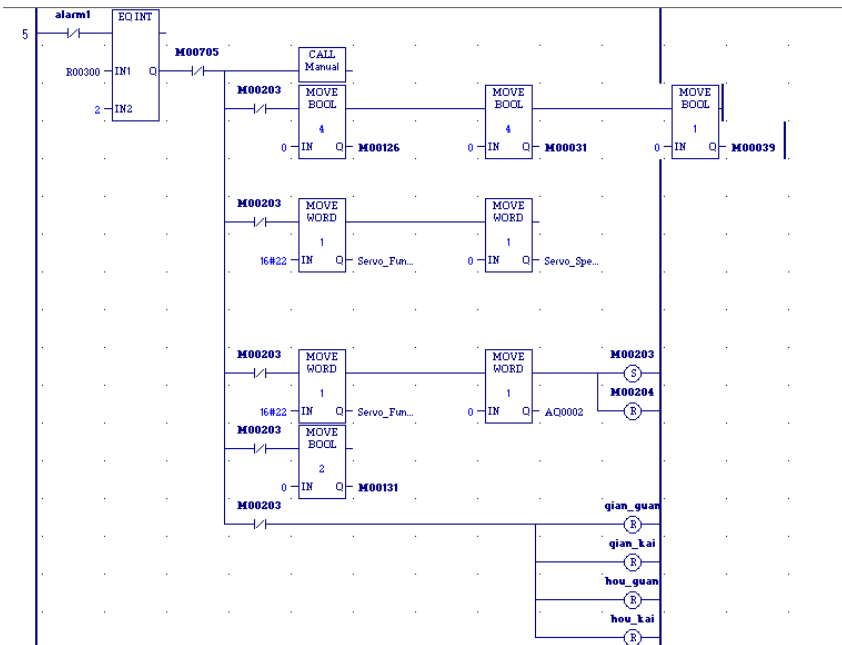
Fig. 4. The program of the passenger management system

3.2   Temperature Control System

Temperature control system is mainly composed of temperature sensor, the PAC for controlling host, and cockpit touch-screen display, etc. In the working of the system, the conditions for turning on air conditioning should be first set, and then the environmental temperature in the bus is detected and the number of passenger parameters is read according to the temperature sensor, so as to judge if air conditioning is turned on. In the work time of the air conditioning, the air conditioner motor speed is adjusted using PID algorithm according to the temperature change inside the bus, and then the amount of the supplied air in the air conditioner is changed. The function flow of the automatic temperature control system is as shown in figure 5. The program of the automatic temperature control system is as shown in figure 6.
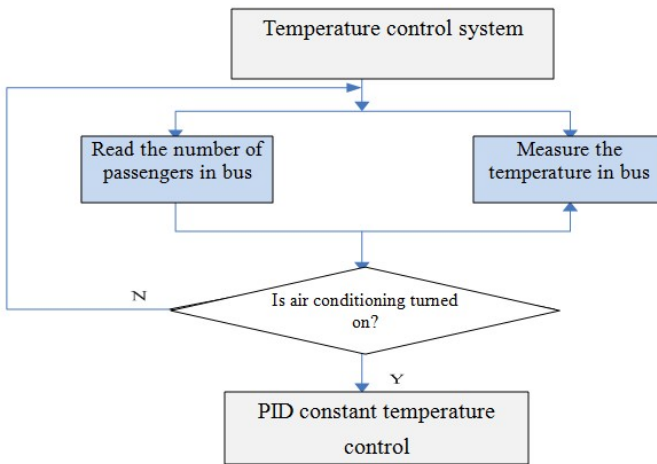
Fig. 5. The function flow of the automatic temperature control system
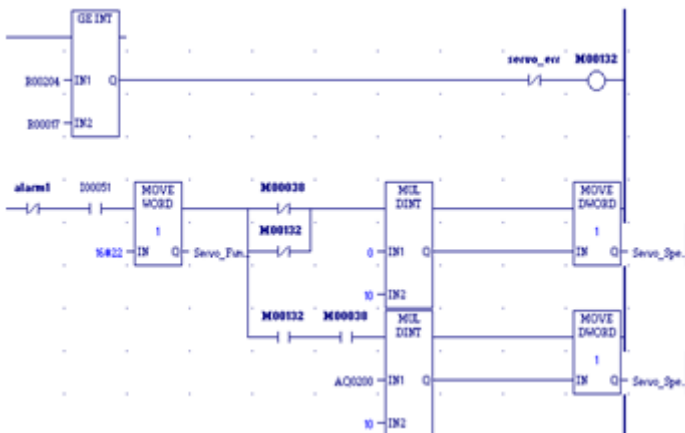
Fig. 6. The program of the automatic temperature control system

3.3   Lighting Control System

Lighting control system consists of the illumination sensor installed in the outside of bus body, the PAC for controlling host, and three sets of LED floodlights. The illumination outside the bus is monitored by illumination sensor in real time, and the intensity of the illumination outside the bus is reported by the illumination sensor: if the intensity is not less than the set value, the detection can go proceed until the value is less than the pre-set, and then light and shade should be decided and only a set of lights goes on in a dark background. If the illumination outside the bus is quite dark, two control logics are designed according to the number of the passengers inside the bus; a set of lights goes on if the number of the passengers inside the bus is less than 50% of the set number, and otherwise two sets will go on. If the illumination outside the bus is very dark, three logics are designed: a set of lights go on if the number of the passengers is less than 50%; two sets of lights go on if the number of the passengers is in 50%~ 80%; three sets of lights go on if the number of the passengers is more than 80%. Thus, the real-time adjustment of the lighting system is implemented and also the idea of energy conservation is reflected. The function flow of the lighting control system is as shown in figure 7, and the program design of the lighting control system is as shown in figure 8.
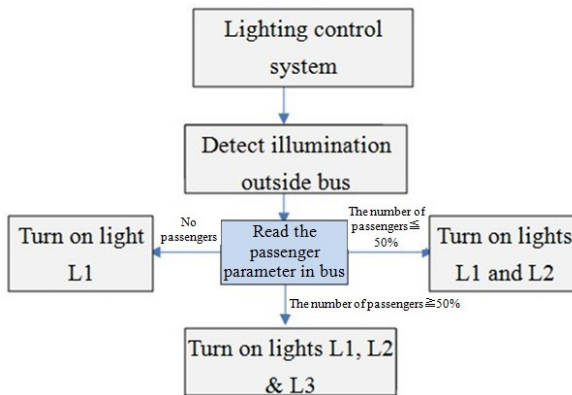
Fig. 7. The function flow of the lighting control system

3.4   Automatic Fire Escape System

Automatic fire escape system consists of smoke sensor, pressure sensor, temperature sensor, the PAC for controlling host and door, and parking enforcement agencies. The smoke sensor inside the bus is to detect the concentration of the smoke in the air: the outbreak of fire in a bus is comprehensively judged in combination with the reading number of pressure and temperature sensor if the concentration of the smoke in the air reaches a certain value: smoke detector alarms and sends voice prompts if there is an outbreak of fire, and the bus will automatically stop and open the escaping door after the tail lights flash for 5s, so that passengers can escape as soon as possible. The function flow of the automatic fire escape system is as shown in figure 9. The automatic fire escape program is as shown in figure 10.
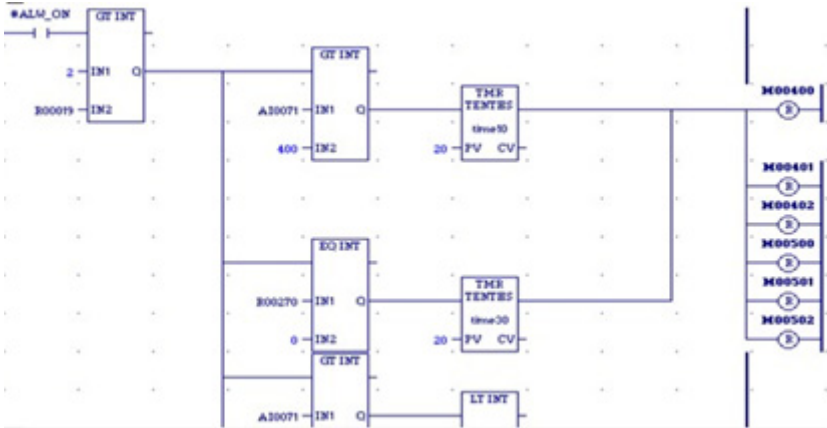
Fig. 8. Part of lighting control system's program design
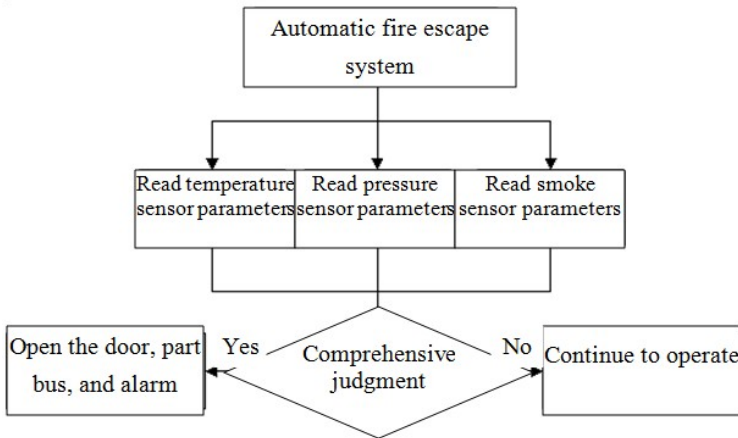
Fig. 9. The function flow of the automatic fire escape system

3.5   Energy Control Management System

Energy control system is made up of the solar receiver on the roof of bus, and the power supply conversion device, so that power supply is comprehensively controlled. The function flow of the automatic energy control system is as shown in figure 11. The automatic energy control program is as shown in figure 12.
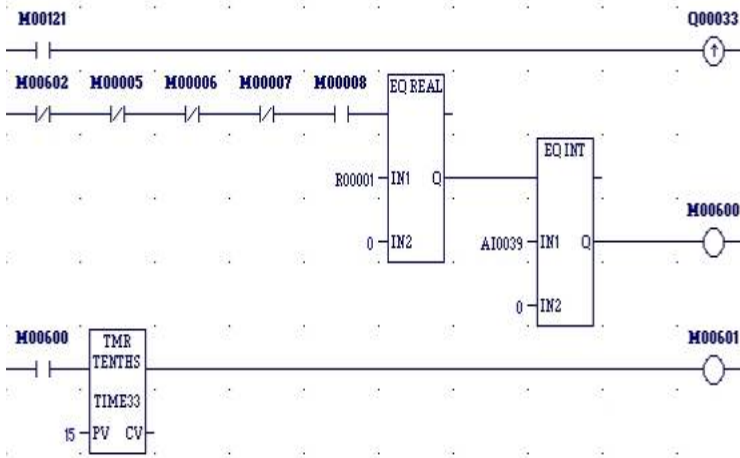
Fig. 10. The automatic fire escape program
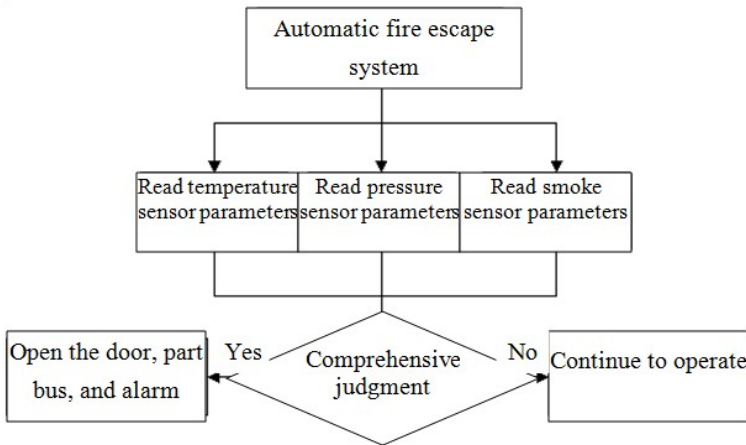
Fig. 11. The function flow of the automatic energy control system

3.6   Intelligent Dispatching System

Intelligent dispatching system consists of the passenger management system in bus, the Ethernet wireless communication technology, and part of the real-time monitoring system in bus station. According to the number of the passengers in bus and the number of passengers in waiting room, if a bus is added or reduced is decided comprehensively, so as to reasonably control the dispatching interval. Thus, intelligent control is achieved. The flow of the dispatching system is as shown in figure 13.

Fig. 12. The automatic energy control program
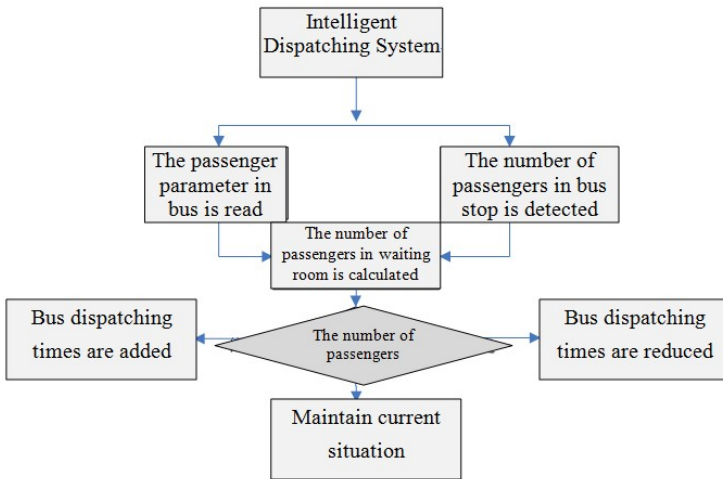
Fig. 13. The flow of the achieve dispatching system

The software system of the intelligent dispatching system is implemented based on GE iFIX configuration software. It is a set of industrial automation software, and provides users with a "process-oriented" window and the real-time data for operators and software applications.

After the program is started up, the system enters the main monitoring interface, as shown in figure 14. Then, click on the button of "Click to Enter" and enter the course monitoring and operation interface, as shown in figure 15.
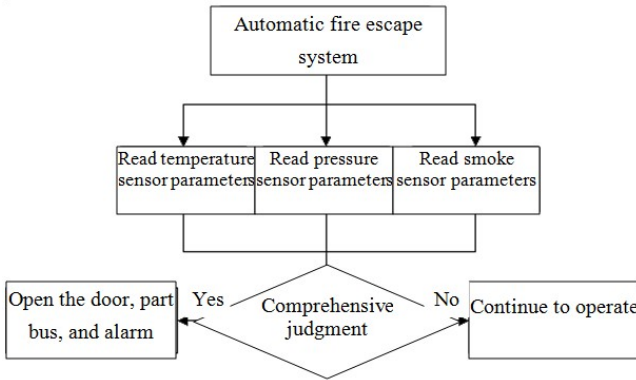
Fig. 14. iFIX main remote monitoring interface
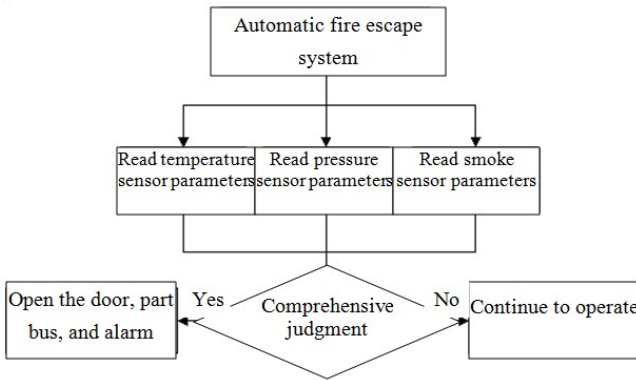
Fig. 15. The iFIX remote monitoring line and dispatching interface

4   Characteristics of the Bus-Mounted Intelligent Control System

The design and production of this bus-mounted intelligent control system have been completed, and also it has been tested successfully. Since the related technologies are developed for more than two years, this system has been in the stage of mature application through the practical application of multiple projects at present. This technology has won a utility patent of "PLC-based bus-mounted intelligent control system" and the software copyright certificate of the "bus-mounted intelligent control system" (registration No.2011SR071802). In the meantime, it received a special praise from the automation control design competition of GE intelligence platform for college students in 2010 ("intelligent trip, and moved from creativity) held by GE Intelligent Equipment (Shanghai) co., LTD.

References

1. Shen, X.: Use of Fire Auto-evacuation System for Public Transportation Vehicles Based on Controller of GE Company. Coal Technology (11), 226–227 (2011)
2. Pan, C.: Development of the Fire Auto-evacuation System for Bus Based on PAC System RX3i Controller. Manufacturing Information Engineering of China 40(7), 61–63 (2011)
3. Wang, W., Ma, H., Sun, G., et al.: Design and Development of Intelligent Public Transportation Terminal Device. Chinese Journal of Scientific Instrument 27(6), 666–668 (2006)
4. Liu, Y., Lu, C., Yu, M., et al.: Research on Ergonomics-based Design Method for City Bus. Machinery Design & Manufacture (9), 4–6 (2006)
5. Lv, Y., Zou, Z.: Design and Study on Hydraulic Transmission System of City Bus. Hydraulics & Pneumatics (11), 1–5 (2007)
6. Wang, S., Lin, B.: A GPS/GSM/GIS Based Intelligent Bus Monitoring System. Journal of Sichuan University (Natural Science Edition) 42(4), 710–713 (2005)
7. Jun, Z., Tan, Y., Ni, A.: Study on Microscopic Traffic Simulation Model of Public Transport Vehicles Operation. Journal of Highway and Transportation Research and Development 25(8), 119–122 (2008)
8. Wei, W., Chen, P., Chen, B.: Information Management System of Buses Based on OPS. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition) 17(2), 244–246 (2005)
9. Chang, X., Shen, L., Hu, J., et al.: Development and Performance Analysis of a Bus-mounted System in Intelligent Transportation. Journal of Southeast University (Natural Science Edition) 39(2), 377–383 (2009)
10. Wang, Y.: China's Urban Public Transport Problems and Reform. The Construction of China (02), 1–2 (2007)
11. Tan, Y., Hao, Z., Wang, L.: Study on Public Bus Dispatching with Single Yard. Railway Transport and Economy 30(11), 87–90 (2008)
12. Huang, K.: The Causes of Spontaneous Bus Combustion and the Prophylactic Measures. Guangdong Traffic (06), 43 (2008)
13. Bing, S., Zhao, Y., Pan, Y., et al.: Intelligent Fire Emergency Lighting and Evacuation Instructions Escape System. China Illuminating Engineering Journal (12), 23–29 (2004)
14. Yang, L., Zhuang, S., Ma, J., et al.: Discussion on the Application of Microwave Technology to Fire Detection. Fire Science and Technology (01), 77–79 (2006)

Track Algorithm for High Speed Target Based on Buffer-Diffluence Model

Liu Hongjian, Cai Zhongxiang, Wu Huan, and Lai Chuanlong

Institute of Geospatial Information, Information Engeering Univercity Zheng Zhou,
450052, Zhenzhou, China

Abstract. The Whole-Track algorithm for high speed target is important for target monitor application, this passage describe the theory of dynamic display, analyze the defect of common Whole-Track algorithm, implement the Whole-Track algorithm based on "Buffer-Diffluence" model, give a solution for the choke point in memory resource allocation of common Whole-Track algorithm.

Keywords: Whole Track Algorithm, "Buffer-Diffluence" Model, Dynamic-Display.

1   Introduction

High-Speed target of survey and control refers to flying targets like the unmanned aerial vehicles (UAV), helicopters, missiles, etc which is tracked, measured and controlled by ground Survery and Control Systrem [1]. Different from conventional moving targets, high-speed targets move much faster, frame rate of packets of target movement obtained by ground measure-control stations is higher, a single target frame rate can be as high as dozens of frames per second [2] . In the application of high-speed target real-time monitoring, the target of whole track real-time display refers to that from the starting point of monitoring, every frame of target position and attitude data should dynamically be displayed on the map, and updated in real time, the requirement of the display is that data frame is not allowed to be thrown away, the target line can't be simplified by means of the vector data compression algorithm [3]. Trajectory real-time display of targets is the core functions of high-speed target monitoring application, in applications of general GPS and beidou navigation monitoring, because of the low target frame rate, tracking points of a single target are controlled by threshold value (if the tracking points of targets are more than threshold, the first frame data will be discarded from the trajectory),the trajectory of the target is the display of part of the trajectory [4-8], it is true clear that the algorithm used in applications of the general GPS or beidou monitoring applications target trajectory does not adapt to application scenarios of the high-speed target whole trajectory display, therefore this paper proposes and implements a display algorithm of High-Speed target of survey and control trajectory based on "shunt cache" model, which performs better in solving all kinds of shortcomings of the general trajectory display algorithm, and has larger significance in practical applications.

2   The Dynamic Display Mechanism for High Speed Target of Survey and Control

The realization mechanisms of High speed dynamic target which are drawn in different GIS and its development platforms also differ from each other. This paper choose ESRI Arcgis Engine (ArcEngine) as the test bed. There are two ways to display high speed targets and their trajectory [9-12]:

(1) Dynamic display based on Graphics Layer. In this method, every dynamic target is a n element, render high speed targets by drawing elements dynamically and updating, which performs better in the effect of displaying, and can update part of the targets through the interface of partial update. The main problem is that in the process of updating targets, the scene of annotations on the geographic base maps moving around influences the effect of display.

(2) Dynamic target display based on Dynamic Display. In this way of dynamic display, every high speed target is a "Dynamic Graphic", which can support extremely high frame rate with quick rendering of targets. But map background is textured when realizing this technology, that is to say geographic base maps will be converted to pixel maps, which greatly reduces the effect display of base maps.
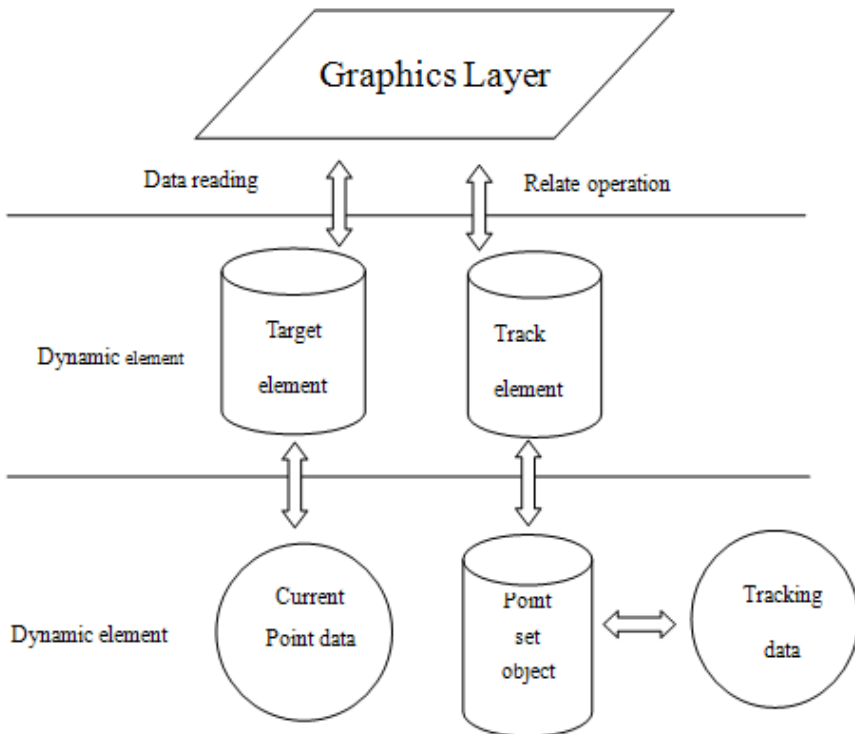
Fig. 1. Dynamic display mechanism based on Graphics Layer

Combining the advantages and disadvantages of the two methods mentioned above. In the application scene of display of high speed targets and its real-time trajectory, both the dynamic rendering of targets and the beauty of geographic base maps should be considered, while no influence on the annotations when using dynamic rendering. Therefore this paper chooses the method of dynamic display based on Graphics Layer, transforming annotations from the form of "Label" to the form of "Annotation", which solves the problem of the annotations moving around in order to avoid the high speed moving target.

Trajectory drawing of targets based on Graphics Layer, considering the dynamic target as an element, through organizing the Geometry(Geometry entities) and Symbol(display of symbol), adding them to the container of Graphics Layer. Graphics Layer is a peculiar map layer which is used in organizing graphs data dynamic data. During the process of dynamic display of targets, Graphics Layer connects with data of real-time dynamic targets through "Element". Two kinds of data need to be displayed in the process of dynamic display of high speed measure and control targets(1)data of targets, mainly described by the geographical coordinates, yaw angle, sidewinder angle and angle of pitch.(2)trajectory of targets, mainly described by the segment collection. Therefore, two dynamic elements need to be set: Object m_pElementTarget corresponds to the current data of point, while Object m_pElementTrack corresponds to the data of targets' trajectory.

3   Whole Track Algorithm for High Speed Target of Survey and Control Based on Buffer-Diffluence

3.1   Whole Track Algorithm of High Speed Target of Survey and Control Targets

When displaying high speed measure and control target and its real-time trajectory, Firstly we resolve the data package of maps, acquiring the current coordinates of target, then judge whether the current data frame is legal, and it will be treated as an outlier if not. We depend on the threshold of global distance and global time:(1)if the time interval between two frames before and after is more than global time threshold, we consider the frame illegal;(2)if the spatial distance between two frames before and after is more than global distance threshold, we consider the frame illegal.

After eliminating the outlier, it's feasible to draw targets dynamically. In terms of real-time element, we can make use of method of UpdateGeometry to update geometry data of Object m_pElementTarget, where the target will move to the place where the current data is. In terms of real-time trajectory of targets, two aspects should be considered: (1)If the current frame of map data is legal, add the current point of target into the object of point set of the current segment of trajectory element; (2) if the current frame of map data is illegal, in this situation connecting the point and its upriver is not allowed, They must be disconnected. More specifically, use the method of AddSegment, then open up a new segment on the Object m_pElementTrack, and add

the data frame of maps to the newly opened-up segment, which guarantee that trajectory data is a unified object while the trajectory breaks naturally at the outlier.

After the update of geometry data of element of target and its trajectory, we need refresh the screen through the refreshing mechanism of Graphics Layer, so all the data hooked by the trajectory element of targets could be displayed. It is visible that the position of target keep changing and the number of tracking points keeps increasing.
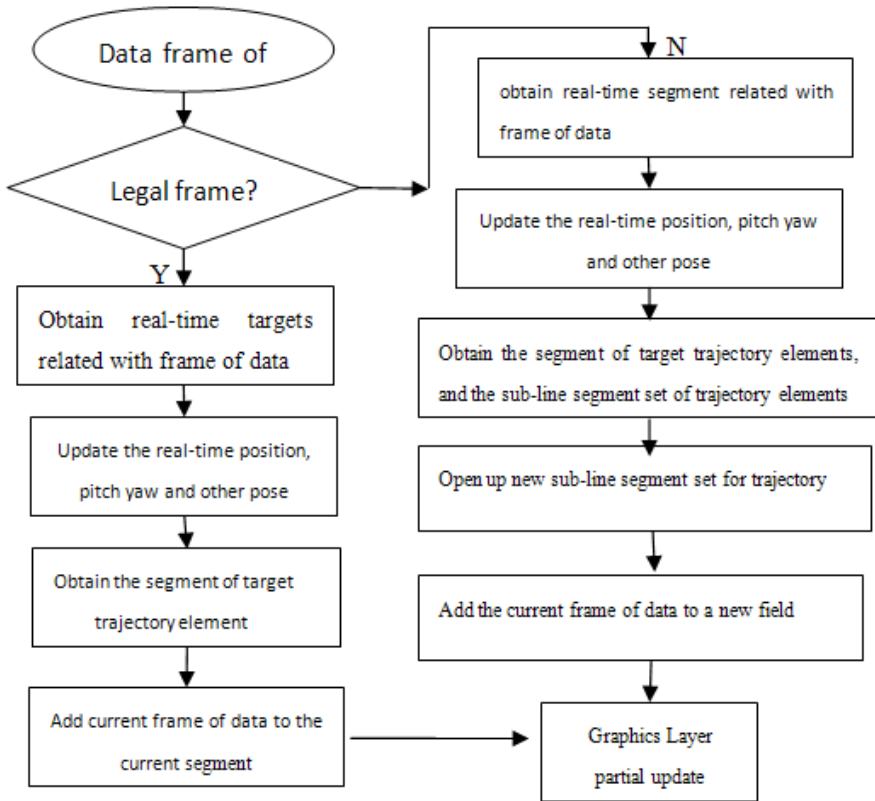
Fig. 2. Algorithm of whole track display

3.2   Limitation of the Algorithm

For general positioning and navigation monitoring systems, because of the low data frame, the number of points of target trajectory data is small. It is feasible to use this algorithm in displaying dynamic trajectory of target, while in displaying high speed measure and control target, if implementing tracking through UAV or helicopters, 12000 tracking points will be displayed considering receiving a target data frame per 50 ms. If continuing tracking for 3 hours, more than 200 thousand points will be displayed on the map.

Proved by real-time flight paths display in the tests of several flights of UAV/helicopter, algorithm referred above will set PF and CPU utilization rate rising fast with the increasing number of tracking points, which at last will cause the breakdown of system. After Analysis, The author has found that this algorithm lack the mechanism of controlling the number of points of target trajectory. When the number of points keeps increasing, data size of point position keeps increasing in the memory. In its essence, the algorithm lack the mechanism of controlling the possession of resource which is used by tracking point set in the memory, and it bring out a bug that memory already used by assigned point set cannot be reclaimed. Those test reflected great contradiction exists in the aspect of resource allocation:

(1) Ever-increasing number of element of target tracking point set makes the need of system memory continue to rise, while memory resource is always limited.

(2) With the increasing number of target tracking point set, processing time needed for updating trajectory geometry entities, after the author's test, 5ms is needed for updating trajectory geometry entities when the number of target tracking point set is 1200; while 853ms is needed for updating trajectory geometry entities when the number of target tracking point set is 30000. Processing time increases by more than 170 times, which highlights the contradiction of resource possession.

The author attempted to improve the algorithm above to eliminate the defects. For example: Algorithm of partial update, every time only the rectangular area formed by current position and its upriver data frame will be updated, while other data points of trajectory stay the same, thus helps improve the performance of algorithm, but still cannot solve the problem of CPU updating geometry entities formed by numbers of tracking points.

Only display current point using Graphics Layer, save the growing numbers of target tracking points to a vector map layer, then add this layer to the geographic base maps to display the trajectory of targets. Therefore the memory resource needed to save tracking points of targets turns to vector map layer's need of outer memory. This idea could solve the problem of CPU updating geometry entities formed by numbers of tracking point, but because of the limitation of the internal mechanism of ArcEngine, vector map layer loaded in memory is not allowed to be edited at the same time. If the idea above is to be realized, trajectory map layer will be unloaded when new frame of target data is added. Editing interface that load map layer of target trajectory after the process of adding current frame of data. This cost is unacceptable to reach the demand of real time.

3.3   Buffer-Diffluence Model

The idea of Buffer-Diffluence Model is as follows

(1) Establish a cache based on target, and   build a sequence of cache where frame of target data for each high speed measure and control target, allocate the frame of data with cache in the principle of first-in first-out (FIFO)

(2) Cache size of single target is controlled by the threshold value of global cache. Through the control of threshold value, make sure that the frames of data loaded into

memory at the same time will not make the system overloaded. The threshold value of cache as the global parameter could be adjusted according to the concrete machine configuration, the author's computer configuration: Intel Core i5 CPU, 2.40GHz, 2.86GB memory, WinXP operation system. Set threshold value of global cache as 6000, that is to say the largest storage of frame is 6000 frames for a single trajectory of target.

(3) Make use of vector map layer to "distribute" dynamic data. In the process of initializing of the algorithm, vector map layer named TargetTrack of target trajectory will be established, if   the frame of data of the trajectory of high speed measure and control target doesn't reach the upper limit of global threshold, target and its trajectory will be drawn on the Graphics Layer. While when the upper limit is exceeded, first unload the map layer of target trajectory TargetTrack, then use the interface of IWrokspaceEdit to edit the map layer of target trajectory TargetTrack, in order to write all the data points of targets' cache into the map layer TargetTrack, then achieve the effect of distributary by converting memory cache to a disk-based cache or the other way around. In the process, Graphics Layer display the current position of current point, as well as become a cache to save the newest track point, and the focus of algorithm is here.

(4) The mutex access of frame of target trajectory in the cache sequence. In the process of importing data of trajectory of target to map layer TargetTrack, The system must make sure that the thread of the real-time data of target mutexes the dump thread of cache, that is, when the thread of cache dumping, the writing thread of cache data must be blocked, dumping data while writing the frame of target data at the same time is not allowed. In the concrete realization, To guarantee the mutex access of cache data, it is feasible to use "semaphore" or "critical code section" to "lock" the sequence of cache.

(5) After the dumping of data in the cache of target, clear the region of target cache, then load the map layer TargetTrack into memory again.

The summarization of core idea is as follows: The sequence of target cache is used to save the newest position of tracking point. When the number of frames of data in cache sequence reach the upper limit of global threshold value of cache, first save all the data into a temporary map layer TargetTrack, if the map layer is loaded in the memory, unload it and dump the data of cache sequence, then reload the map layer, finally clear the sequence, free the memory. In this way, although the number of tracking point keeps increasing,but the number of frames of data in the cache sequence always stays below the load of CPU and memory. In a word, by replacing the need of cache sequence of target to memory with the use of outer memory through TragetTrack, solving the problem of only distributing without reclaiming. This idea could solve the main problems in the conventional whole track algorithm.

4   Whole Track Algorithm Based On Buffer-Diffluence Model

The flow-process diagram of algorithm is as figure 3

(1) The check of availability of frames of target data. If frames are legal, the position of point is in the view window of current active map and not the same point compared with the upper frame, The real-time position of target will be updated.

(2) Forming the trajectory of target. Add the frames of measure and control target data to the corresponding cache of trajectory in proper order, forming the trajectory line of current time.

(3) The display of elements of target and its trajectory. Acquire the geographic range, implement partial update for elements of target and its trajectory with partial mechanism in Graphics Layer.

(4)Distributary of dynamic data. Judge whether the data of tracking point in the cache sequence of target data has reached the threshold value of global cache. If so, unload the existing TargetTrack map layer or create a new vector map layer TargetTrack when no map layer like this exists, at the same time save the points in the sequence of cache. At last, load the map layer of target trajectory on the geographic base map.
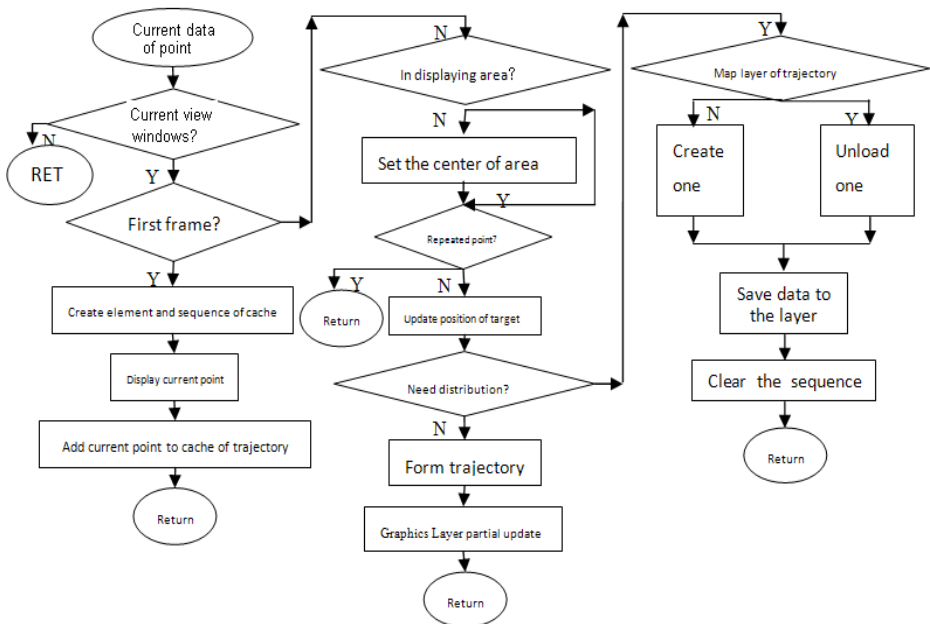
Fig. 3. Algorithm flow chart

In the realization of this algorithm, two timer is used, one is real-time updating timer (as figure 5), mainly in charge of updating the attribute of Geometry entities of element of target and its trajectory. The other one is update timer, mainly in charge of implementing partial update to Graphics Layer. In order to meet the demand of real time, the interval of data updating keeps pace with the frame rate of target data, while the interval of update timer may be 3~5 times of the minimum interval.
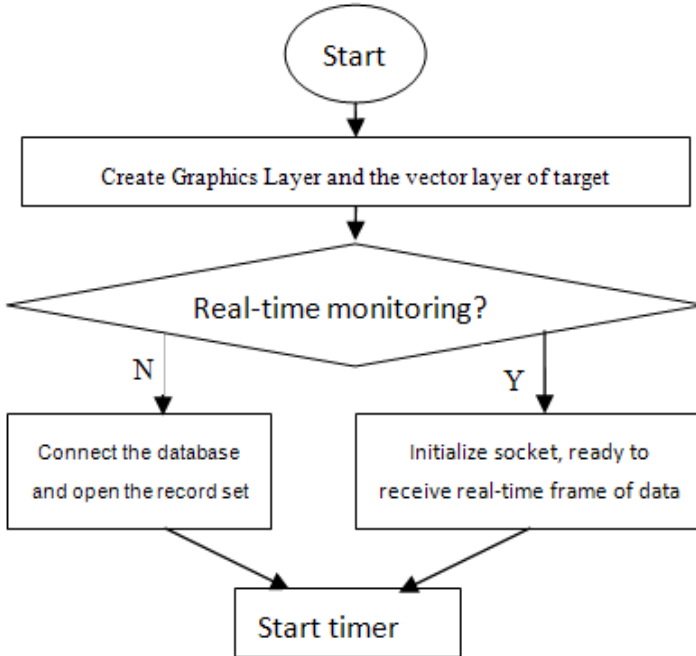
Fig. 4. Upstream process of Initialization of algorithm

In the same condition of test, comparison between the conventional algorithm and Whole Track Algorithm Based On"Buffer-Diffluence", the results are as follows:

(1) Conventional algorithm continued to work for 15 minutes, displayed 18000 tracking points, the usage rate of PF rose from 184M at start to 2.1G in the end, almost exponentially, and the system will be broken down at last.

(2) Whole Track Algorithm Based On "Buffer-Diffluence" continued to work for 8 hours, displayed 576000 tracking points of target, the usage rate of PF rose from 184M at start to 267M in the end, rising gently, The CPU kept 5% when working with this algorithm. It is clear that   Whole Track Algorithm Based On "Buffer-Diffluence" has improved a lot in the performance of display, from the average memory consumed by every frame, about 1/252 what Conventional algorithm need to spend.
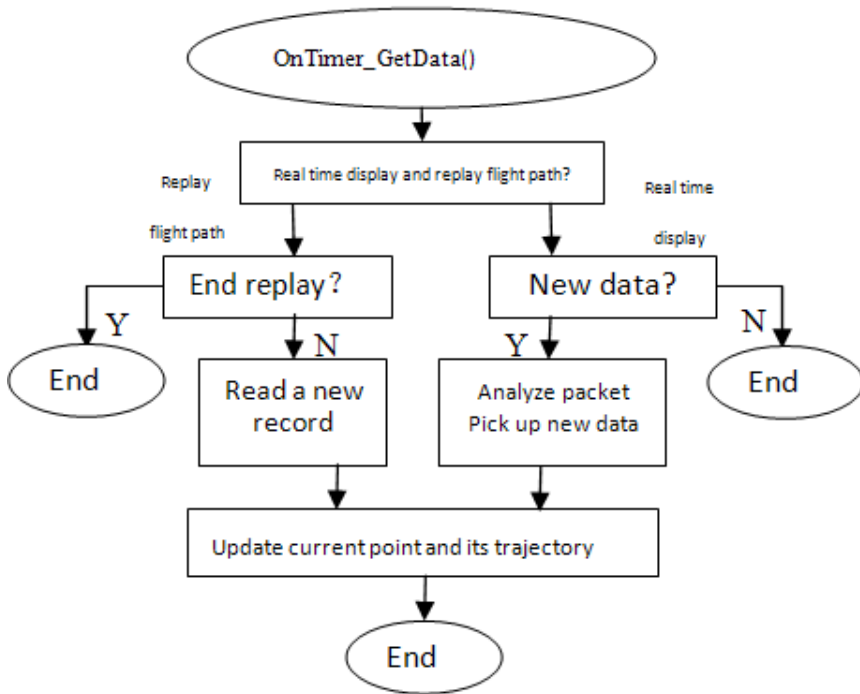
Fig. 5. Work process of data updating timer

5   Conclusion

Whole Track Algorithm for High Speed Target of Survey and Control is the important content of real-time Target of Survey and Control monitoring research and application, this paper makes research on the conventional Whole Track Algorithm, and analyzes its limitations. Put forward and realize the Whole Track Algorithm based On "Buffer-Diffluence" Model, which proved to be better in practical application.

References

1. Sheng, C.: Conventional equipment test. National industry press, Beijing (2007)
2. Xue, Y.: Methods and application of Conventional equitment test in intersection measurement. Xi'an Jiaotong University Press (2010)
3. Tang, G., Zhao, M.: Geographical information system. Science Press, Beijing (2000)
4. Lun, W., et al.: geographical information system. Science Press, Beijing (2000)
5. Feng, Z.: Application of Dynamic Track Drawing Arithmetic Used in Ship Monitoring system (5), 68–71 (2008)
6. He, S., Wen, F.: On Application of GSM/GPS to Public Transportation, vol. (5), pp. 35–37 (2003)

7. He, Y.B., Fan, S.: Updating system of moving objects database. Computer Engineering and Applications 44(27) (2008)
8. Song, S., Jiao, J.: Improved Algorithm of the Real-time Trajectory Dsiplay of Remote Sensing UAV Based on MapObjects. Remote Sensing, 12 (2007)
9. Zheng, G., Yan, P., Ding, M., Sun, F.: National defence industry press 12 (2008)
10. Crino, J.R.: A Group Theoretic Tabu Search Methodology for Solving Theater Distribution Vehicle Routing and Scheduling Problems. Air Force Institute of Technology (2012)
11. Wiley, V.: The Aerial Fleet Refueling Problem. The University of Texas, Austin (2011)
12. Jia, Y., Gu, H., Xi, Y.: Quick taboo search algorithm for solving PDPTW Problems. Control and Decision 19(1), 57–60 (2004)

Research on Monitoring and Early Warning
of Sustainable Development of Regional Economy

Lin Lin

School of Economics and Management, Wuhan University, Wuhan 430072

Abstract. The paper makes comparative research on common index systems, comprehensively uses systematic analysis method, frequency statistics method and theoretical analysis method to explore, and establishes early-warning index system and early-warning system of sustainable development of regional economy concerning economy, society, environment, resource and 21 indexes. And the paper takes Zhejiang as an example, and collects the relevant data from 1996 to 2011 to make empirical study and predicts that developing sustainability degree of Zhejiang is in the state without policeman in the next few years. But the sustainable development doesn't achieve the best level. Based on the predicting results and combined with obfuscation rules, the paper proposes specific adjustment direction.

Keywords: regional economy, sustainable development, self-organize data mining, early-warning system.

1 Introduction

Sustainable development system of regional economy is the core subsystem of regional sustainable development system (a compound system consisting of three interactive and interdependent subsystems including economy, society and ecology) [1-3]. Its healthy operation plays an essential role in realizing regional sustainable development. During the "Twelfth Five-year Plan" period, according to the requirements of the country implementing sustainable development strategy and the trend of regional sustainable development study, economic system is selected as the research focus to explore the construction of sustainable development early-warning system of regional economy, which extends the theoretical view of sustainable development research [4]. The paper combines the economic development status of Zhejiang to make empirical study, which is good exploration on studying the field in Zhejiang. And the early-warning results provide reference for relevant decisions.

2 Establishment of Early-Warning System of Sustainable Economic Development in Zhejiang

2.1 Specifying Warning Source

According to the steps of establishing early warning system of sustainable development of regional economy, the step of specifying risk source includes the determination of

early warning indicator s of sustainable development and the collection of relevant data, dimensionless treatment of indicators and determination of index weight [5].

Analytic hierarchy process factor minor empowerment data method is used to calculate the weight of various indicators of early warning system of sustainable economic development in Zhejiang. And the concrete calculation process is as follows.

The objective of the paper is to study sustainable economic development at present and in the future in Zhejiang. There are many factors influencing sustainable economic development in Zhejiang [6, 7]. From economic structure, economic scale, economic benefit, energy consumption and life quality, the paper observes sustainable economic development in Zhejiang. There are 23 evaluation indexes in index level. Hierarchical structure is established and divided into objective layer A, criteria layer B and index layer C.

Statistical software SPSS is used. Normal P-P is drawn to realize.

X1 GDP is taken as an example. The original data is imported into SPSS worksheet. Normal P-P process of Graphs is called, and the achieved results are as shown in Fig. 1. And Fig. 2.
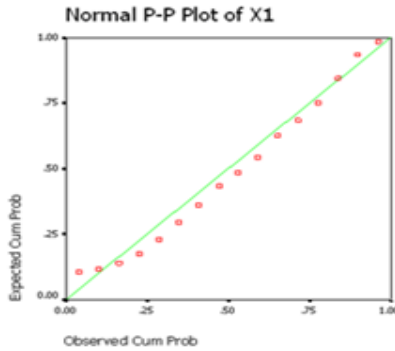
Fig. 1. X1 Nomal P-P

Fig. 2. X1 Normal P-P without trend

Figure 1 shows that X1 is close to Normal P-P line, and Figure 2 indicates that the hypothesis that X1 disobeys normal distribution is false, from which we can see that X1 obeys normal distribution.

The checkout procedure of the original data for the other 22 indicators is the same to it.

After checkout, the data after dimensionless treatment receives the calculation of the Weight. Index X20, X21, X22 and X23 are taken as an example, and the function is used,

$$y_{s4} = x_{s20}^{a_{20}} \cdot x_{s21}^{a_{21}} \cdot x_{s22}^{a_{22}} \cdot x_{s23}^{a_{23}} \quad s = 1, 2, 3, \cdots 16 \tag{1}$$

(s means the number from 1996 to 2011, and the number of 1996 is 1)

Index X20, X21, X22 and X23 in index layer C in the formula correspond to the weight of criteria layer B. And X20, X21, X22 and X23 are the value of data after dimensionless treatment corresponding to component. The way of acquiring the value of y_{s4} is that arithmetic mean of X20, X21, X22 and X23 in each year are achieved firstly, the interval consisting of the minimum and maximum of arithmetic mean is divided into 9 parts according to the length. After calculation, the minimum of arithmetic mean of X20, X21, X22 and X23 is 0, and the maximum is 0.9702. After the division based on the above principle, the results are shown in Table 1.

Table 1. Interval value division and interval value

Interval value of arithmetic average	0~0.1078	0.1078~0.2156	0.2156~0.3234
Interval number	1	2	3
Interval value of arithmetic average	0.3234~0.4312	0.4312~0.539	0.539~0.6468
Interval number	4	5	6
Interval value of arithmetic average	0.6468~0.7546	0.7546~0.8624	0.8624~0.9702
Interval number	7	8	9

According to Table 1, the arithmetic averages of each year are observed. The results are shown in Table 2.

Table 2. Value of y_{s4}

Value	$x_{s20}, x_{s21}, x_{s22}, x_{s23}$ arithmetic averages in every year	Value of y_{s4}
1990 (s=1)	0.0000	1
1991 (s=2)	0.1255	2
1992 (s=3)	0.2755	3
1993 (s=4)	0.5166	5
1994 (s=5)	0.6479	7
1995 (s=6)	0.7172	7
1996 (s=7)	0.7806	8

Table 2. (*continued*)

1997 (s=8)	0.8309	8
1998 (s=9)	0.8627	9
1999 (s=10)	0.8848	9
2000 (s=11)	0.8947	9
2001 (s=12)	0.8859	9
2002 (s=13)	0.8982	9
2003 (s=14)	0.9206	9
2004 (s=15)	0.9372	9
2005 (s=16)	0.9702	9

Two sides of (1) take natural logarithm at the same time and are transformed into linear regression, $\ln y_{s4} = a_{20} \ln x_{s20} + a_{21} \ln x_{s21} + a_{22} \ln x_{s22} + a_{23} \ln x_{s23}$.

$\ln y_{s4}, \ln x_{s20}, \ln x_{s21}, \ln x_{s22}, \ln x_{s23}$ is imported into SPSS worksheet for multiple linear regression analysis. And the results are shown in Fig. 3.

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.230	.027		81.694	.000
	LNXS20	.795	.561	.797	1.417	.187
	LNXS21	1.032	.540	1.153	1.912	.085
	LNXS22	.163	.144	.284	1.129	.285
	LNXS23	-.927	.364	-1.239	-2.545	.029

a. Dependent Variable: LNYS4

Fig. 3. Regression coefficient

Calculating the weight of X20, X21, X22 and X23 corresponding to criteria layer B. And $\dfrac{|a_{20}|}{|a_{20}|+|a_{21}|+|a_{22}|+|a_{23}|}$, $\dfrac{|a_{21}|}{|a_{20}|+|a_{21}|+|a_{22}|+|a_{23}|}$, $\dfrac{|a_{22}|}{|a_{20}|+|a_{21}|+|a_{22}|+|a_{23}|}$ and $\dfrac{|a_{23}|}{|a_{20}|+|a_{21}|+|a_{22}|+|a_{23}|}$ are represented by $W_{20}, W_{21}, W_{22}, W_{23}$. The results are shown in Table 3.

Table 3. Weights of X_{20}, X_{21}, X_{22} and X_{23}

W20	W21	W22	W23
0.2725	0.3538	0.0559	0.3178

The weights of other indexes corresponding to criteria layer B can be calculated by following the above calculation method.

Table 4. Weight of each index in index layer C

W_{20}	W_{21}	W_{22}	W_{23}
0.2725	0.3538	0.0559	0.3178

Using formula 2:

$$z_S = y_{s1}^{\beta_1} \cdot y_{s2}^{\beta_2} \cdot y_{s3}^{\beta_3} \cdot y_{s4}^{\beta_4} \cdot y_{s5}^{\beta_5} \quad s = 1, 2, 3, \cdots 16 \tag{2}$$

(s means the number from 1996 to 2011,and the number of is 1)

The step (2) is repeated to calculate the weight Ui of each criterion in criterion layer B corresponding to objective layer A, and the method is the same to the above method. The results are shown in Table 5.

Table 5. Weights of criterions in criterion layer B

Name of criterion layer	Weight value (UI)
Y1 economic structure	0.125060
Y2 economic scale	0.312289
Y3 economic benefit	0.313976
Y4 energy conusumption	0.073494
Y5 life quality	0.175181

2.2 Dividing Warning Degree

(1) Economic sustainable development degree of each year in Zhejiang

According to the results of dimensionless treatment and the weights of indicators, formula (3) is used to calculate regional economic sustainable development degree of each year, as follows.

Sustainable development of regional economy=SUM(fraction of Yi×weight of Yi) (3)

2011 is taken as an example. And the calculation process of economic sustainable development degree in Zhejiang and the relevant data are as shown in Table 6. And

Fraction 1 is the data of indicators after dimensionless treatment in 2011.

Wi is the weights of indicators in index layer C.

Fraction 2= fraction 1×Wi, and fraction 3 is the sum of fraction 2 of the same category.

Ui is the weight of criterions in criterion layer B.

According to Formula 3, we can get ESDI of Zhejiang in 2011 =SUM(fraction 3   × corresponding Ui)=0.5781×0.12506+0.5321×0.312289+0.8104×0.313976+0.9732×0.073494+0.4688×0.175181=0.64654.

According to the same method, we can get ESDI of Zhejiang from 1996 to 2011, as shown in Table 7.

Table 6. Computation sheet of fractions of each index

Category	Name of indexes	Fraction 1	Wi	Fraction 2	Fraction 3	Ui
Y1 economic structure	X3 per capita agricultural output value	0.5615	0.3781	0.2123		
	X4 Per capita industrial output value	0.6247	0.2433	0.152		
	Percentage of X8R&D to GDP	0.6892	0.1958	0.1349	0.5781	0.12506
	X13Percentage of tertiary industry to GDP	0.52	0.1094	0.0569		
	X14 Percentage of environmental investment to GDP	0.3	0.0734	0.022		
Y2 economic scale	X1 GDP	0.5896	0.5697	0.3359		
	X9 Steel output	0.3675	0.2273	0.0835		
	X10 Total investment in fixed assets	0.5944	0.0596	0.0354	0.5321	0.312289
	X11 Total export	0.5958	0.0208	0.0124		
	X15 Urban infrastructure investment	0.5973	0.0658	0.0393		
	X17 Local fiscal expenditure	0.4496	0.0568	0.0255		
Y3 economic benefit	X2 Per capita GDP	0.8432	0.4122	0.3476		
	X5 Social labor productivity	0.3754	0.0723	0.0271	0.8104	0.313976
	X7 Per capita revenue	0.8451	0.5155	0.4356		
Y4 energy consumption	X20 Industrial solid waste production per GDP	0.9732	0.2725	0.2652		
	X21 Industrial wastewater waste production per GDP	0.9948	0.3538	0.352	0.9732	0.073494
	X22 Industrial waste gas discharge per GDP	0.9619	0.0559	0.0538		
	X23 Energy consumption per GDP	0.951	0.3178	0.3022		
Y5 life quality	X6 Total retail sales of consumer goods	0.3559	0.3920	0.1395		
	X12 Inflation rate	0.4343	0.0246	0.0107		
	X16 Consumer price index	0.25	0.1680	0.042	0.4688	0.175181
	X18 Urban and rural residents saving deposits	0.6175	0.3584	0.2213		
	X19 unemployment rate	0.9677	0.0571	0.0553		

Table 7. ESDI of Zhejiang in each year

Year	ESDI	Year	ESDI
1996	0.031235	2004	0.286647
1997	0.049179	2005	0.318528
1998	0.082875	2006	0.352951
1999	0.111094	2007	0.367793
2000	0.150537	2008	0.402612
2001	0.198493	2009	0.465116
2002	0.253027	2010	0.568942
2003	0.28389	2011	0.646540

(2)Warning division of ESDI in Zhejiang

1 Cluster analysis for dividing warning degree

SPSS software is used to make Q-type analysis of ESDI in Zhejiang from 1990 to 2005. And the clustering process and results are as follows.

Table 8. Analysis of Q-type clustering results

Category	Year	Minimum	Maximum
1	1996-2000	0.031235	0.150527
2	2001-2004	0.198493	0.286647
3	2005-2009	0.318528	0.465226
4	2010	0.568942	
5	2011	0.646540	

According to table 8, as index interval is continuous and ESDI is the number between 0 and 1, the warning degree   of ESDI in Zhejiang is determined, as follows, (0.00,0.03] , (0.03,0.17], (0.17,0.30],   (0.30,0.52], (0.52,1.00).

2 Dividing warning degree of ESDI in Zhejiang

According to warning division results and the theory in 2.3.2, ESDT of Zhejiang can be divided into five warning degree levels, and each warning degree level corresponds to different sustainable development. When there is no warning, sustainable development can be divided into two levels, good and very good. And the specific division is shown in Table 9.

Table 9. Division of sustainable development and warning degree grade of ESDI in Zhejiang

ESDI	Warning degree grade	Sustainable development	Sustainable or not sustainable
(0.00,0.03)	Heavy warning	Very bad	Not
(0.03,017)	Severe warning	Bad	
(0.17,0.30)	Meduim warning	Worse	
(0.30,0.52)	Light warning	Better	Yes
()0.52,0.80	No warning	Good	
(0.80,1.00)		Very good	

3 Warning degree grade of ESDI of Zhejiang from 1996 to 2011

From the data of ESDI of Zhejiang in each year and Table 9, we can get warning degree grade and sustainable development of Zhejiang from 1996 to 2011, as shown in Table 10.

Warning degree grade and sustainable development of ESDI of Zhejiang in each year

Table 10. Warning degree grade and sustainable development of ESDI of Zhejiang in each year

Year	Warning degree grade	Sustainable development	Year	Warning degree grade	Sustainable development
1996	Severe warning	Bad	2004	Medium warning	Worse
1997	Severe warning	Bad	2005	Light warning	Better
1998	Severe warning	Bad	2006	Light warning	Better
1999	Severe warning	Bad	2007	Light warning	Better
2000	Severe warning	Bad	2008	Light warning	Better
2001	Medium warning	Worse	2009	Light warning	Better
2002	Medium warning	Worse	2010	No warning	Good
2003	Medium warning	Worse	2011	No warning	Good

3   Implementing Economic Sustainable Development Strategy in Zhejiang

As a part of complicated system of sustainable development, economy has close relationship with natural and social system. As there is difference for each economic subject in space and time , regional economic activity is highly complicated, and there are a series of problems and contradictions in the development process. To solve these problems, it needs to realize economic sustainable development and singly use some means including economic means, legal means and technical means to achieve the objectives, which needs system control.

Therefore, implementing economic sustainable development strategy in Zhejiang must formulate strategic objectives and corresponding measures. In the implementation process, we must focus on priorities to achieve an ideal optimal combination among economic development, social progress and resource environment, which can realize economic sustainable development in Zhejiang.

Optimizing urban form layout. It needs to quicken central business district and greenbelt construction, accelerate shading cooperation and function adjustment of central city area, speed up urbanization area construction in suburbs, and six suburbs should be on the basis of the requirements of constructing medium-sized cities. And planning should be in accordance with high starting point and high standard. And it needs to rationally expand urban area and population scale and construct corresponding municipal infrastructures and social construction.

Quickening industrial distribution adjustment. Urban centers should attract various services at home and abroad and improve sophisticated functions. The suburbs need to drive on constructing distinctive productive service industry agglomeration areas.

Increasing strategic adjustment of industrial structure, forming new economic growth point of supporting the city's economic development, and gradually improving overall efficiency and anti-fluctuation capability of economic development. It needs to manage large enterprises well while relaxing control over small ones to promote to optimize organizational structure of the enterprise.

Quickening the construction of agricultural modernization. Enhancing overall agricultural production capacity, improving agricultural infrastructure, improving agri-organizations and marketization level, improving system of rural cooperative economic organizations and implementing agricultural brand strategy. And it needs to improve the policies supporting country and agricultural development. According to the distribution, focus and product structure of agricultural development, different agricultural investment supportive policies are   determined, investment and financing means are innovated and all kinds of funds are attracted to participate in suburban construction. It needs to establish long-term mechanism which makes farmer increase income, persist in the policy of giving more taking less and loosening control, and improve income level of farmer.

We should adhered to the basic national policy of family planning, stabilize low birthrate of population, establish new mode of population management which are suitable for the characteristics of large cities in Zhejiang, keep bearing capacity of population size and urban resources and environment adjusting with economic and social development level.

Implementing the strategy of science and education encouraging city, developing the advantage of science, technology and talents in Zhejiang, promoting the science and technology to face market, establishing the mechanism combining scientific research, development, production and market, and ensuring that scientific-technical progress becomes active force of economic development in Zhejiang.

Reforming and adjusting the existing economic and financial incentives, which makes it accord with sustainable development objectives.

Establishing green GDP accounting system and price system according with the objectives of sustainable development.

Perfecting systems and mechanisms of environmental protection and management, enhancing environmental protection law enforcement and supervision, perfecting environmental quality monitoring system and key pollution sources on-line monitoring network. And strictly implementing environmental impact assessment.

Strengthening capacity construction of sustainable development., specifying the construction of policy system, laws and regulations system and index system of strategic targets, and improving sustainable development consciousness and implementing competence.

4   Summary

Based on index system of sustainable development and economic early-warnings of sustainable development at home and abroad, and combined with the characteristics of regional economic environment in our country, the paper establishes early-warning index system of sustainable development of regional economy, and economic sustainable development

early-warning system in Zhejiang, evaluates sustainable development and warning degree of ESDT in Zhejiang from 1996 to 2011, predicts ESDI of Zhejiang from 2012 to 2016, uses systematic analysis method, frequency statistics method and theoretical analysis method to screen out sustainable development early-warning index of regional economy and establishes early-warning index system concerning four categories. When index weight is determined, hierarchical analysis factor seating arrangement method is used, and expert decision is replaced by arithmetic mean value assignment method. From pure quantitative point, the paper gets the sorted results and transforms the regression coefficients into weight. On determining warning degree limitation, the paper uses Q-type clustering analysis method and SPSS statistical software to realize the process. The paper establishes the database of sustainable development early-warning index, uses complete data mining software to establish the operation process of sustainable development early-warning system of regional economy, and simplifies early-warning process, which realizes the warning and prediction of economic sustainable development in different regions more efficiently and accurately.

References

1. Wu, N., Zhang, L., Ning, B.: Study on formation mechanism of resource-based enterprises core competitiveness. Gold (05) (2010)
2. Li, Y., Weng, M., Yang, C., Chen, Y.: Study on constraints and countermeasures of sustainable development for resource-based enterprises in China. Chinese Population, Resources and Environment (S1) (2010)
3. Fang, Z.: Analysis on sustainable development ways of resource-based enterprises. Productivity Research (07) (2010)
4. Ye, L., Yi, Z.: Selection of resource-based enterprises transformation models based on sustainable development. Resources & Industries (02) (2009)
5. Ye, L., Yi, Z., Jihui, G., Lu, P.: Study on transformation of resource-based enterprises. Resources & Industries (04) (2009)
6. Ke, X., Yang, C., Zhu, K.: Study on evaluation methods of core competence for resource-based enterprises. Mining Research and Development (04) (2009)
7. Chen, M., Wang, H.: Study on countermeasures of sustainable development for resource-based enterprises. Theory Monthly(11) (2009)

An Enhancing K-Means Algorithm
Based on Sorting and Partition

Yin Jun-wei, Chen Jian-ming, Xue Bai-li, and Zhang Jian

Department of Computer Science and Technology,
Soochow University, 215006, Suzhou, China
`yin1junwei7@163.com`

Abstract. The accuracy and efficiency as the two main evaluation indexes for k-means algorithm are influenced by the choice of initial clustering centers and the partition method of data points. In this paper, in view of the deficiency of direct k-means algorithm which chooses initial centers randomly, we propose a novel method about initial clustering centers based on sorting and partition and apply it to real data as well as simulated data, which show that this is an efficient method to improve the clustering accuracy and efficiency.

Keywords: k-means, sorting and partition, clustering analysis, initial centers.

1   Introduction

Clustering is a process of grouping data objects into disjointed clusters so that the data in the same cluster are similar, yet greatly different from those in others. Recently, clustering techniques are applied in many application areas, such as data mining, pattern recognition, statistics, artificial intelligence, image processing and information retrieval. There are many clustering methods, which can be divided into the following categories depending on the method used: Partition-based, Hierarchy-based, Density-based, and Model-based. In this paper we concentrate on partitional clustering, and in particular a popular partitional clustering method called K-means clustering.

In many practical applications, K-means algorithm has been proved to be effective, and can make satisfactory clustering result. However, the traditional K-means is a local optimization strategy, which is easy to fall into local optimization solution [1] and sensitive to initial centers. Besides, the K-means algorithm execute by a large number of iterations so that it is inefficient. Therefore, in consideration of the deficiency of the direct K-means algorithm, this paper propose a novel method to choose the initial clustering centers based on sorting and partition, which is simple and easy to be implemented. The experiment on real data and simulated data shows that this kind of methods can effectively improve the clustering accuracy and efficiency.

This paper is organized as follows: section 2 first introduces some related studies, and then describes the K-means algorithm; in section 3, we propose our algorithm which is based on sorting and partition, and make a detailed analysis. The experimental results, along with the data sets, are discussed in section4. Conclusions are drawn at last.

2   Related Works

2.1   Related Study

One of the earliest references to improve the K-means was by Forgy, after that, researchers and experts proposed many different improved methods from two directions: the choice of initial clustering centers and the iterative optimizing method. In the first direction, there are some enhancing methods as follows. MacQueen introduced a method that selects the first K points as the initial centers. Tou and Gonzales suggested the Simple Cluster Seeking (SCS) method [2-5]. Linde et al. proposed a Binary Splitteing (BS) [3]. Babu and Murty published a method of near-optimal seed selection using genetic programming [4]. Huang and Harris proposed the Direct Search Binary Splitting (DSBS) method [6-8]. Recently, Khan and Ahmad have described their Cluster Centre Initialization Method (CCIA) [6]. In the other direction, there are three classic improved algorithms, the reference [9, 10] suggested a Fixed Object Algorithm (FOA), the reference [8, 11] proposed a Integrating Genetics Into K-means (IGKM) and the reference [9, 13, 14] presented a Filter Algorithm (FA).

Those algorithms mentioned above, either just improved the selection of initial centers or just optimized the time complexity, and the most important is that the implement of those algorithms are very complex. By the revelation of the reference [15-18], we propose the method based sorting and partition in the selection of initial cluster centers, which can discretely select the initial centers so as to achieve a better cluster result and to make convergence faster.

2.2   Direct K-Means Algorithm

Direct K-means algorithm is a kind of hard clustering method, in which n data points are partitioned into k clusters in d-dimensional space R^d [19, 20]. This algorithm consists of two separate phases: the first is to define the cluster number k, one for each cluster, and randomly select k data objects as the initial center. The next phase is to take other objects to the nearest center according to their distance from each center, and then recalculate the cluster center, repeat this until convergence criterion is met. The description for K-means clustering is listed as algorithm 1.

Algorithm 1: the direct k-means algorithm
Input:

```
D = {x₁,x₂,......xₙ}  //set of n data objects.
k //Number of clusters
    δ  //convergence accuracy
```

Output:
A set of K clusters.
Steps:

```
1. Randomly choose K data objects(C₁,C₂,......Cₖ) from D as
initial centers.
2. Assign each xᵢ to its nearest cluster center Cₖ.
3. Update each cluster centre Cₖ as the mean of all xᵢ that
have been assigned as closest to it.
4. Calculate the convergence accuracy δ.
5. If the value of δ has converged, then return(C₁,C₂,......Cₖ);
else go to Step2.
```

This algorithm results in different types of clusters depending on the random choice of initial centers. Thus, as to this disadvantage, we propose a novel method to choose the initial cluster centers.

3   Proposed Algorithm

The initial centers of direct K-means algorithm are chosen randomly, so this kind of algorithm exists problems. Generally speaking, the random selecting of data objects with small sample cannot express the trend of the dataset distribution in space that is say, it is difficult to reflect the dataset's distribution, clustering shape and the change state of data density with random selecting [21-23]. Therefore, in the paper, we propose a K-partition algorithm based on sorting (KPBS for short) for initializing cluster centers, which is a heuristic algorithm being able to avoid the disadvantages mentioned above, and compared with some complex algorithm for choosing initial cluster centers, KPBS algorithm can be more Simple and efficient.

3.1   Algorithm Thought

The basic idea of *KPBS* is to determine the initial centers of the clusters in a heuristic manner so as to ensure that the centers are chosen in accordance with the distribution of data, so it can avoid falling into local optimization to some extent. The algorithm first sort the input dataset, and then partition the sorted dataset into *k* subsets, where *k* is the number of clusters to be defined, finally, the mean value of each these subsets are taken as the initial cluster centers. To further describe the algorithm idea, we can see fig. 1, the distribution diagram of the dataset *D*, which fit Gaussian distribution, and from which we can determine the cluster shape of dataset. Our algorithm thought is implemented as follows.
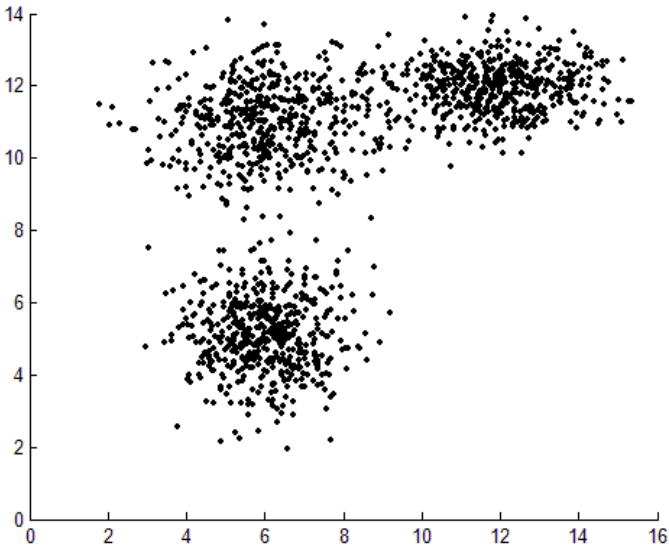
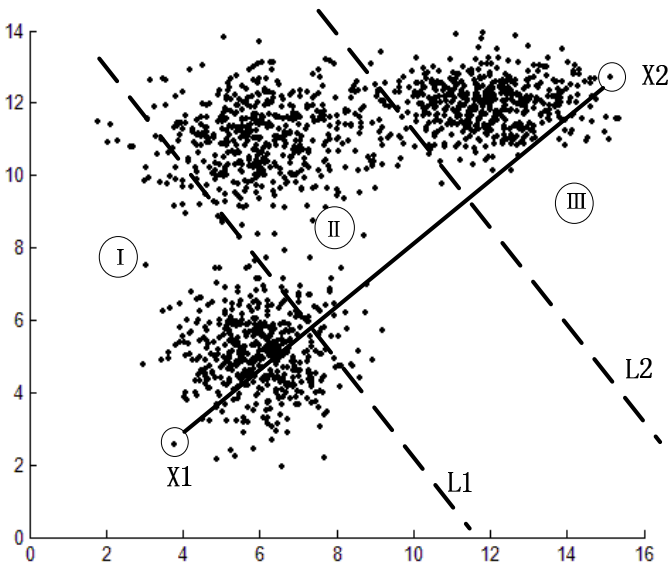**Fig. 1.** The dataset distribution with 3 clusters, which fit Gaussian distribution

Fig. 2. The partition state in data space

First, sort data according to dispersion. In this way, the difference between the first data object and the last data object is minimal, which means the two points disperse extremely or at a furthest distance, just as the points X_1 and X_2 in Fig. 2. Then partition the dataset into *k* parts. Here we make the value of *k* be 3. In the Fig. 3, the line segment x_1x_2 is equally divided into 3 segments. Line L_1 and L_2 are the vertical segment lines of x_1x_2, dividing the space into three parts (as ① ⑪ ⑫), from which we can see that the data points are accurately partitioned into each space. At last, calculate the mean value of each data space as initial clustering center.

3.2   Algorithm Design

However, there is another problem, that is, how to deal with the multi-dimensional data. In general, the data we need to analyze are all multi-dimensional data and each data object X_i has many data features, such as $X_{i1}, X_{i2}, \ldots\ldots X_{id}$, where *d* is the number of object attribute or dimension. As to this case, we should first to determine sorted-dimension; here we adopt the means of maximal dispersion to determine sorted-dimension, which means the more discrete this attribute is, the more information about distribution of dataset it can reflect. We calculate the variance of each attributer dimension for whole dataset, and take the attribute with max variance as the sorted-dimension. The entire dataset are then sorted in non-decreasing order, using the Heap Sort algorithm, based on the sorted-dimension. The sorted data points are then divided into 'k' equal subsets. Finally, the arithmetic means of each of these 'k' subsets are computed, and these means are taken as the initial clustering center. The proposed algorithm is described below as Algorithm 2.

Algorithm 2: Proposed algorithm (KPBS)
Input:

```
D = {x₁,x₂,......xₙ} //set of n data objects.
k     //Number of clusters
```
δ `//convergence accuracy`

Output:
A set of K clusters.
Steps:

```
1. Calculate the variance of each attributer dimension
for whole dataset
```

$$D(d_j) = \frac{\sqrt{\left(X_1 - E\left(d_j\right)\right)^2 + \left(X_2 - E\left(d_j\right)\right)^2 + \ldots + \left(X_n - E\left(d_j\right)\right)^2}}{n} = \frac{\sqrt{\sum_{i=1}^{n}\left(X_i - E(d_j)\right)^2}}{n}, 1 \le i \le n, 1 \le j \le d \text{ ,}$$

`d as the number of dimension,` $E(d_j)$ `as the mean value.`
```
2. Determine the attribute with max variance
```
$d_k (1 \le k \le d)$ `as`
`sorted-partition.`

3. Sort the entire dataset in non-decreasing order based on sorted-partition d_k.

4. Divide the sorted dataset into k equal subsets.

5. Calculate the mean value of each subset and take it as the initial cluster center $(C_1, C_2, \ldots\ldots C_k)$.

6. Assign each x_i to its nearest cluster center C_k.

7. Update each cluster centre C_k as the mean of all x_i that have been assigned as closest to it.

8. Calculate the convergence accuracy δ.

9. If the value of δ has converged, then return $(C_1, C_2, \ldots\ldots C_k)$; else go to Step6.

Unlike the direct k-means algorithm in which the initial centers are chosen randomly, the proposed KPBS algorithm determines the initial centers in a more meaningful way, in accordance with the dataset distribution and clustering shape. Consequently, KPBS algorithm can not only improve the accuracy of clustering results, but also greatly reduce the number of iterations and speed up the clustering convergence. Besides, because of the use of sorting method, it also can reduce the comparison time between the object so as to reduce the time complexity and improve efficiency. We will analyze the time complexity of proposed algorithm as follows.

3.3   Time Complexity Analysis

In the direct k-means algorithm, the initial centers are chosen randomly and the time consuming are mainly used to assign each data object to its nearest centers, which means the time is used before the algorithm converges, the procedure takes time $O(nkl)$, where n is the number of data objects, k is the number of clusters and l is the number of iterations.

In the proposed algorithm discussed in this paper, the first step is to calculate each dimension's variance require $O(n)$ time, where n is the number of data objects, the required time to find the attribute dimension with max variance as sorted-partition dimension is $O(nd)$ where d is the number of attributes, the next step of sorting the dataset based on the sorted-partition dimension can be performed in $O(n \log n)$ time using Heap Sort, the time for partitioning the n data objects into k equal parts and computing the mean of each part is $O(n)$. Since m is much less than n, the time $O(n \log n)$ is more than $O(nd)$, thus the overall time complexity for determining the initial clustering centers is $O(n \log n)$.

The second phase of assigning data points to clusters is the same as direct k-means algorithm, the loop consisting of the assignment of data objects to the nearest clusters and the subsequent recalculation of centers is executed repetitively until reaching the convergence criterion, this procedure takes time $O(nkl)$, where n is the number of data objects, k is the number of clusters and l is the number of iterations. Nevertheless, the proposed algorithm converges in much less number of iterations as

the initial centers are computed in a strategic manner in tune with the data distribution. So the overall time complexity of the proposed algorithm is $O(n \log n + nkl)$.

4   Experiment Results

For testing the accuracy and efficiency of the proposed algorithm(KPBS), we use three available datasets (Breast Cancer-Wisconsin, Thyroid, E-Coli) as experimental data in the UCI repository of machine learning database. Their data features are described as in Table 1. In addition to making a comparison with direct k-means algorithm, we also compare E-Kmeans algorithm in reference [11] with KPBS algorithm mainly on time cost.

Table 1. The feature of each dataset

Data set	Object number	Dimension number	Standard cluster number
Breast-Cancer	669	10	2
Thyroid	215	5	3
Ecoli	336	7	8

The experiment scheme we used to evaluate the accuracy and efficiency of KPBS algorithm is discussed as follows. In the phase of choosing initial centers, direct k-means algorithm and E-Kmeans algorithm randomly choose initial centers in the first phrase, and in the iteration phase they are same. The accuracy is got by comparing the cluster result with the UCI standard cluster result, which is defined as follows:

$$\text{Accuracy:} f = \frac{\Sigma_{t=1}^{p}\alpha_t}{n}\times 100\%$$

n is the number of data objects, p is the number of correct cluster, α_t is the number of data objects which drop into each cluster.

The results of this experiment are the average value after executing many times, the detail result listed as Table 2. Just comparing the direct E-Keans algorithm with E-Kmeans, there is little difference in the accuracy rate because they all apply the random methods to select initial cluster centers. As for time cost, E-Kmeans algorithm is better than direct k-means algorithm. Then making comparison with KBPS algorithm proposed in this paper, the accuracy rate have a better improvement, and especially it has obvious results in the Thyroid dataset. However, they have both high accuracy rate and few differences for the Breast Cancer dataset mainly because Breast Cancer itself just only be divided into two kinds. The phenomenon that KBPS algorithm is equal to E-Keans in time cost explains that KBPS algorithm has higher accuracy rate and efficiency in comparison with k-means algorithm.

Table 2. The comparation of experiment result

| Data sets | Algorithm | | | | | |
| | k-means | | E-Kmeans | | KPBS | |
	Accuracy(%)	Time(ms)	Accuracy (%)	Time(ms)	Accuracy (%)	Time(ms)
Breast Cancer	92.4	78	94.2	66	96.6	62
Thyroid	77.8	60	76.4	52	85.5	51
Ecoli	72.3	64	74.5	48	79.7	45

The experiment above applies to real dataset. In order to further inspect the performance in running time of the proposed algorithm in this paper, we generated eight datasets with the mode of Multivariate Gaussian Distribution. The number of data objects in each dataset is in ascending order. Applying the three algorithms to these eight datasets, we get the relationship between clustering running time and the size of data set as shown in Fig. 3.

Compared with direct k-means algorithm, the algorithms in this paper have much improvement in performance which shows that the choice of initial clustering centre corresponds to the data distribution, as a result of that, it can reduce the numbers of iteration, and accelerate the cluster convergence.
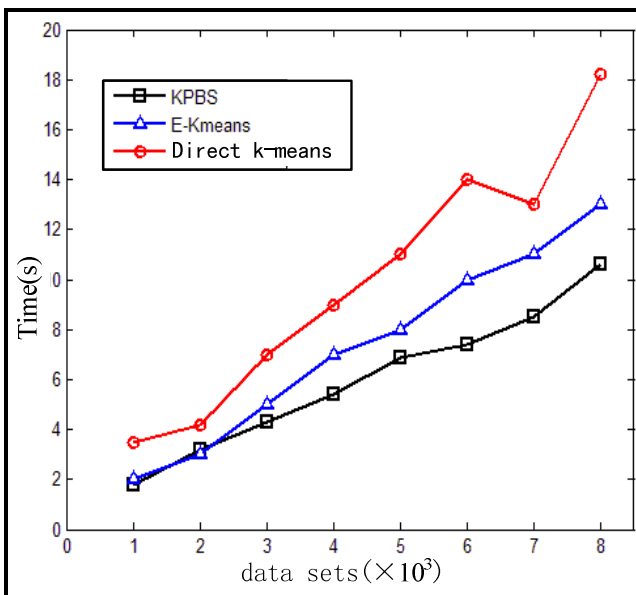

Fig. 3. The running time of different datasets

5   Conclusion

K-means algorithm has been applied to data clustering extensively, but traditional k-means algorithm can't guarantee the accuracy of clustering result since the result can be different as a result of different initial cluster center. Moreover, the complexity of calculation is high since each object should be distributed many times in each iteration. Therefore, as to this case, this paper proposed a cluster method based on sorting and partition that can guarantees the generated initial centre to represent the distribution of datasets and improve the accuracy and efficiency. Compared with direct k-means, the experiment results have already shown that the algorithms in this paper can generate more accurate clustering results with less time cost.

There is a limitation for the proposed method that the value of k should be an input, and because of the distribution of dataset, the selection of k has a great influence on the accuracy of clustering result. In addition, the isolated points in dataset have much effect on the performance of clustering. Both aspects are main contents to be researched in the future.

References

1. Xu, R., Wunsch Il, D.: Survey of clustering algorithms. IEEE Trans. on Neural Networks 16(3), 645–678 (2005)
2. Tou, Gonzaales, R.: Pattern Recognition Principles, pp. 54–57. Addison-Wesley, New Jersey (1974)
3. Linde, Y., Buzo, A., Gray, R.M.: An algorithm for vector quantizer design. IEEE Transaction on Communication 28(2), 84–95 (1980)
4. Babu, G.P.: A near-optimal initial seed value selection in k-means algorithm using a genetic algorithm. Pattern Recognition Lett. 14(10), 763–769 (1993)
5. Huang, C., Harris, R.: A comparison of several codebook generation approaches. IEEE Transaction on Image Process. 2(1), 108–112 (1993)
6. Khan, S.S., Ahmad, A.: Cluster center initialization algorithm for k-means clustering. Pattern Recognition Letters 25(11), 1293–1302 (2004)
7. Birgin, E.G., Martinez, J.M., Ronconi, D.P.: Minimization Subproblems and Heuristics for an Applied Clustering Problem. European Journal of Operational Research 146(1), 19–34 (2003)
8. Corral, A., Alnendros, J.M.: A Performance Comparison of Distance-Based Query Algorithms Using R-Trees in Spatial Databases Information Sciences. An International Journal 177(11), 2207–2237 (2007)
9. Kangngo, T., Mount, D.M., Netanyaha, N.S.: An Efficient K-Means Clustering Algorithm: Analysis and Implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 881–892 (2002)
10. Redmod, S.J., Heneghan, C.: A method for initializing the K-means clustering algorithm using kd-trees. Pattern Recognition Letters 28, 965–973 (2007)
11. Fahim, A.M., Salem, A.M., Torkey, A., Ramadan, M.A.: An Efficient enhanced k-means clustering algorithm. Journal of Zhejiang University 10(7), 1626–1633
12. Fogel, D.B.: An introduction to simulated evolutionary optimization. IEEE Trans. Neural Networks 5(1), 3–14 (1994)

13. Holland, J.H.: Adaptation in Natural and Artificial Systems. Univ. of Michigan Press (1975)
14. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice-Hall (1989)
15. Klein, R.W., Dubes, R.C.: Experiments in projection and clustering by simulated annealing. Pattern Recognit. 22, 213–220 (1989)
16. Selim, S.Z., Alsultan, K.: A simulated annealing algorithm for the clustering problem. Pattern Recognit. 10(24), 1003–1008 (1991)
17. Babu, G.P., Murty, M.N.: Simulated annealing for selecting initial seeds in the k-means algorithm. Ind. J. Pure Appl. Math. 25, 85–94 (1994)
18. Bhuyan, J.N., Raghavan, V.V., Elayavalli, V.K.: Genetic algorithm for clustering with an ordered representation. In: Proc. 4th Int. Conf. Genetic Algorithms. Morgan Kaufmann (1991)
19. Jones, D.R., Beltramo, M.A.: Solving partitioning problems with genetic algorithms. In: Proc. 4th Int. Conf. Genetic Algorithms. Morgan Kaufmann (1991)
20. Babu, G.P., Murty, M.N.: A near-optimal initial seed selection in K-means algorithm using a genetic algorithm. Pattern Recognit. Lett. 14, 763–769 (1993)
21. Babu, G.P., Murty, M.N.: Clustering with evolution strategies. Pattern Recognit. 27(2), 321–329 (1994)
22. Babu, G.P.: Connectionist and evolutionary approaches for pattern clustering, Dept. Comput. Sci. Automat., Indian Inst. Sci. (1994)
23. Goldberg, D.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley (1989)

B2C Trading Platform Security Strategy Based on the Optimal Stopping Theory

Wei Li[1], Hongtu Zhang[1], and Tingting An[2]

[1] Northeast Dianli University, Jilin, China
[2] Beijing University of Chemical Technology, Beijing, China
npuliwei@126.com, {zhtpld,att200851012}@163.com

Abstract. In order to find out an optimal stopping rule for the B2C trading platform security risk, we propose a quantitative random process decision-making model. Two factors were regarded as independent variables which were investment for search security risk and profit of the B2C enterprise. In the same time, the optimal search times of enterprises' B2C trading platform was a dependent variable, and the loss resulted by the B2C trading platform security risk as a random process. The existence of the optimal solution of this problem was proved, and the solution of the computer program was given, by which we can solve the problem conveniently. Therefore, the search times has strongly maneuverability. It also provides the theoretical and practical guidance for the safety decision-making of the B2C trading platform.

Keywords: optimal stopping, B2C, trading platform, security strategy.

1   Introduction

With the increasing amount of network e-commerce transactions, B2C trading platform security issues have become increasingly prominent. Many domestic e-commerce trading platform was attacked by hackers, causing major economic losses to the network business and customers. For example: in December 2011, CSDN company servers were attacked by illegal invasion, which caused core data was leaked; in March 2012, the famous shopping site—Dangdang whose user accounts were stolen by hackers which brought to the consumer and Dangdang significant losses. However, due to the openness of the Internet and the illegal invasion to attack the interests of the network attacker B2C trading platform, there are always security risks in the B2C trading platform , therefore, improving the B2C trading platform of security has been a hot and difficult academic research. Zhi-Jun Cheng [1], Xiao-Qiang Wu [2], etc., through the improvement of computer network security technology, to strengthen the trading platform security from a technical point of view; Yu DanDong, Dahai [3], Sun Xiang, Zhang Shuo Yang [4], etc., by studying the sources of risk to consumers in the B2C transactions, the point of view of the consumer risk is primarily derived from the B2C e-commerce virtual interface ", and accordingly put forward the security suggestions on improving B2C trading platform. Xie Hongyan [5], Liu Ye Fei [6],etc, under the premise of a variety of factors to

consider legal, ethical and management, put forward suggestions to build e-commerce platform security system from the system's point of view. Network business and its customers in the B2C transactions Introduction B2C trading platform security risks mainly produce various security vulnerabilities in the platform itself, how both accurate and economical to find a trading platform for security vulnerabilities, reducing illegal invasion of the economic losses caused by network businesses and their customers is an important decision for the B2C trading platform security policy. During searching Security risks, increasing lookup times, and will no doubt improve the security of the platform, thereby reducing the probabilities of the network business and its customers' economic losses caused by security risks; at the same time, the increase in costs associated will inevitably lead to cost improved. In the contrary, reducing lookup times of security risks , and reducing the cost of checking the safety of trading platform would likely not be guaranteed, therefore it can increase the possibility of the losses come from the network business and its customers. According to the security issues in B2C trading platform above, starting from the theory of optimal stopping, setting the target of the minimum economic losses from customers caused by B2C trading platform security risks, to make the optimal number of security vulnerabilities as decision variables, regard the customers' losses as a random event cause by B2C trading platform security issues, at the same time considering the cost of the B2C trading platform security risks of the network enterprise solutions and the loss of eliminated due to the security risks to the enterprise and its customers, to construct a quantitative random process of decision-making model. In accordance with the characteristics of the security risks under the assumption that the distribution obey the Poisson distribution and binomial distribution, to prove the existence of the optimal solution of the problem, and take advantage of the C + + programming language gives the algorithm of solving the decision problem and its program that achieved, making the solution of the problem has stronger operability.

2   Construct the Model

Supposed there are many security which caused by security vulnerabilities in the B2C trading platform structured by one corporation. There are N pieces of security risk in them. Here set N as a nonnegative random variable, and its expected value satisfies $EN < \infty$. That's to say the quantity of the security risk is limited. In order to improve the accuracy of the trading platform, we should do some job to find and repair those security vulnerabilities. Assume that each mistake which has been found can be rectified to enhance the quality of the platform, which means some profit can be acquired from customers and corporation. Here, we call this reduce losses "profit for corporation". Provided that the profit which we can acquired from the proofread is w , the cost each time we should spend $b_k \left(b_k > 0, k = 1, 2, 3 \cdots \right)$. The losses which caused by the each mistake is $d_k \left(d_k > 0, k = 1, 2, 3 \cdots \right)$, and let $d = \sup_k d_k < \infty$. That means the losses which can't be found through k times

collation is d_k, and the maximum is less than infinite. Defined x_k is the number of security risk which can't be found through $k-1$ times before at the k time collation. If the N *given* through the $x_1, x_2, x_3, \cdots x_{k-1}$ times collation, Let x_k submits to binomial distribution conditional probability. That is to say $(x_k/x_1, x_2, x_3, \cdots x_{k-1}) \sim B(N_{k-1}, p_k)$. Here, $N_{k-1} = N - \sum_{i=1}^{k-1} x_i$, and p_k is the probability of the mistakes which have been found through k time collation.

Based on the aforementioned hypothesis conditions, let y_k denote the profit which acquired from the k times proofreading.

$$y_k = w \sum_{i=1}^{k} x_i - g(k) - d_k E\{N_k | x_1, x_2, x_3, \cdots x_k\} \tag{1}$$

In formula （1）, $w \sum_{i=1}^{k} x_i$ denotes the profit which has been acquired through the k times proofread $g(k) = \sum_{i=1}^{k} b_i$ denotes the total cost which have been spent through the k times proofread.

$E\{N_k | x_1, x_2, x_3, \cdots x_k\}$ denotes the expected value of the mistakes' quantity which has not been found yet through k times collation. $d_k E\{N_k | x_1, x_2, x_3, \cdots x_k\}$ denotes the losses which have been caused by the security risk which has not been found yet through k times collation.

3   Model Solution and Optimal Solution Condition

3.1   Solved for the Model and the Existence Condition of Optimal Solution at the Poisson Distribution

When the number of loopholes submits to Poisson distribution with parameter $P(\lambda)$, supposed when

$$p_n = p, \quad p \in (0,1), \quad d_n = d < \infty, \quad n = 1, 2, \ldots, \tag{2}$$

Then,

$$(x_n | x_1, x_2, \cdots, x_n, N) \sim B(N_n, p). \tag{3}$$

To profit sequence equation (*), it has been shown that,

$$E\{N_n \mid x_1 = a_1, x_2 = a_2, \cdots, x_n = a_n\} = \lambda(1-p)^n \tag{4}$$

Then,

$$E\{y_{n+1} \mid F_n\} = y_n + (w+d)p\lambda(1-p)^n - C_{n+1}$$

$$E\{y_{n+1} \mid F_n\} = y_n + (w+d)p\lambda(1-p)^n - C_{n+1} \tag{5}$$

Let,

$$n_0 = \inf\{n \geq 1, (w+d)\lambda p(1-p)^n \leq C_{n+1}\} \tag{6}$$

To arbitrary $\forall k \geq n_0$, when $n \in [n_0, k]$, then $\lambda_n = y_n$;

When, $n \in [1, n_0]$, then $\lambda_n = y_n + \sum_{m=n}^{n_0 - 1}[(d+w)p\lambda(1-p)^m - C_{n+1}]$.

Then, $\sigma = n_0$, σ can be defined as optimal stopping.

That is to say, if the number of security loopholes submits to Poisson distribution, the optimal number of proofreading can be determined by the σ above. σ is a positive integer, that is the optimum proofreading time.

3.2   Solved for the Model and the Existence Condition of Optimal Solution at the Binomial Distribution

When the number of security loopholes submits to binomial distribution with parameter, $N \sim B(M, a)$, where M is the sum of the loopholes we have been collected, a is the probability of each piece of loopholes. $0 < a < 1$, then:

$$(N_n \mid x_1, x_2, \cdots, x_n) \sim B(M_n, a_n) \tag{7}$$

Here

$$M_n = a\prod_{i=1}^{n}(1-p_i) \Big/ \left[(1-a) + a\prod_{j=1}^{n}(1-p_j)\right], \tag{8}$$

Let $p_j \equiv p, d_j \equiv d, j = 1, 2, \cdots 0 < p < 1$, where $q = 1 - p, s = 1 - a$, then

$$E\{N_n \mid F_n\} = \frac{aq^n}{s + aq^n} M_n \tag{9}$$

So the profit sequence equation (*) turns to this sequence:

$$y_n = w \sum_{i=1}^{n} x_i - g(n) - d \frac{aq^n}{s + aq^n} M_n \tag{10}$$

When $\{C_n\}$ is non-increasing, $\{y_n, F_n\}_1^\infty$ is a monotone stochastic sequence.
Let

$$\sigma_1 = \inf\{n \geq 1, E\{y_{n+1} \mid F_n\} \leq y_n\}$$

$$= \inf\{n \geq 1, \frac{aq^n}{s + aq^n} M_n \leq \frac{C_{n+1}}{(w + p)d}\}, \tag{11}$$

So, when $N \sim B(M, a)$, $p_j = p$, $d_j = d$, $0 < p < 1$, $0 < a < 1$, and $\{C_n\}$ is non-increasing, $\sigma_1 = \inf\{n \geq 1, E[y_{n+1} \mid F_n]\}$ can be defined optimal stopping.

Through the derivation process above, we can also conclude that if the number of security risk information submits to binomial distribution, the optimal proofread numbers can be calculated only by σ_1.

3.3   The Existence Condition of the General Solution

Supposed the probability space is (Ω, F, P). N is measurable to F, and $F_n = \sigma(x_1, x_2, x_3, \cdots, x_n)$, $F_n \in F$ .Let t denote the random variable, $\{t \in I\} = 1$, and $\forall t = n(n = 1, 2, \cdots,) \in F_n$, then , t can be defined stopping.

Define $\Im = \{t, stopping t, P(t < \infty) = 1, Ey_t^- < \infty\}$, when $t^* \in \Im$, and $Ey_{t^*} = \sup y_t, Ey_{t^*} < \infty$, then t^* can be defined optimal stopping.

To profit sequence that is mentioned with the equation (1) y_k ,
Set

$$t_n = \{t \in \Im, t \geq n\}, \gamma_n = ess \sup_{t \in \tau_n} \{y_t / F_n\}, v_n = \sup_{t \in t_n} Ey_t,$$

$$v = v_1, v_n' = \lim_{k \to \infty} v_n^k, \gamma_n' = \lim_{k \to \infty} \gamma_n^k. \tag{12}$$

The conclusion can be obtained from reference paper [7], if $d = \sup_k d_k < \infty$, then

$$v_n' = v_n, \quad \gamma_n' = \gamma_n$$

That's to say, if the maximum loss which is less than infinite in all the mistakes that has not been discovered, then the limit must be existence, which are $v'_n = \lim\limits_{k\to\infty} v_n^k$, $\gamma'_n = \lim\limits_{k\to\infty} \gamma_n^k$ [7].

For stochastic sequence y_k, if $n \geq 1$, $t, s \in T$, when $n \in \{s > n\}$, $E(y_s | F_n) \geq y_n$; when $\{s = n, t \geq n\}$, $E(y_t | F_n) \geq y_n$, so $Ey_s \geq Ey_t$ [8,9].

Set $A_n = \{E(y_{n+1} | F_n) \leq y_n\}$, $n = 1, 2, \cdots$ If $A_1 \subseteq A_2 \subseteq \cdots \subseteq A_n \subseteq \cdots$ and $P\left(\bigcup\limits_n A_n\right) = 1$

From the view of optimal stopping, the Stop variable make $y_n \geq E(y_{n+1} | F_n)$ reasonable.

If $s \in T$ and $\liminf\limits_{n} \int_{s>n} y_n^{+} = 0$

Then, any $t \in T$ who can make $\liminf\limits_{n} \int_{t>n} y_n^{-} = 0$ reasonable will be $Ey_s \geq Ey_t$. Above all, $Ey_t < \infty, t \in T$ when $Ey_{t^*} = \sup\limits_{t} y_t$ and $t^* \in T$, we call t^* the optimal stopping time.

From Double expectation formula, we can know that:

$$E(y_s) = E[E(y_s | F_n)] = \int E(y_s | F_n) dF_n .\qquad(13)$$

4   The Model Implementation Steps and Application Examples

4.1   Algorithm Steps

Certain theoretical assumptions for a specific B2C trading platform, in accordance with above model, we can achieve the optimal choice of its security policy, the specific algorithm steps are as follows:

Step One: Determine the security vulnerabilities and find obedience probability distribution, check the constraints of the model whether it meets the requirements.

During finding the security risks in B2C trading platform, binomial and Poisson distribution can both describe the actual problem objectively, therefore, we generally based on the actual issues, choose one as probability distribution described that security vulnerabilities obey.

Step Two: According to the problem, estimate and project the parameters in the model.

It can be used sampling assumption estimation method for the probability of occurrence of the error data information in the model; b_k that costs required for each test, based on the platform maintenance experience, set it to a fixed value; Because of the randomness of the gains w caused by finding a security flaw and correcting it and

the loss d_k caused by each vulnerability that through finding work legacy, and set w and d_k to a random number.
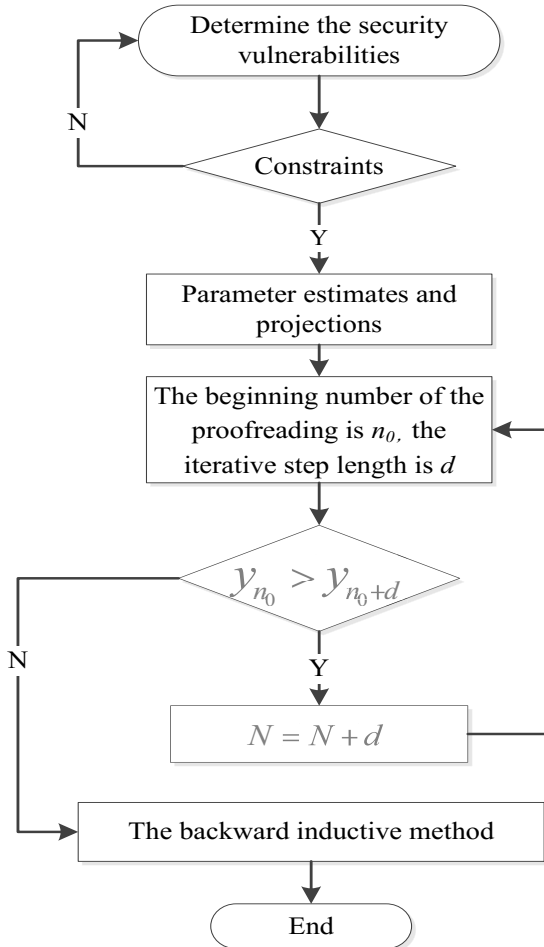
Fig. 1. Flow chat of application steps

Step Three: Given the number N of initial safety hazards check, let n_0 be the beginning number of the proofreading, d be the iterative step length, calculating the value of profit through the revenue function y_{n_0} and y_{n_0+d}, if $y_{n_0} > y_{n_0+d}$, turn to Step 4;otherwise to turn to Step5.

Step Four: Set $N = N + d$, then turn to Step 3;

Step Five: To solve the calculation by using the backward inductive method. Then end.

The flow is shown in figure 1.

4.2   Examples

Assuming there are some security risks in a B2C trading platform, and existing some security loopholes which will lead to a hacker attack and losses to the businesses and consumers. So, we try to find an optimal security strategy, which can help us to find those loopholes with maximum probability by using minimum searching times which can lead to maximize corporate earnings (Network business and customers suffered minimal loss).

Assuming that the event of "Loophole is found" on the platform follows Poisson distribution, and the number of loopholes is N = 100. According to the experience of platform maintenance, we set a fix cost -- 1000 Yuan -- to per test; Because the current B2C trading platform single amount maximum to 20000 Yuan, and according to the importance of loopholes, the different income is random, so find loophole the profits we will set its to 0-20000 between random number. According to the above assumptions, the result of algorithm routine operation are -- suppose there are 100 loopholes, and they need the number of test for 11, income 78640.1Yuan.
It must be pointed out that we set a random number during the process of search each enterprise obtained income, thus, every time we run the program we may get different results. After running the program several times we can get the average mean, which is as the optimal value of lookup times.

Algorithm routine and operation results are in the appendix.

4.3   Appendix Program

```
#include <iostream>
#include <time.h>
#include <math.h>
using namespace std;
//#define N 100
void main()
srand((unsigned)time(NULL));
float p=(float)(rand()%10000)/10000;
float q=1-p;
float a=(float)(rand()%20000)/20000;
float s=1-a;
int k=0;
int N=100;
int w=(float)(rand()%20000);
int b_k=1000;
int x_k=0;
int sum_x_k=0;
int d_k=(float)(rand()%20000);
long double y_k=0;
long double py_k=0;
do
```

```
k++;
py_k=y_k;
x_k=p*(N-sum_x_k);
sum_x_k=sum_x_k+x_k;
y_k=w*sum_x_k-k*b_k- d_k*a*pow(q,k)/(s+a*pow(q,k));
while(y_k>=py_k);
cout<<"If there "<<N<<"loopholes"<<endl;
cout<<"number of testing"<<k<<endl;
cout<<"profit after k times collation"<<y_k<<endl;
System ("pause");
```

5   Conclusion

Considering the above model in search for B2C trading platform security vulnerabilities expense and due to the network security vulnerabilities to reduce enterprise and its customers the benefits of two factors premise, using the optimal stopping time theory based on the random process model, finding the B2C trading platform security hole to find the optimal stopping time, namely the best number of search for B2C trading platform safe hidden trouble to find the optimal number of decision-making work provides a quantitative basis. At the same time, this paper puts forward the model and the algorithm is realized by C+ + language program, and combining with an actual example given in the computer to realize the results, to validate the model's maneuverability, makes the result of this paper to B2C trading platform security hidden danger find decision to the solution of the problem has certain guiding significance, and has a broad prospect of application.

References

1. Cheng, Z.-J., Xu, H.-Y.: Research and Realization of B2C Web Security Based on XML. Science Technology and Engineering 8(15), 4148–4151 (2008)
2. Wu, X.-Q., Liu, J., Zhu, S.-P., Zhou, R.-X., Qiu, W.-H.: Research on payment authentication model based on trusted third party. Computer Integrated Manufacturing Systems 11(5), 690–695 (2005)
3. Yu, D., Dong, D.-H., Liu, R.-M., Yuan, Y.-D.: Study of types, resources and their influential factors of perceived risks in purchase online. Journal of Dalian University of Technology (Social Sciences) 28(2), 13–19 (2007)
4. Sun, X., Zhang, S.-Y., You, D.-R., Chen, Y.-W., Wang, E.-P.: The Source of Consumers Risk and Their Perception in B2C E-Commerce. Chinese Journal of Management 2(1), 45–54 (2005)
5. Xie, H.-Y.: Study on security and counter measures of electronic commerce. Journal of Harbin University of Commerce: Natural Science Edition 23(3), 350–358 (2007)
6. Liu, Y.-F., Zhao, D.-A., Shan, Z.-Y.: Probe into Electronic Commerce Security. China Safety Science Journal 16(2), 109–113 (2006)
7. Li, B.-C., Luo, J.-S.: Optimal Stopping of proofreading problems with Gain and Costs. J. 17(1), 117–121 (1995)

8. Luo, J.-S.: Optimal Stopping of Proofreading Problems with Costs. J. 15(2), 99–104 (1993)
9. Stabile, G.: Optimal timing of the annuity purchase: combined stochastic control and optimal stopping problem. International Journal of Theoretical and Applied Finance 9(2), 151–170 (2006)
10. Chen, A., Wu, Y.-N.: Design of Termination Mechanism in Emergency Management Based on Optimal Stopping Theory. Chinese Journal of Management Science 18(4), 173–182 (2010)
11. Lempa, J.: Optimal stopping with random exercise lag. Mathematical Methods of Operations Research 7, 53 (2012)
12. Lempa, J.: Optimal Stopping with Information Constraint. Applied Mathematics & Optimization 6, 62 (2012)
13. Ivanov, R.V.: On the problem of optimal stopping for the composite Russian option. Automation and Remote Control 718 (2010)
14. Kawano, H., Hattori, T., Takeda, K., Izumi, T.: Early structural change detection as an optimal stopping problem: solution theorem and its proof using reduction to absurdity. Artificial Life and Robotics 154 (2010)
15. Kuznetsov, Y.A.: An optimal stopping rule in the problem of discrimination between hypotheses in a scheme with possible switching. Journal of Communications Technology and Electronics 5512 (2011)
16. May, P., Ehrlich, H.-C., Steinke, T.: Structure Prediction Pipeline: Composing a Complex Biological Workflow through Web Services. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) Euro-Par 2006. LNCS, vol. 4128, pp. 1148–1158. Springer, Heidelberg (2006)
17. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
18. Czajkowski, K., Fitzgerald, S., Foster, I., Kesselman, C.: Grid Information Services for Distributed Resource Sharing. In: 10th IEEE International Symposium on High Performance Distributed Computing, pp. 181–184. IEEE Press, New York (2001)
19. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The Physiology of the Grid: an Open Grid Services Architecture for Distributed Systems Integration. Technical report, Global Grid Forum (2002)
20. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov

Microblog Comments Sentiment Analysis
Based on Extended Emotional Lexicon

Xu-ming Wang[1,2], Wan-li Zuo[1,2,*], Ying Wang[1,2], and Xiang-lin Zuo[1]

[1] College of Computer Science and Technology, Jilin University, Changchun, 130012, China
[2] Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, Changchun, 130012, China
`wxmsmile@126.com, {wangying2010,zuowl}@jlu.edu.cn`
`zuoxl_jlu@163.com`

Abstract. Text sentiment analysis is a technology of high practical value and has been widely applied in spam filtering, recommendation system and automatic text summarization. This paper presents a text classification method based on extended emotional lexicon for microblog. With the help of the Sina Weibo official API, we extract the comments under the hot topics. By means of the existing emotional lexicon and the extended microblog emoticons lexicon, the cyberwords lexicon, and the interjection lexicon etc., taking into account the negation rules, the modification of the degree words, the effect of sentence patterns and so on, we design a contrast experiment of six groups and figure out the promoting effect of the various impact factors on the text sentiment classification accuracy. In addition, this paper puts forward the detailed computational formula to analyze the emotional intensity. By means of existing lexicons and extended lexicons, considering the variety of impact factors, experimental results from the microblog comments emotional polarity evaluator (MCEPE) developed in C++ show that the classification accuracy can reach up to 80% or more.

Keywords: microblog, extended emotional lexicon, emotional intensity, sentiment classification.

1 Introduction

With the advent of the era of Web 2.0, Internet users and network has more interactive behavior. People turn into writing the Internet from simply reading it. The protagonist of the Internet has changed into living persons from static pages. People on the network record their life, write their mood and express their views. Thus a lot of valuable texts have been accumulated on the Internet. Whether it is the evaluation of the quality of goods, or discussion of the film and television works, or the concern of the current hot events, the data have a potential value of mining.

---

* Corresponding author.

Microblog as the most popular way of socializing at present is a set of social inter-course, social media and communication. People not only record their daily chores in the microblog, but also communicate with others and more are focused on the current hot topic to express their views. These views, whether it is for businesses to improve product quality, or for the government to conduct opinion polls, are of great signific-ance. The paper is based on this consideration; it extracts reviews on hot topics as a source of data, analyses emotional tendencies (ET) of each comment, and finally gets the support number, antilog and neutral number held by microbloggers to determine the overall attitude about the hot topic.

2   Related Work

Sentiment analysis, also known as the analysis of viewpoint, opinion mining, is the analysis, processing, induction and reasoning of the text with emotional colors, in order to determine the overall emotion expressed in the text, whether positive, negative or neutral [1]. The sentiment analysis study began in 2000 or so, mainly in the English documents [2, 3].

Chinese text sentiment analysis study is relatively few, most researchers use a sim-ilar method to do Chinese text sentiment analysis which is usually used in English [4, 5], while ignoring the differences between Chinese and English.

Text sentiment analysis utilizing microblog-based data abroad are mainly concen-trated on twitter [6, 7]. Due to the late development of the Chinese domestic microblog such as Sina Weibo [8], microblog emotional study is in its infancy [9].

To sum up the relevant research of the past 10 years, the sentence level sentiment analysis method can be concluded in the following two points: 1. unsupervised learning methods, by means of emotional lexicon getting emotion words extracted from the text to determine the text's corresponding emotional polarity (positive, negative or neutral); 2. supervised learning method based on the training corpus, by means of a specific classification algorithm, such as naive Bayes classification algorithm, support vector machines, k-nearest neighbor learning algorithm to learn classification from the training data set. Both methods have their own advantages, as well as shortcomings. For instance, method 1 is faced with lexicon coverage and low accuracy, while method 2 takes a lot of time on the training corpus manual annotation. The paper is just to overcome the shortcomings of method 1, utilizing extended emotional lexicon as well as the calculation method of emotional strength for sentiment analysis, in order to improve the classification accuracy.

3   Mining Method

3.1   Data Collection

The hot topic and its related comments can be obtained by calling the Sina Weibo official API. Through the collection of topics of interest, call GetFavorites interface to

get these topics' corresponding microblog IDs, and finally get the topic's following comments by its ID.

3.2   Data Preprocessing

The data obtained through calling official API are all comments for a specific topic until the call time. We write the comments to a text file, each line representing a comment.

Before extracting the emotional words from the comment text, the sentence needs to be segmented, while before segmentation we remove redundant information in the microblog as follows.

In Sina Weibo, "//@someone" means reposting someone's message. Reposting includes pure reposting (no comments) and comments-with reposting (just write your comments before "//@someone"). Any reposting, with comments or not will be displayed on personal homepage; separate comment without reposting will only be displayed below the original message. "//@someone:" may begin a sentence (no comments reposting), also may appear in the middle of a sentence (comments-with reposting or more than one reposting); may appear once (only one reposting), may also appear several times (more than one reposting). We believe that if a person just reposts with no comments, he by default agrees with the views of the first reposted one, so we retain only the views of the first reposted and remove the other ones'; as with comments-with reposting, the comments may be contrary to the views of those reposted ones, so we retain only the current reposter's reviews, and remove all the reposted ones' since they will be appeared in the previous reviews.

"Reply@" in the microblog indicates the interaction between the reviewers. Although these replies are somehow related with the original topic, they do not directly express the views for the original microblog, just interaction among the reviewers, and analysis on these replies will increase the complexity of the algorithm and the running time, therefore we delete the comments which contain "Reply@", and only consider the comments directed at the original microblog.

In Sina Weibo "#topic#" indicates that this message is issued for a topic. Since the topic is selected manually according to one's interest, there is no need to do segmentation processing on "#topic#"; we remove the tag "#topic#" from the comment.

"@someone", is intended to enable a person to know this message. It is not a direct expression of emotion and attitude; therefore we remove these characters from a comment.

Short chains like "http://t.cn/zOttm2O" belong to hyperlinks. They cannot directly show the author's emotional attitude, and would cause interference to the Chinese word segmentation program, so they are removed from the comments.

3.3   Existing Lexicon-Based Approach

The basic idea of the method is that based on the existing lexicon, get the number of emotional words appeared in the text, and compare the number of positive emotion words and negative ones. If positive emotion words are more, the emotional polarity of

the text belongs to positive; if negative emotion words are more, the sentiment polarity belongs to negative; if they are equal, then neutral.

In this paper, we use "sentiment analysis word set" as the original emotion lexicon published by HowNet [10]. The word set contains six categories: positive emotion words, negative emotion words, positive evaluation words, negative evaluation words, the degree words and perception words. Each category contains Chinese and English; since we do Chinese sentiment analysis, we select only the Chinese version to use. Positive emotional words such as "love", "appreciation", "happy" sum to 836; positive evaluation words like "indispensable", "talented", "beauty" sum to 3730; negative emotion words including "sadness", "dubitation", "contempt" sum to 1254; negative evaluation words like "ugly", "bitter", "excessive", "specious" sum to 3116. Due to the identical emotional polarity, we merge the positive emotion words and positive evaluation words into one file, collectively referred to as positive emotions lexicon (PEL); similarly we merge negative emotion words and negative evaluation words into another file, collectively called negative emotions lexicon (NEL).

After the merger, we found that raw emotion lexicon is not perfect, and some words are even given the wrong polarity, such as the negative idiom "The thighs grow fleshy again" (a sigh of regret about one's idleness) assigned to the positive emotions lexicon. Therefore we make some amendments to the combined lexicon. After amendments, we get a total of 4446 positive emotional words and 4308 negative emotions words, which constitute our initial emotional lexicon (IEL).

To count the number of emotional words in the text of a comment, we first need to do word segmentation processing. We use ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) [11] developed by Dr. Hua-ping Zhang from Chinese Academy of Sciences. The segmentation system has achieved excellent results in a number of domestic and international segmentation contests, and has high accuracy and fast segmentation speed. Furthermore, it supports calling by C, C++, C#, Delphi and Java etc. and can be customized according to the users' needs.

3.3.1   Not Consider Negation Rules

In this case, consider only the number of positive emotional words and negative emotional words in sentences, regardless of the negation. In order to speed up processing, we eliminate the stop words in the comment text, such as punctuation, auxiliary words, interjections like "ah, oh, fie, ouch", conjunctions like "although, however, not only, but also" etc.; these words do not appear in the initial emotional lexicon (IEL), and they would not affect the experiment results in case of not considering negative rule.

3.3.2   Consider Negation Rules

When dealing with the emotions of the text, one problem we have to face is "negative transfer" phenomenon, that is, sentiment analysis about the sentence with a negative word or phrase in it. For instance, "the image quality of this camera is not satisfying", due to the use of the negation word, makes the emotional attitude into negative; as to sentence containing emotional word and ending with a question mark, like "What he has done is right?", we also consider this belong to the "negative transfer" phenomenon. As it relates to the negative words and punctuation, we do not remove stop words after the text segmentation while make experiments directly on the comments after preprocessing.

In order to facilitate the calculation, we give a score for each emotion word. Positive emotion word score 1, while negative emotional word score -1. For the "negative transfer" phenomenon, we process the following algorithm:

- Read each word in the comment after segmentation into the temporary variable word;
- See whether the word is in the negative lexicon;
- See whether the word is in positive emotions lexicon (PEL) to decide whether clause emotional score plus one;
- See whether the word is in negative emotions lexicon (NEL) to decide whether clause emotional score minus one;
- If you encounter a clause mark, process as follows;
- If the clause contains an odd number of negative words and contains emotional words, the emotional polarity reverses;
- If the sentence clause mark is the question mark, the emotional polarity reverses;
- Add the clause emotional score to the whole comment emotional score;
- If the whole comment has not been finished, go back to (1);
- Compare the whole comment emotional score with 0 to determine its emotional polarity.

Negation words lexicon (NWL) contains some negative words and phrases containing negative meanings like "no", "hardly", "deny", "negative", "none", "lack of" etc.. The clause mark lexicon contains ",", "，", ".", "。", "?", "？".

The main idea of the above algorithm is to consider a comment message composed by several clauses, if we respectively calculate the clauses emotional score, then add them up and we can get the whole emotional score for the whole. When calculating the clause emotional score, we consider the negative words as well as the question mark; the emergence of an odd number of negative words would reverse the emotional polarity; if emotional words and a question mark appear in one clause, it will also reverse the emotional polarity.

3.4   Emotional Lexicon Extension

In section 3.3, though the emotional words in positive lexicon or negative lexicon are more than four thousand after amendment, the lexicons provided by HowNet do not encompass network buzzwords, and do not consider the Sina Weibo emoticons. Therefore we extend the emotional lexicon from the two aspects: Sina Weibo emoticons and network buzzwords.

Sina itself provides the emoticons of various emotions, these emoticons show in the form of an image on the page, while the corresponding text is "[emoticons description]". Such as this expression, the corresponding text is "[Billd messy]". If we do not import the Sina Weibo emoticons description lexicon into the segmentation system ICTCLAS, ICTCLAS will cut the description into "Billd" and "messy "and not treat it as a whole. So we need to get the emoticons description between "[" and "]" to do emoticons processing. We get all the emoticons description by calling Sina Weibo official API, and manually divide them into positive emoticons and negative emoticons. Thus we get a positive emoticons lexicon (PEL, including 345 emoticons

description) and negative emoticons lexicon (NEL, containing 400 emoticons description). Before ICTCLAS makes word segmentation, we import all the emoticons lexicons so that ICTCLAS would treat the Sina Weibo emoticons description as a whole; after segmentation, when processing each line, simply compare the words between "[" and "]" with those in the emoticons lexicons. What we should do is to add emoticons processing part to internal cycle of the algorithm in section 3.3.2.

For the network buzzwords such as "niubility", "smilence", "gelivable", and so on, these words have distinctive characteristics of the times, and can express one's emotions accurately. From major forums and blogs we collect common network buzzwords in recent years and divide them into the positive network buzzwords lexicon (PNBL, 187 words in all) and negative network buzzwords lexicon (NNBL, 410 words in all). How to process these buzzwords is similar to that of Sina Weibo emoticons description.

3.5   Emotional Intensity Calculation

The text sentiment classification researches in the past just determine the emotional polarity of the text (positive or negative or neutral), but ignore the analysis of the emotional intensity. For example, "I hate you so much" expresses strong discontent, while "Today's lecture is so so" just shows a relatively mild emotional attitude. In some cases, reviewers may express both a positive and negative attitude toward a subject, such as "This cellphone's standby time is not very long, but photograph very well and the software runs smoothly. Overall, it's very good ", and then you need to capture the overall emotional intensity and focus. We consider the calculation of the emotional intensity from the following three aspects:

The HowNet provides the degree words lexicon. The lexicon includes "extreme/most", "very", "more", "-ish", "insufficiently", "over" six levels. Taking into account the similarity of degree, we merge "-ish", "insufficiently" level lexicon into one lexicon, collectively referred to as "-ish" level lexicon. Some words in "over" level lexicon coincide with negation words, such as "too much", "too far" and "too", we will add these words to the negation word lexicon (NWL) in section 3.3.2 and treat them as the "negation transfer" case; the remaining words like "far more than", "cannot…too", "super", belong to the modification words, we add them to the "very" level lexicon. So we finally get a total of four degree words lexicon (DWL) - "extreme / most", "very", "more" and "-ish" level. The four lexicon corresponding weights are 2, 1.5, 1.2 and 0.2. The weight indicates that if an emotional word is modified by a degree word, we multiply its original emotional value 1 (or -1) by the degree word's weight to get its current emotional value.

Consider the use of "？", "!" and interjection. For example, "Huh, should I teach you?" the use of interjection "Huh" and the question mark express discontent;   while "Ha ha, I finally learn how to drive the car!" interjection "Ha ha" and the use of the exclamation mark express a strong positive emotion. For this reason, we collect "ha ha, oh, ah, bah, ouch, cough" and other words as interjection lexicon and divide them into positive interjection lexicon (PIL, including "ha ha", "hey" etc.) and negative interjection lexicon (NIL, including "bah", "alas" etc.), and respectively consider their effect on the sentence sentiment score. We believe that if the emotional words and interjection or exclamation or question mark coexist in one clause, then the clause's sentiment score double.

Take the effect of sentence pattern into account. We consider the following relations: the coordinative relation like "both ... and ...", the climactic relation like "not only ... but also ...", both of which are called "paralleling clauses" type , for both clauses' emotional intensity should be considered; alternative relation like "rather … than ...," adversative relation like "... but ...", conditional relation like "Whatever ... ", causal relation like "... so … ", summary statement like "... In conclusion ... ", we call all of these "emphasis postposition" type, of which we consider the emotional intensity of the second half of the sentence while ignore the first half; conditional relation like "Only ... can ...", which we call the "emphasis preposition" type, while ignoring the latter part of the sentence we only consider the emotional intensity of the first half.

According to the above three aspects, we design the new algorithm based on section 3.3.2 algorithm to calculate the emotional intensity; the entire emotional score of the text is calculated as follows:

$$globalScore = \sum_{j=1}^{n} \sum_{i=1}^{t} \alpha^{DegreeWordsCnt} \cdot eScoreOfWord_i \cdot (-1)^{NegationWordsCnt} \cdot$$

$$2^{InterjectionWordsCnt-1+ExclamationMarkCnt} \cdot f(EffectOfQuestionMark)$$

(1)

In this formula, n represents the number of Clauses included in a comment, t for the number of words contained in a clause. $DegreeWordsCnt$ is the number of occurrences for the degree word and α is the corresponding weight. $eScoreOfWord_i$ is the emotional score of the $Word_i$, 1 for positive and -1 for negative. $NegationWordsCnt$, $InterjectionWordsCnt$ and $ExclamationMarkCnt$ respectively represent the frequency of occurrence of negation word, interjection and exclamation mark.

$f(EffectOfQuestionMark)$ indicates the role of a question mark. We believe that when a clause there contains an emotional word and ends with a question mark, its emotional polarity reverses, and the emotional score gets doubled with the increasing of question marks; otherwise the emotional score stays unchanged. The sign of the emotional score for the whole message is on behalf of the emotional polarity, while the size of the score represents the strength or intensity of the emotion.

4   Experimental Results and Analysis

Of Sina Weibo hot topics, we select three hot topics: "The act of Guang-biao Chen playing Feng Lei has been questioned, Chen claiming to be filled with sacredness", "Xiaomi Phone, it's not easy to say love you" and "Is court intrigue play junk", which respectively represents celebrities, products, film and television, and the corresponding original microblog links are: "http://weibo.com/1638782947/y8rjOhxA7", "http://weibo.com/1796445350/yesE4Fff8" and " http://weibo.com/1821898647/yilu UlUEE".

After the preprocessing of the data, the numbers of comments under three topics are 1342, 585 and 950. According to the analysis in section 3, we carry out a total of six experiments as follows:

Based on the initial emotion lexicon (IEL), without considering the negation rules, remove stop words;

Based on the initial emotion lexicon (IEL), consider the negation rules;

Based on the initial emotion lexicon (IEL) and the emoticons description lexicon (EDL), consider the negation rules;

Based on the initial emotion lexicon (IEL) and the network buzzwords lexicon (NBL), consider the negation rules;

Based on the initial emotion lexicon (IEL), the emoticons description lexicon (EDL) and the network buzzwords lexicon (NBL), consider the negation rules;

Based on the initial emotion lexicon (IEL), the emoticons description lexicon (EDL) and the network buzzwords lexicon (NBL), consider the negation rules and the three factors discussed in section 3.5.

4.1   The Emotional Polarity Classification Accuracy and the Impact Factor's Effect

After statistical calculations, we have come to the classification accuracy of the three topics in six cases as shown in the following table:

Table 1. The Emotional Polarity Classification Accuracy

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Topic 1	44.04%	46.20%	52.76%	73.40%	76.53%	79.96%
Topic 2	43.25%	52.82%	57.61%	70.26%	73.85%	81.20%
Topic 3	51.68%	57.79%	60.21%	76.63%	77.68%	84.42%

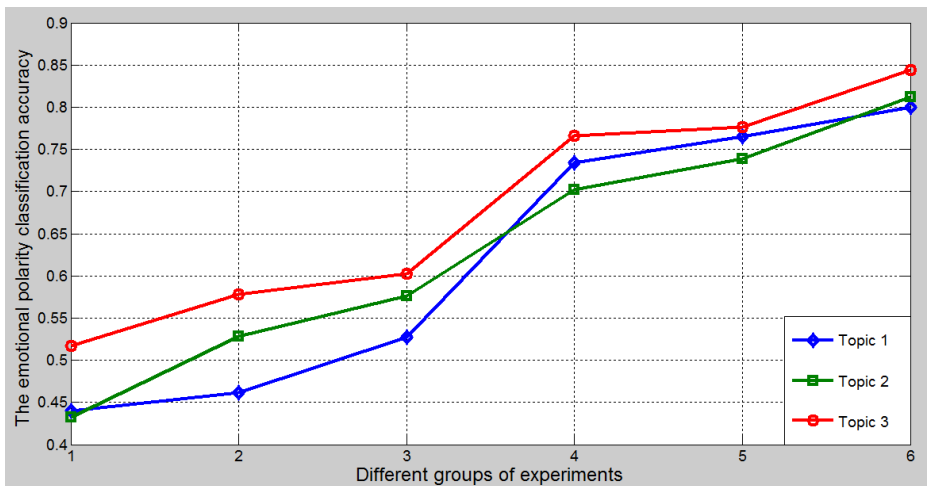The corresponding curves are shown in Figure 1.

Fig. 1. The Emotional Polarity Classification Accuracy Curves

Figure 1 visually shows that the emotional polarity classification accuracy gradually increases with the use of negation rules and the expansion of the emotional lexicon. The improvement between the third set of experiments and the fourth set of experiments is the maximum, showing that the introduction of the network buzzwords lexicon (NBL) greatly improve the classification results, which agrees with the fact that the microblog contains a large number of network buzzwords. With the experimental results of the three topics, we can find the promoting effect of the various impact factors (IF) as shown in Table 2.

From Table 2 we can see that the contribution to promote classification accuracy made by the network buzzwords lexicon (NBL) is the greatest. The effect of the negation rules and the combined three factors discussed in section 3.5 are approximately equal. The promoting effect of the emoticons description lexicon (EDL) is the smallest, only 4%, this may be due to the fact that emoticons are not common in the comments since not every message contains an emoticon.

Table 2. The Promoting Effect of the Various Impact Factors

	Negation Rules	EDL	NBL	Degree Words + Interjection + Sentence Pattern
Improvement	5.95%	3.59%	20.16%	5.84%

4.2 The Overall Attitude of the Comments on Each Topic

We also summarize the specific classification results of the three topics in the sixth set of experiments which is the best classification result, and form the classification result matrix [12], as shown respectively in Table 3 (corresponding to topic 1), Table 4 (corresponding to topic 2), Table 5 (corresponding to topic 3):

Table 3. The Classification Result Matrix of Topic 1

Manual annotation	Positive	Negative	Neutral
Positive	364	46	25
Negative	94	480	67
Neutral	41	12	213

Table 4. The Classification Result Matrix of Topic 2

Manual annotation	Positive	Negative	Neutral
Positive	169	12	11
Negative	43	204	9
Neutral	29	6	102

Table 5. The Classification Result Matrix of Topic 3

Manual annotation	Positive	Negative	Neutral
Positive	261	24	3
Negative	28	362	14
Neutral	17	62	179

From Table 3, Table 4 and Table 5 we can calculate the respective proportion of positive emotional comments, negative emotion comments and neutral emotion comments in correctly classified instances, as shown below:
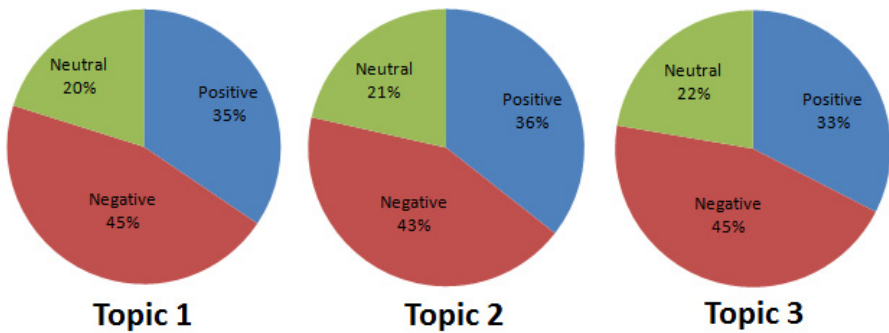

Fig. 2. The Respective Proportion of Positive, Negative and Neutral Comments

From Figure 2 we can see approximate view proportion of the comments on three topics: about 35% of the people hold positive opinions; about 45% hold negative opinions; the remaining 20% or so hold neutral opinions.

4.3   Precision Ratio and Recall Ratio Comparison

We also calculate the precision (P) and recall (R) of the various emotions for the three topics, as shown in Table 6:

Table 6. The Precision (P) and Recall (R)

Topics	Positive		Negative		Neutral	
	P	R	P	R	P	R
Topic 1	72.95%	83.68%	89.22%	74.88%	69.84%	80.08%
Topic 2	70.12%	88.02%	91.89%	79.69%	83.61%	74.45%
Topic 3	85.29%	90.63%	80.80%	89.60%	91.33%	69.38%

According to Table 6, the precision rate of negative emotion is higher than that of positive and neutral emotion in topic 1 and 2, which is inseparable with the fact that the negative emotion instances share the highest proportion of total (Table 3, Table 4

shows). Although the actual negative emotion instances are the most in topic 3 (Table 5 shows), the neutral emotion precision ratio is the highest, which is caused by the low probability to classify the positive emotion or negative emotion as the neutral since topic 3 is "Is court intrigue play junk", and the vast majority of reviewers clearly expressed their point of view, "yes" or "no". Whether in topic 1, topic 2 or topics 3, positive emotion recall is the highest, which is caused by the low probability of positive emotions classified as negative or neutral, for people mostly use positive or affirmative words and these words can be easily identified with the emotional lexicon, while negative emotions are often misidentified as positive or neutral due to the impact of rhetoric and sentence pattern, and neutral emotions can be easily divided into the positive or negative direction because of the subtle changes in emotional scores. In addition, high precision is often obtained in the case of sacrificial recall, while high recall is also at the expense of precision, which can be concluded from Table 6.

4.4   Error Analysis

In order to better understand the inadequacies of the experiment, we also extract the misclassification samples. After analysis, we summarize the following situations:

The text does not contain emotional words, emoticons, network buzzwords, interjection or other characteristics to express point of view, such as in topic 1, "This modeling ....". The author did not use any emotional words, but expressed the emotions of a "silent" feeling by playing a long series of dots. Again "There are no gyri in the man's head" indicates a negative emotion, but our algorithm misjudges it as neutral because of no emotional words extracted.

The multiple roles of question mark. We think that the question mark reverses the emotions polarity. For instance, "Do you think doing like this is right?" The fact is "Doing like this is inappropriate". This reversal is more common in rhetorical question, while the question mark can also be used in general interrogative sentence, such as "Carrying guns illegally???" containing negative emotional word "illegally"; if we consider the reversal effect of the question marks, then the emotion of the sentence would be determined as positive, resulting in an error.

The use of irony, contrast and other rhetorical devices. For example in topic 2, there is a review which is "Buying Xiaomi Phone? You must have confidence in its after-sale services! Ha ha!". Actually it's irony and ridicule about Xiaomi Phone's poor after-sale services. If we extract positive emotion word "confidence" and positive interjection "Ha ha", there will be a misjudgment. Another example is "Meizu doesn't have this problem. My director's Xiaomi Phone for Telecom, just has no signal after plugging in the SIM card, only to return to the plant for a new!" This review uses the contrast between Xiaomi Phone's instability and Meizu Phone's stability, which expresses negative emotions about Xiaomi Phone; if we simply extract a positive emotion word "new", then, the sentence emotional polarity will be misidentified as positive.

Subjects extraction. Although our topic is determined, one topic can involve other topics, as in topic 3 "This society is full of intrigue. 'Zhen Huan biography' is still worth seeing." The first part is critique of society, while the last part is the praise of the court intrigue play "Zhen Huan biography"; the author holds a positive attitude toward

the play. The algorithm will get negative emotional word "intrigue" and positive emotional word "worth seeing" extracted; because of no recognition of the emotional word modified subjects, the sentence emotional polarity is misjudged as neutral.

Dialects. Dialects have a certain probability of the occurrence in the network buzzwords. As in topic 1, "Such a tiger-like person even exists in this world!" If we simply understand the word "tiger-like" literally, it describes a person as burly and tall, so it is a positive emotion word. However, in the dialects of Northeast China it means "reckless, without considering the consequences", so we would make misclassification if we have no knowledge of the dialect.

5   Conclusion and Future Work

In this paper, we use Sina Weibo official API to extract comments of hot topics; after preprocessing of these data, we do text sentiment classification mainly based on emotional lexicons. On the basis of initial emotional lexicon (IEL), we introduce the emoticons description lexicon (EDL), the network buzzwords lexicon (NBL), the interjection lexicon and so on, and consider the "negative transfer" phenomenon, the modification of the degree words, the sentence pattern and other impact factors, and finally design six different groups of the experiments. The experiments show that microblog comments text sentiment classification accuracy can reach 80% or more. The consideration of negation rules, the introduction of emoticons description lexicon and network buzzwords lexicon, and the calculation of the emotional intensity have different contributions to the improvement of emotion classification accuracy. The paper also analyzes the reason of classification error.

Although our emotional polarity classification accuracy can reach up to 80%, there are still some problems to be settled, such as the impact of the multiple roles of the question mark, the extraction of the subjects the emotional word modifies.

References

1. Zhao, Y.Y., Qin, B., Liu, T.: Text Sentiment Analysis. Journal of Software 21(8), 1834–1848 (2010)
2. Bo, P., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86 (2002)

3. Read, J.: Using Emoticons to reduce Dependency in Machine Learning Techniques for Sentiment Classification. In: ACL Student 2005 Proceedings of the ACL Student Research Workshop, pp. 43–48 (2005)
4. Tang, H.-F., Tan, S.-B., Cheng, X.-Q.: Comparative Study of Chinese Emotion Classification Technique based on Supervised Learning. Journal of Chinese Information Processing 21(6), 88–94 (2007)
5. Xu, J., Ding, Y.-X., Wang, X.-L.: Sentiment Automatic Classification for News using Machine Learning Methods. Journal of Chinese Information Processing 21(6), 95–100 (2007)
6. Van Wanzeele, R., Verbeeck, K., Vorstermans, A., et al.: Extracting emotions out of twitter's microblogs. In: Proceedings of the 23rd Benelux Conference on Artificial Intelligence (2011)
7. Pak, A., Paroubek, P.: Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In: Proceedings of LREC 2010 (2010)
8. Sina Weibo, http://www.weibo.com
9. Liu, Z.-M., Liu, L.: Empirical Studies of the Chinese Microblog Sentiment Classification based on Machine Learning. Computer Engineering and Applications (1), 1–4 (2012)
10. HowNet, http://www.keenage.com/html/c_index.html
11. ICTCLAS, http://ictclas.org/
12. Liu, B.: Web Data Mining, pp. 57–59. Tsinghua University Press, Beijing (2009)

Research on Fast Loading Large Scale
Point Cloud Data File

Jiansheng Zhang[1,2]

[1] School of Manufacturing Science and Engineering, Southwest University of Science
and Technology, 621010 Mianyang, China
[2] Key Lab of Reverse Engineering and Rapid Manufacturing in Sichuan,
Southwest University of Science and Technology, 621010 Mianyang, China
zjs_1980@163.com

Abstract. To meet the requirement of fast loading large scale point could data
file, proposed a new method based on memory pre-allocation and multiple-point
concurrency writing technologies which depend on thread pool and memory
mapping file mechanism under Windows platform. Test result shows that the
new method based on memory pre-allocation and multiple-point concurrency
writing technologies improve the performance of loading large-scale point-could
data file notably with 220%~300% in loading efficiency, and along with the file
size increment, the extra loading time caused by the method is the least compared
with the other three methods.

Keywords: Point Cloud, Fast Loading, Large Scale, Memory Pre-allocation,
Multiple-point Concurrency Writing.

1   Introduction

With the rapid development of reverse engineering technology, 3D scanning hardware
gains great improvement in both scale and performance [1-6]. On the one hand, 3D
point cloud data representing the object surface could be easily acquired, and on the
other hand the size of acquired 3D point cloud data goes up in a geometric progression.
For processing point cloud data, loading those point data from disk files into system
memory is the first thing for any application [7-12]. It is significantly important to load
large scale point cloud data file as fast as possible because users' experience for the
application will be seriously affected by the time delay during loading points into
system memory. For the problem about fast loading point cloud data, the attention is
concentrated on how to minimize the time delay of reading disk file in the condition
that system memory is under certain size.

The paper first introduced three common ways of loading point cloud data file, and
then proposed the notion of memory pre-allocation and multiple-point concurrency
writing technologies which were also implemented, and at last made comparison be-
tween this new method and other methods with conclusions drawn.

2   Three Ways of Loading Point Cloud Data File

There are basic three ways of loading point cloud data file for the present [13-16], and among all the methods, sequence reading method should be the simplest and most straightforward. The basic principle of this method is to read one line of data each time in the data file, and parse the line of data as x, y, and z coordinates, at last save these coordinates as a point into data array.

Block reading method uses the notion of memory pre-allocation, by which a certain size of data is read from disk file through a single operation and these data are saved into pre-allocated memory, at last parsing operation is performed on these data. This kind of method improves the throughput of IO operation, and greatly speeds up the acceleration of loading point cloud data file.

The memory mapping file under Windows platform is a kernel object which could be used as an alternative mechanism for conventional IO operation, and it could be utilized in many applications [17-21]. The key idea of memory mapping file mechanism is to make full use of the APIs provided by operating system to complete the IO operation in memory which should have been completed in external disk files, so the performance will be greatly improved compared with conventional IO operation.

All points in the point cloud data file form a point set P, and for the definition of fast loading, it is to choose a best method R which satisfies the time of operation $T=R(P)$ is the least when R is applied on the point cloud data file compared with other methods [22-26].

All the three loading methods above could be easily implemented using basic programming techniques and operating system knowledge, and loading performance could be tested respectively and compared. However, there are some questions left for us. Are these methods appropriate for large scale point cloud data file since the loading methods are all in a sequential way? Could we make full use of the power of modern CPUs with mutli-cores since each core of CPU with hyperthread technology could execute a separate computation line? How about the memory? Could we read or write more than one address point of memory at the same time in order to speed up memory access? All these questions are important clues to bring up the method of fast loading large scale point cloud data file.

Twenty-one point cloud data files were automatically generated by program for test. The size of files varies from 10MB to 1GB while the number of points in the files varies from 374456 to 38349417 correspondingly. By applying each loading method on these point cloud data files, we could find out the differences of loading performance among these methods [27-32]. For the just the three loading methods above, the comparison result could be illustrated as Fig. 1.
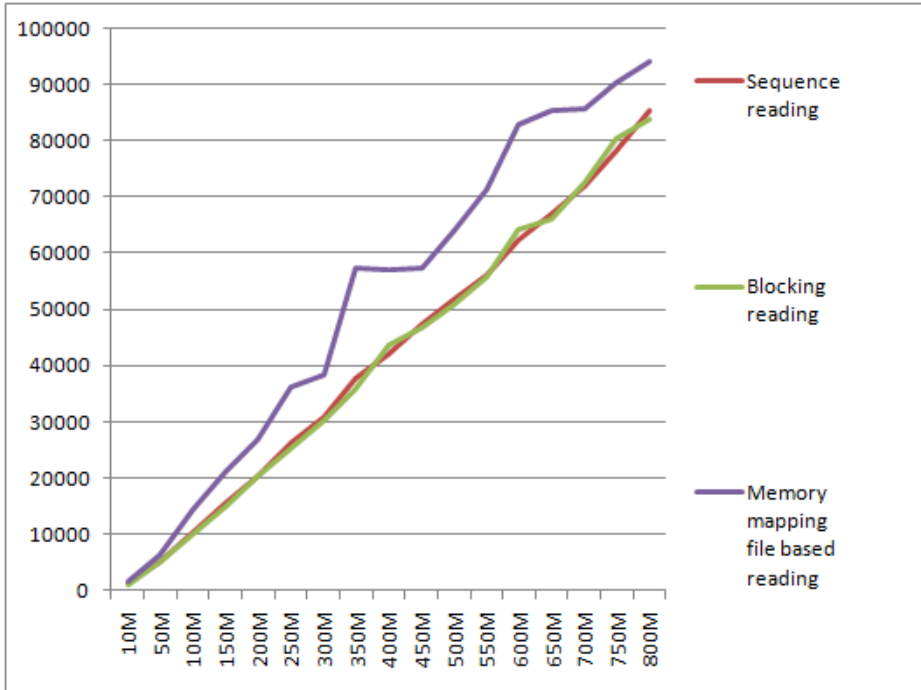
Fig. 1. Performance comparison among the three reading methods

3   Principle of Memory Pre-allocation and Multiple-point Concurrency Writing

There are two main problems about loading point cloud data file for the three methods described above [33-36].

3.1   Dynamic Memory Allocation Caused by Storage of Each Read Point Data

Since the number of points in point cloud data could be very huge, frequent dynamic memory allocation will seriously slow the application performance down.

3.2   The Sequence of Reading Operation of Point Cloud Data File

For the sequence reading method and block read method mentioned above, the operation principle are both based on the file handle, and it limits that the reading operation should be done in a sequential way. For the memory mapping file based method, after the point cloud data file on disk is mapped to memory, concurrent operation could be executed on those memory.

Based on the analysis made above, the paper proposed the memory pre-allocation and multiple-point concurrency writing technologies which solve the problems of loading point cloud data files, especially those are in a large scale, so the application performance will be greatly improved. Both the memory pre-allocation technology and multiple-point concurrency writing technology are based on thread pool and memory mapping file mechanism, and thread pool provides the technical support for concurrency operations while memory mapping file mechanism provides the technical support for concurrency IO operations.

3.3   Principle of Memory Pre-allocation

The problem that memory pre-allocation technology aims to solve is to avoid frequent dynamic memory allocation caused by saving operations during loading point could data file. Since data type of the three coordinates of each point is the same, so the size of memory for one point is fixed. As long as the number of points is determined, the total memory required by all points is also determined. So application could allocate exact memory for all points in advance, which could avoid dynamic memory allocation.
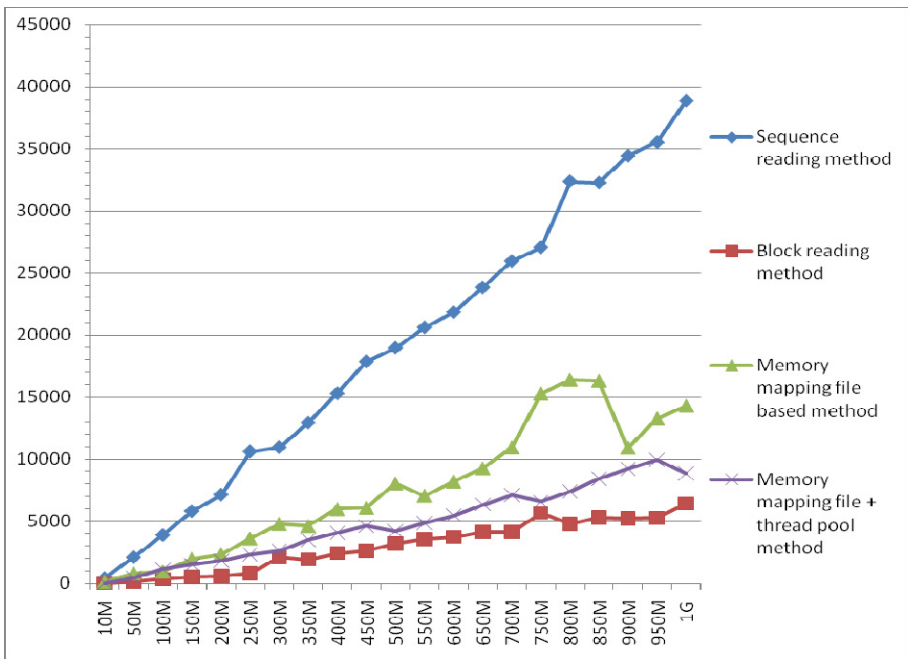

Fig. 2. Comparison among methods of number of points determination

There are three ways to determine how many points in a point cloud data file corresponding to the sequence reading method, block reading method and memory mapping file based method, whereas in implementation the coordinates are not parsed.

The fourth method which is multi-threaded and based on memory mapping file combines both blocking method and memory mapping file based method, and it takes the advantages of thread pool. The key of the method is each thread in the thread pool reads a block of data of disk file which has been mapped into memory, and then determines the number of points in the block of data. After that, the thread will continue to do the same thing until all of data have been processed. At last, accumulates the number of points of each block of data, and the result is the total number of points of the point cloud data file.

The four methods described above which could be used to determine the number of points in point cloud data file were tested, and the result is illustrated as Fig.2.

It could be found in the figure that compared with the other three methods, the block reading method can determine the number of points in the minimum duration, and the following two are memory mapping file based method and multi-thread based method. The result is different from the expectation that multi-thread and memory mapping file based method should be the fastest. However, taking the application condition of memory mapping file and the computer system environment into consideration, we noticed that during the point number determination, only the comparison operation between '\n' and several characters in the block data was performed, and there are no operations of parsing coordination happen. So the result could be acceptable and reasonable. Why the memory mapping file based method and multi-thread based method is slower in determining number of points than that of block read method should charge upon the many operations in processing system resources and maintain the internal state both for thread pool and memory mapping file. The memory mapping file based method depends upon operating system heavily, and the testing results could be contrary to expectations. But from a global point of view, the performance of this method tends to be stable.

3.4   Principle of Multiple-point Concurrency Writing

The kernel of multi-point concurrency writing technology is to convert the IO operation from conventional sequence operation to concurrent operation by making use of memory mapping file mechanism. Just as the description above, the mechanism of memory mapping file could map external disk files into system memory, so it is possible for users to process memory in a concurrent manner. The principle of multiple-point concurrency writing technology based on memory mapping file and thread pool could be illustrated as Fig.3.

It could be found that because of concurrency support by memory access, the concurrency IO operations could be applied as soon as the disk file is mapped into system memory. What's more, there is no conflicts among memory writings (storing points) caused by threads after coordinates are parsed, so the performance could be greatly improved in loading point cloud data files, especially the ones those are very large.
Pseudo-code of the multiple-point concurrency writing technology could be illustrated as follows:

```
pre-allocate memory buffer gbuf for point storage;
open a point-cloud data file on disk;
```

```
mapping file data into memory;
get address pbuffer of mapped memory;
create a thread pool object tp;
while all data in memory are not processed
generate a task where data in memory and how much data should
be processed for parsing point coordinates, and write them
correspondingly in gbuf;
assign a thread to do such a task;
end while
wait each thread to complete its task;
close the opened file;
```
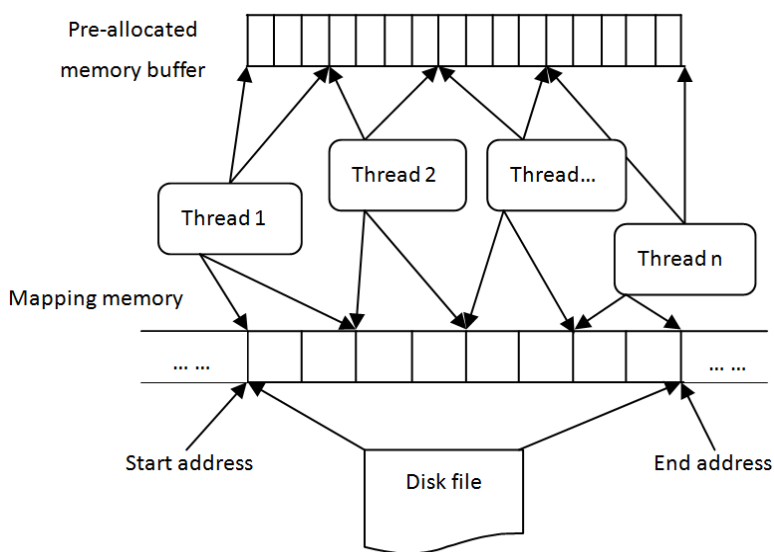

Fig. 3. Principle of multiple-point concurrency writing

One of the key problems in implementing the multiple-point concurrency writing technology is to parse the data mapped into system memory to point coordinates, and writing them into the pre-allocated memory. It is necessary to calculate the address of memory where and how many each thread will access, and to calculate the address of pre-allocated memory where the thread will write, both these information should be passed to each thread before its execution.

4   Test Results

The paper made tests using the four different methods to load point cloud data files, and the result could be illustrated by Fig.4. Test result shows that the new method based on

memory pre-allocation and multiple-point concurrency writing technologies improve the performance of loading large-scale point-could data file notably with 220%~300% in loading efficiency, and along with the file size increment, the extra loading time caused by the method is the least compared with the other three methods.
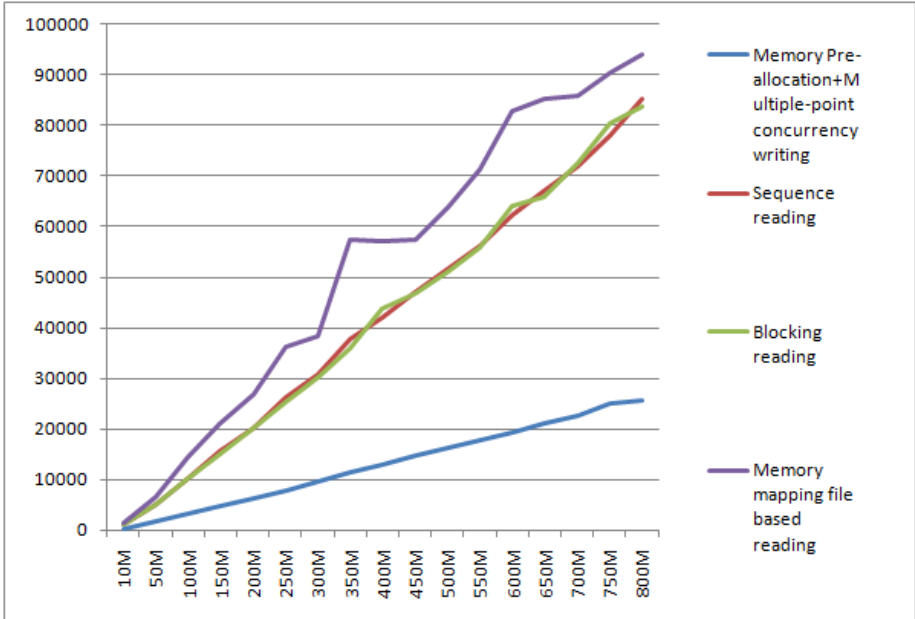

Fig. 4. Performance comparison among difference reading methods

The scale of processing point cloud data varies from hundreds of points to hundreds of millions of points in different applications. In narrow sense, large scale point cloud data confines to the number of points to be processed such as hundreds of points or millions of points while in broad sense large scale point cloud data involves with system memory. Processing hundreds of millions of points in a computer system with large physical memory in TBs is not very difficult, while it is very hard to process tens of thousands of points in a computer system with less physical memory. For the paper, the attention is concentrated on how to minimize the time delay of reading disk file in the condition that system memory is under certain size.

5   Conclusion

Aim at the requirements of loading point cloud data file especially those are very large, the paper proposed a new method of fast loading point cloud data file based on memory pre-allocation and multiple-point concurrency writing technologies after analyzing and comparing several commonly used loading methods. On the other side, in broad large-scale point cloud data concept, all of the point cloud data may not be mapped into system memory, and under the condition it is necessary to use blocking mapping

method to map the entire point cloud data file into system memory in many times for concurrent processing. Test results show that adopting the new method to loading point cloud data file could greatly improve the performance of application, and it provides good environment for point processing later. The method could be used almost in all applications those are point-based.

References

1. Adams, B., Wicke, M., Dutré, P., Gross, M., Pauly, M., Teschner, M.: Interactive 3D painting on point-sampled objects. In: Proceedings of the Eurographics Symposium on Point-based Graphics, pp. 57–66 (2004)
2. Kazhdan, M.: Reconstruction of solid models from oriented point sets. In: Symposium on Geometry Processing, Vienna, Austria, pp. 73–82 (2005)
3. Hu, W.-Z., Liu, N., Liu, R.-Y.: Quick-read Means for Large Scale Images Based on Storage Mapping File Technique. Computer Application Research 2, 111–112, 107 (2005)
4. Zhu, G.-S., Zhou, T.-R., Pang, H.-P.: A New Rapid-reading method for Point Cloud of Laser Scanner. Computer System Application 1, 102–104 (2007)
5. Yang, N.-X., Zhu, C.-Q., Nie, A.-L.: Memory-mapped Files and It's Application on Fast Read/Write from Mass Data Files. Application Research of Computers 8, 187–188 (2004)
6. Li, X.-G., Wang, Z.-M., Huang, Z.-Q.: on memory mapping file technique to visualize large scene watershed model quickly. Application Research of Computers 26, 527–528 (2009)
7. Liu, S., Ye, Y.: Source: Point cloud segmentation using gradient vector flow snake. In: 2011 International Conference on Information Science and Technology, pp. 1114–1118 (2011)
8. Hou, W., Chan, T., Ding, M.: Denoising point cloud. Inverse Problems in Science and Engineering, 287–298 (2012)
9. Liu, S., Xie, X.: Research on algorithm of point cloud MapReduce registration. In: 2011 IEEE International Conference on Cloud Computing and Intelligence Systems, pp. 338–341 (2011)
10. Bouvier, D.J., Gordon, C., McDonald, M.: An approach for occlusion detection in construction site point cloud data. In: Proceedings of the Congress on Computing in Civil Engineering, pp. 234–241 (2011)
11. Li, F., Tang, R., Liu, C., Yu, H.: A method for object reconstruction based on point-cloud data via 3D scanning. In: Proceedings of the 2010 International Conference on Audio, Language and Image Processing, pp. 302–306 (2010)
12. Chen, J.-Y., Lai, H.-J., Lin, C.-H.: Point cloud modeling using algebraic template. International Journal of Innovative Computing, Information and Control, 1521–1532 (2011)
13. Levoy, M., Whitted, T.: The use of points as a display primitive. Technical report, Computer Science Department, University of North Carolina at Chapel Hill (January 1985)
14. Hoppe, H., DeRose, T., Duchamp, T., McDonald, J., Stuetzle, W.: Surface reconstruction from unorganized points. In: Catmull, E.E. (ed.) Computer Graphics, pp. 71–78 (1992)
15. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Symposium on Geometry Processing, pp. 43–52 (2006)

16. Tian, H.-S., He, Y.-J., Cai, H.-M.: Survey of Point-based Computer Graphics. Journal of System Simulation, 42–45 (2006)
17. Qiu, H., Chen, L.-T.: Research and Development of Point-based Computer Graphics. Computer Science, 10–15 (2009)
18. Thomas, K., Tatiana, T., Tobias, S., Frank, Z.: GPU-accelerated 2D point cloud visualization using smooth splines for visual analytics applications. In: Proceedings - SCCG 2008: 24th Spring Conference on Computer Graphics, pp. 97–104 (2010)
19. Ardagna, D., Casolari, S., Colajanni, M., Panicucci, B.: Dual time-scale distributed capacity allocation and load redirect algorithms for cloud systems. Journal of Parallel and Distributed Computing, 796–808 (2012)
20. Ma, H., Zongyue, W.: Distributed data organization and parallel data retrieval methods for huge laser scanner point clouds. Computers & Geosciences, 193–201 (2011)
21. Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R.H., Konwinski, A., Lee, G., Patterson, D.A., Rabkin, A., Stoica, I., Zaharia, M.: Above the clouds: A berkeley view of cloud computing, University of California, Berkeley, Tech. Rep. USB-EECS-2009-28 (February 2009)
22. 104th United States Congress, Health Insurance Portability and Accountability Act of 1996 (HIPPA) (1996), `http://aspe.hhs.gov/admnsimp/pl104191.htm`
23. Harney, H., Colgrove, A., McDaniel, P.D.: Principles of policy in secure groups. In: Proc. of NDSS 2001 (2001)
24. McDaniel, P.D., Prakash, A.: Methods and limitations of security policy reconciliation. In: Proc. of SP 2002 (2002)
25. Yu, T., Winslett, M.: A unified scheme for resource protection in automated trust negotiation. In: Proc. of SP 2003 (2003)
26. Li, J., Li, N., Winsborough, W.H.: Automated trust negotiation using cryptographic credentials. In: Proc. of CCS 2005 (2005)
27. Kallahalla, M., Riedel, E., Swaminathan, R., Wang, Q., Fu, K.: Scalable secure file sharing on untrusted storage. In: Proc. of FAST 2003 (2003)
28. Goh, E., Shacham, H., Modadugu, N., Boneh, D.: Sirius: Securing remote untrusted storage. In: Proc. of NDSS 2003 (2003)
29. Ateniese, G., Fu, K., Green, M., Hohenberger, S.: Improved proxy re-encryption schemes with applications to secure distributed storage. In: Proc. of NDSS 2005 (2005)
30. di Vimercati, S.D.C., Foresti, S., Jajodia, S., Paraboschi, S., Samarati, P.: Over-encryption: Management of access control evolution on outsourced data. In: Proc. of VLDB 2007 (2007)
31. Goyal, V., Pandey, O., Sahai, A., Waters, B.: Attribute-based encryption for fine-grained access control of encrypted data. In: Proc. of CCS 2006 (2006)
32. Blaze, M., Bleumer, G., Strauss, M.: Divertible protocols and atomic proxy cryptography. In: Nyberg, K. (ed.) EUROCRYPT 1998. LNCS, vol. 1403, pp. 127–144. Springer, Heidelberg (1998)
33. Wang, Q., Wang, C., Li, J., Ren, K., Lou, W.: Enabling public verifiability and data dynamics for storage security in cloud computing. In: Backes, M., Ning, P. (eds.) ESORICS 2009. LNCS, vol. 5789, pp. 355–370. Springer, Heidelberg (2009)
34. Yu, S., Ren, K., Lou, W., Li, J.: Defending against key abuse attacks in KP-ABE enabled broadcast systems. In: Chen, Y., Dimitriou, T.D., Zhou, J. (eds.) SecureComm 2009. LNICST, vol. 19, pp. 311–329. Springer, Heidelberg (2009)
35. Naor, D., Naor, M., Lotspiech, J.: Revocation and tracing schemes for stateless receivers. In: Kilian, J. (ed.) CRYPTO 2001. LNCS, vol. 2139, pp. 41–62. Springer, Heidelberg (2001)
36. Atallah, M., Frikken, K., Blanton, M.: Dynamic and efficient key management for access hierarchies. In: Proc. of CCS 2005 (2005)

Fast Distributed Mining Algorithm of Maximum Frequent Itemsets Based on Cloud Computing

Bo He[1,2,3]

[1] School of Computer Science and Engineering, Chongqing University of Technology, ChongQing, 400054, China
[2] State Key Laboratories for Novel Software Technology, Nanjing University, 210093, China
[3] Shenzhen Key Laboratory of High Performance Data Mining, Shenzhen, 518055, China
heboswnu@sina.com

Abstract. The paper proposed a fast distributed mining algorithm of maximum frequent itemsets based on cloud computing, namely, FDMMFI algorithm. FDMMFI algorithm made nodes compute local maximum frequent itemsets by cloud computing, then the center node exchanged data with other nodes and combined, finally, global maximum frequent itemsets were gained by cloud computing. Theoretical analysis and experimental results suggest that under the same minimum support threshold, communication traffic and runtime of FDMMFI decreases while comparing with CD and FDM. The less the minimum support threshold, the better the three performance parameters of FDMMFI.FDMMFI algorithm is fast and effective.

Keywords: Data Mining, Cloud Computing, Maximum Frequent Itemsets.

1 Introduction

There are some traditional algorithms for mining maximum frequent itemsets [1] in local database, such as Max-Miner [2] and DMFIA [3]. There are also some algorithms for mining global frequent itemsets, include, CD [4], FDM [5], etc. However, the database for mining maximum frequent itemsets is generally distributed; these mining algorithms do not suit mining of global maximum frequent itemsets. Most of them adopt Apriori-like algorithm framework, so that a lot of candidate itemsets are generated and database is scanned frequently [6-10]. This causes a large amount of communication traffic among nodes. Aiming at these problems, this paper proposes a fast distributed mining algorithm of maximum frequent itemsets based on cloud computing, namely, FDMMFI algorithm.

2   Related Definition and Theorem

2.1   Description of Distributed Mining of Maximum Frequent Itemsets Based on Cloud Computing

The global transaction database is *DB* [11-15], the total number of tuples is *M*. Suppose $P_1, P_2, \ldots, P_n$ are *n* nodes, node for short, there are M_i tuples in DB_i, if DB_i $(i=1,2,\ldots,n)$ is

a part of *DB* and stores in P_i, then $DB = \bigcup\limits_{i=1}^{n} DB_i$, $M = \sum\limits_{i=1}^{n} M_i$.

Distributed mining of maximum frequent itemsets based on cloud computing can be described as follows: each node P_i deals with local database DB_i by cloud computing, and communicates with other nodes, finally, global maximum frequent itemsets of global transaction database are gained by cloud computing.

2.2   Related Definition

Definition 1. For itemsets *X*, the number of tuples which contain *X* in local database $DB_i(i=1,2,\ldots,n)$ is defined as local frequency of *X*, symbolized as *X.si* .

Definition 2. For itemsets *X*, the number of tuples which contain *X* in global database is global frequency of *X*, symbolized as *X.s* .

Definition 3. For itemsets *X*, if $X.si \geq min\_sup*M_i(i=1,2,\ldots,n)$, then *X* are defined as local frequent itemsets of DB_i, symbolized as F_i. If $|X|=k$, then *X* symbolized as F_i^k , and *min_sup* is the minimum support threshold.

Definition 4. For itemsets *X*, if $X.s \geq min\_sup*M$, then *X* are defined as global frequent itemsets, symbolized as *F*. If $|X|=k$, then *X* symbolized as F_k.

Definition 5. For global frequent itemsets *X*, if all superset of *X* are not global frequent itemsets, then *X* are defined as global maximum frequent itemsets, symbolized as *FM*.

2.3   Related Theorem

Theorem 1. If itemsets *X* are local frequent itemsets of DB_i, then any nonempty subset of *X* are also local frequent itemsets of DB_i.

Proof: If itemsets *X* are local frequent itemsets of DB_i, then $X.si \geq min\_sup \times M_i$, if $Y \subseteq X$, then $Y.si \geq X.si$, hence $Y.si \geq X.si \geq min \, sup \times M_i$, *Y* are local frequent itemsets of DB_i.

Corollary 1. If itemsets *X* are not local frequent itemsets of DB_i, then the superset of *X* must not be local frequent itemsets of DB_i.

Theorem 2. If itemsets X are global frequent itemsets, then X and all nonempty subset of X are at least local frequent itemsets of a certain local database.

Proof: Suppose n local database are n pigeon nests, the times of itemsets X appearing in DB is concord with the number of pigeons. Because X are global frequent itemsets of DB, $X.s \geq (M_1 + M_2 + ... + M_n) \times \min\_ \sup$. According to pigeon nest principle, there's at least one local database DB_i which makes $X.si \geq \min\_ \sup \times M_i$, namely, X are local frequent itemsets of DB_i. According to theorem 1, any nonempty subset of X is local frequent itemsets of DB_i.

Theorem 3. If itemsets X are global frequent itemsets, then any nonempty subset of X are also global frequent itemsets.

Corollary 2. If itemsets X are not global frequent itemsets, then superset of X must not be global frequent itemsets.

Theorem 4. If itemsets X are global maximum frequent itemsets, then X must be global frequent itemsets.

Theorem 5. If itemsets X are global maximum frequent itemsets, then X and all nonempty subset of X are at least local frequent itemsets of a certain local database.

Proof: If itemsets X are global maximum frequent itemsets, then X are global frequent itemsets. According to theorem 2, X are at least local frequent itemsets of a certain local database.

Corollary 3. If itemsets X are not local frequent itemsets of any local database, then X must not be global maximum frequent itemsets.

Theorem 6. If itemsets X are global maximum frequent itemsets, then X are at least the subset of local maximum frequent itemsets in a certain local database.

Proof: If itemsets X are global maximum frequent itemsets, then X are at least local frequent itemsets of a certain local database according to theorem 5. Hence X are at least the subset of local maximum frequent itemsets in a certain local database.

Corollary 4. If itemsets X are not the subset of local maximum frequent itemsets of any local database, then X must not be global maximum frequent itemsets.

2.4   FP-Tree and FP-Growth Algorithm

Definition 6. FP-tree is a tree structure defined as follow.

(1) It consists of one root labeled as "null", a set of itemset prefix subtrees as the children of the root, and a frequent itemset header table.

(2) Each node in the itemsets prefix subtree consists of four fields: item-name, count, parent and node-link.

(3) Each entry in the frequent-item header table consists of three fields: i, Itemname. ii, Side-link, which points to the first node in the FP-tree carrying the item-set. iii, Count, which registers the frequency of the item-name in the transaction database.

FP-growth algorithm adopts a divide-and-conquer strategy. It only scans the database twice and does not generate candidate itemsets. The algorithm substantially reduces the search costs. The study on the performance of the FP-growth shows that it is efficient and scalable for mining both long and short frequent patterns, and is about an order of magnitude faster than the Apriori algorithm.

3   FDMMFI Algorithm

3.1   Design Thoughts of FDMMFI Algorithm

FDMMFI sets one node P_0 as the center node; other nodes P_i send local maximum frequent itemsets FM_i to the center node P_0. P_0 gets local maximum frequent itemsets FM' ($FM' = \bigcup_{i=1}^{n} FM_i$). FM' is pruned by the searching strategy of top-down [17-21]. Setting of the center node avoids repetitive calculation which caused by local frequent itemsets existing in many nodes.

FM' are pruned by the searching strategy of top-down, Pruning lessens communication traffic.

Each node adopts DMFIA [22-25] algorithm to compute local maximum frequent itemsets in FDMMFI. Adopting FP-tree structure, DMFIA algorithm greatly reduces database scanning times and runtime compared with Apriori-like algorithm.

The acquirement of global frequent items is the first step of FDMMFI. P_i scans DB_i once and computes local frequency of local items E_i. P_0 collects global frequency of all items E_i from each node and gets all global frequent items E. Finally, E is sorted in the order of descending support count. P_0 sends E to other nodes P_i.

Using global frequent items E, FDMMFI makes each node P_i construct $FP\text{-}tree^i$. P_i computes local maximum frequent itemsets FM_i independently by DMFIA algorithm and $FP\text{-}tree^i$, then the center node P_0 exchanges data with other nodes and combines by the searching strategy of top-down, finally, global maximum frequent itemsets are gained. According to theorem 6, global maximum frequent itemsets are at least the subset of local maximum frequent itemsets of one local database; hence the union of each node's local maximum frequent itemsets FM_i must be the superset of global maximum frequent itemsets FM.   Computing local maximum frequent itemsets may be carried out asynchronously, and synchronization is implemented only twice. This lessens synchronization.

3.2   Description of FDMMFI Algorithm

The step of FDMMFI is described as follows.

Each node adopts DMFIA [3] algorithm to compute local maximum frequent itemsets FM_i independently by cloud computing. At first, each node P_i acquires global frequent items E according to section 3.1. Second, each node P_i constructs $FP\text{-}tree^i$ using global frequent items E. At last, P_i computes local maximum frequent itemsets FM_i by DMFIA algorithm and $FP\text{-}tree^i$.

Each node sends their local maximum frequent itemsets FM_i to the center node P_0.

P_0 gets FM' ($FM' = \bigcup_{i=1}^{n} FM_i$).

P_0 adopts the searching strategy of top-down to get global maximum frequent itemsets FM by cloud computing.

The pseudocode of FDMMFI is described as follows.

Alogrithm FDMMFI

```
Input: The local transaction database DBᵢ which has Mᵢ tuples
```
and $M = \sum_{i=1}^{n} M_i$, `n nodes Pᵢ(i=1,2,…n), the center node P₀, the`
```
minimum support threshold min_sup.
Output: The global maximum frequent itemsets FM.
Methods: According to the following steps.
Step1. /*each node adopts DMFIA[3] algorithm to produce local
maximum frequent itemsets by cloud computing */
for(i=1;i<=n;i++) /*gaining global frequent items by cloud
computing */
{Scanning DBᵢ once;
Computing local frequency of local items Eᵢ;
Pᵢ sends Eᵢ and local frequency of Eᵢ to P₀;
P₀ collects global frequent items E from Eᵢ;
E is sorted in the order of descending support count;
P₀ sends E to other nodes Pᵢ;    /*transmitting global frequent
items to other nodes Pᵢ */
for(i=1;i<=n;i++)
{creating the FP-treeⁱ;    /*FP-treeⁱ represents FP-tree of DBᵢ
*/
FMᵢ =DMFIA(FP-treeⁱ, min_sup); /* node adopts DMFIA algorithm
to produce local maximum frequent itemsets FMᵢ aiming at
FP-treeⁱ*/
```

```
step2./* P_0 gets the union of all local maximum frequent
itemsets */
for(i=1;i<=n;i++)
```

P_i sends FM_i to P_0; /* FM_i represents local frequent itemsets
of P_i */

P_0 combines FM_i and produces FM';   /* $FM' = \bigcup_{i=1}^{n} FM_i$, represent

```
the union of all local maximum frequent itemsets*/
step3./*P_0 gets global maximum frequent itemsets according to
the searching strategy of top-down and cloud computing */
```

$FM = \varnothing$;

While $FM' \neq \varnothing$

{P_0 confirms the largest size k of itemsets in FM';

```
for all itemsets Q∈ local frequent k-itemsets in FM'  /* P_0
collects global frequency of Q from  other nodes P_i; */
if Q are not the subset of any itemsets in FM
{ P_0 broadcasts Q;
```

P_i sends $Q.si$ to P_0;   /* P_i computes local frequency $Q.si$ of
Q according to $FP\text{-}tree^i$ */

$Q.s = \sum_{i=1}^{n} Q.si$;   /* $Q.s$ represents global frequency of Q*/

```
if Q.s>=min_sup*M  /* Q are global maximum frequent
itemsets*/
{ FM=FM∪ Q;
P_0 deletes Q and any nonempty subset of Q from FM';
else  /* Q are not global maximum frequent itemsets*/
P_0 deletes Q and any nonempty subset of Q from FM';
for all item x∈ Q
if Q −{x} are not the subset of any itemsets in FM
FM'= FM' ∪{Q −{x}};
```

4   Experiments of FDMMFI

This paper compares FDMMFI with classical distributed mining algorithms CD and
FDM. The experimental data comes from the sales data in August 2009 of a
supermarket. The total number of tuples is about 20000.

Comparison experiment: It is a way of changing the minimum support threshold while adopting fixed number of nodes. FDMMFI compares with CD and FDM in terms of communication traffic and runtime. The results are reported in Fig.1
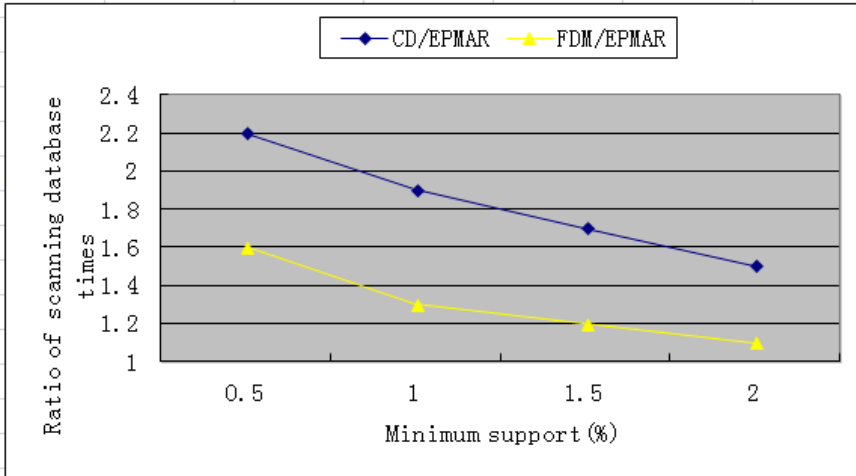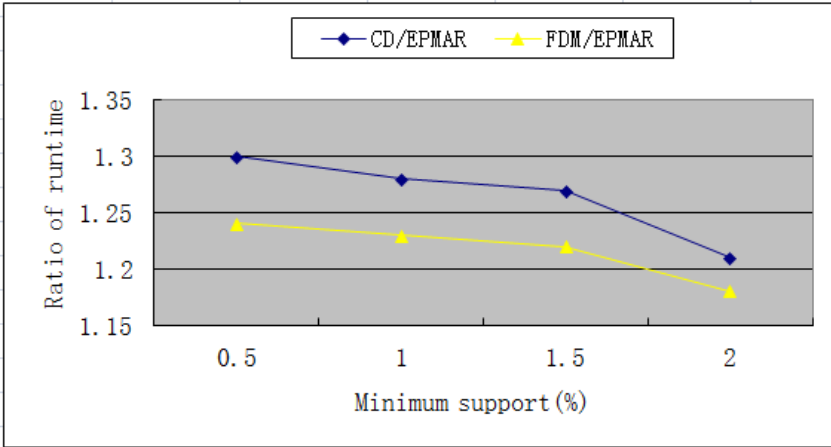

(a)Ratio of scanning database
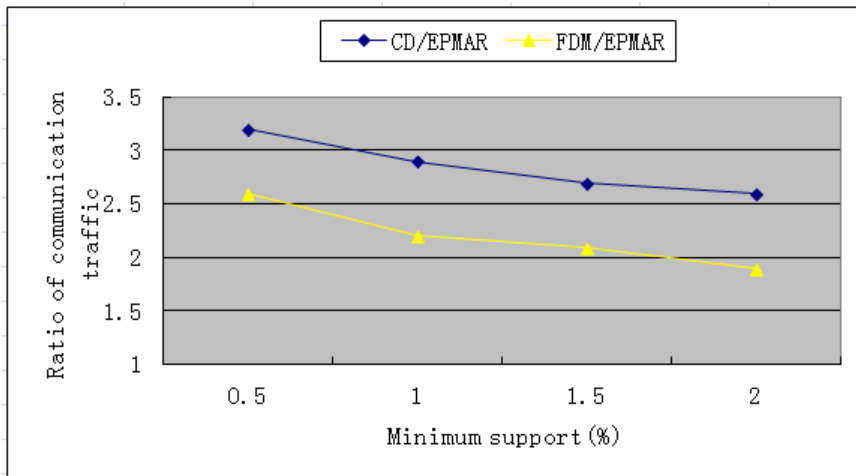
(b)Ratio of communication

Fig. 1. Comparison of communication traffic (a) Ratio of scanning database (b) Ratio of communcation (c) Ratio of runtime (d) Ratio of each node's

(c) Ratio of runtime
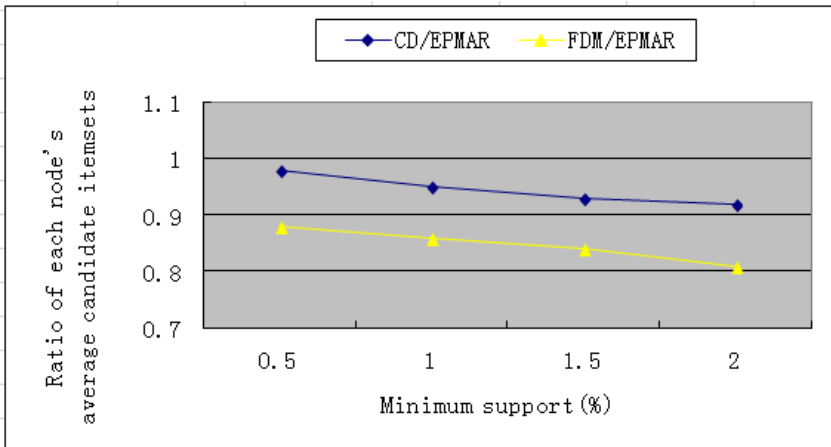
(d) Ratio of each node's

Fig. 1. (*Continued.*)

The comparison experiment results indicate that under the same minimum support threshold, communication traffic and runtime of FDMMFI decreases while comparing with CD and FDM. The less the minimum support threshold, the better the three performance parameters of FDMMFI.

5   Conclusion

FDMMFI makes nodes calculate local frequent itemsets by cloud computing, then the center node exchanges data with other nodes and combines by the searching strategy of top-down. At last, global maximum frequent itemsets are gained by cloud computing.

Theoretical analysis and experimental results suggest that FDMMFI is fast and efficient.

References

1. Mao, Y.X., Le Shi, B.: AFOPT-tax: An efficient method for mining generalized frequent itemsets. In: Nguyen, N.T., Le, M.T., Świątek, J. (eds.) ACIIDS 2010. LNCS, vol. 5990, pp. 82–92. Springer, Heidelberg (2010)
2. Bayardo, R.J.: Efficiently mining long patterns form databases. In: Haas, L.M., Tiwary, A. (eds.) Proc. of the ACM SIGMOD International Conference on Management of Data, pp. 1–12. ACM Press, Dallas (2000)
3. Song, Y.Q., Zhu, Z.H., Chen, G.: An algorithm and its updating algorithm based on FP-tree for mining maximum frequent itemsets. Journal of Software 14(9), 1586–1592 (2003) (in Chinese with English abstract)
4. Agrawal, R., Shafer, J.C.: Parallel mining of association rules. IEEE Transaction on Knowledge and Data Engineering 8(6), 962–969 (1996)
5. Cheung, D.W., Han, J.W., Ng, W.T., Tu, Y.J.: A fast distributed algorithm for mining association rules. In: Proceedings of IEEE 4th International Conference on Management of Data, Miami Beach, Florida, pp. 31–34 (1996)
6. He, B.: Fast Mining of Global Maximum Frequent Itemsets in Distributed Database. Control and Decision 26(8), 1214–1218 (2011) (in Chinese with English abstract)
7. Li, J., Khuller, A.D.S: On computing compression trees for data collection in wireless sensor networks. In: Proc. of the IEEE INFOCOM 2010, pp. 2115–2123. IEEE Press, Washington (2010)
8. He, B., Yan, H.: Incremental Updating Algorithm of Global Maximum Frequent Itemsets in Distributed Database. Journal of Sichuan University (Engineering Science Edition) 44(3), 112–117 (2012) (in Chinese with English abstract)
9. Park, J.S., Chen, M.S., Yu, P.S.: Efficient parallel data mining for association rules. In: Proceedings of the 4th International Conference on Information and Knowledge Management, Baltimore, Maryland, pp. 31–36 (1995)
10. He, B.: Distributed Algorithm for Mining Association Rules Based on FP-tree. Control and Decision 27(4), 618–622 (2012) (in Chinese)
11. Tao, L.M., Huang, L.P.: Cherry: An Algorithm for Mining Frequent Closed Itemsets without Subset Checking. Journal of Software 19(2), 379–388 (2008) (in Chinese with English abstract)
12. Wang, L.H., Zhao, H.: Algorithms of Mining Global Maximum Frequent Itemsets Based on FP-Tree. Journal of Computer Research and Develpment 44(3), 445–451 (2007) (in Chinese with English abstract)

13. Aggarwal, C., Yu, P.: Outlier detection for high dimensional data. In: Proc. of SIGMOD 2001, pp. 37–47 (2001)
14. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proc. of VLDB 1994, pp. 487–499 (1994)
15. Barnett, V., Lewis, T.: Outliers In Statistical Data. John Wiley, Reading (1994)
16. Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J.: OPTICS-OF: Identifying local outliers. In: Żytkow, J.M., Rauch, J. (eds.) PKDD 1999. LNCS (LNAI), vol. 1704, pp. 262–270. Springer, Heidelberg (1999)
17. Breunig, M., Kriegel, H.-P., Ng, R., Sander, J.: Lof: Identifying density-based local outliers. In: Proc. of SIGMOD 2000, pp. 93–104 (2000)
18. Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc. of KDD 1996, pp. 226–231 (1996)
19. Guha, S., Rastogi, R., Shim, K.: Cure: An efficient clustering algorithm for large databases. In: Proc. of SIGMOD 1998, pp. 73–84 (1998)
20. Hawkins, D.: Identification of Outliers. Chapman and Hall, Reading (1980)
21. Hussain, F., Liu, H., Tan, C.L., Dash, M.: Discretization: An enabling technique. Technical Report TRC6/99, National University of Singapore, School of Computing (1999)
22. Jin, W., Tung, A.K., Han, J.: Mining top-n local outliers in large databases. In: Proc. of KDD 2001, pp. 293–298 (2001)
23. Knorr, E., Ng, R.: Algorithms for mining distance-based outliers in large datasets. In: Proc. of VLDB 1998, pp. 392–403 (1998)
24. Knorr, E., Ng, R.: Finding intensional knowledge of distance-based outliers. In: Proc. of VLDB 1999, pp. 211–222 (1999)
25. Merz, G., Murphy, P.: Uci repository of machine learning databases. Technical Report, University of California, Department of Information and Computer Science (1996), http://www.ics.uci.edu/mlearn/MLRepository.html

Information Technology Education Based on Cloud Computing

Jiangyan Zheng[1], Shijie Cao[2], Jianhui Zheng[3], Zhihui Chen[1], and Liang Hong[1]

[1] Qinggong College, Hebei United University
[2] Tangshan Kaiyuan Welding Automation Technology Institute
[3] Qian'An College, Hebei United University
zjymm2005@126.com

Abstract. Cloud computing era has brought new development ideas for the education information of our country. This paper has summarized and analyzed the definition and computing features of cloud computing combined with the information technology curriculum features. The impact of the cloud computing on information technology education in four aspects has been concluded, and taking Baihui office cloud product as an example, this paper has briefly introduced application and significance of the cloud computing in education of information technology.

Keywords: Cloud computing, the education of information technology, Baihui office, Application.

1 Introduction of Cloud Computing

1.1 Definition

Computing is being transformed to a model consisting of services that are commoditized and delivered in a manner similar to traditional utilities such as water, electricity, gas, and telephony. In such a model, users access services based on their requirements without regard to where the services are hosted or how they are delivered. Several computing paradigms have promised to deliver this utility computing vision and these include cluster computing, Grid computing, and more recently Cloud computing. Cloud computing is a new computing paradigm after the service and grid computing, which are gradually being recognized by the industry. Wikipedia's definition of cloud computing: cloud computing makes IT(Information Technology)-related capabilities available to users by the way of services, the users have access to necessary services through the Internet without the technology, knowledge and equipment operational capability [1-5]. The more general definition of cloud computing [6-8]: cloud computing is a business model. Its computational tasks are distributed in a large number of computer resources pools in which various application systems obtain computing power, storage space and a variety of software services according to different needs.

The basic principles of cloud computing is that the processing data are not stored locally but in the data center on the Internet [9, 10]. Enterprises who offer cloud computing services are responsible for the management and maintenance of the normal operation of these data centers to ensure a strong enough computing power and a large enough storage space for users [11-15]. Users only need to get access these services using any terminal equipment at any time and any place, overlooking the storage or calculations on which "cloud".

In 1969, Leonard Kleinrock [3], one of the chief scientists of the original Advanced Research Projects Agency Network (ARPANET) project which seeded the Internet, said: "As of now, computer networks are still in their infancy, but as they grow up and become sophisticated, we will probably see the spread of 'computer utilities' which, like present electric and telephone utilities, will service individual homes and offices across the country". This vision of the computing utility based on the service provisioning model anticipates the massive transformation of the entire computing industry in the 21st century whereby computing services will be readily available on demand, like other utility services available in today's society. Similarly, computing service users (consumers) need to pay providers only when they access computing services.

The creation of the Internet has marked the foremost milestone towards achieving this grand 21st century vision of 'computer utilities' by forming a worldwide system of computer networks that enables individual computers to communicate with any other computers located elsewhere in the world. This internetworking of standalone computers reveals the promising potential of utilizing seemingly endless amount of distributed computing resources owned by various owners. As such, over the recent years, new computing paradigms (shown in Fig. 1) have been proposed and adopted to edge closer toward achieving this grand vision.
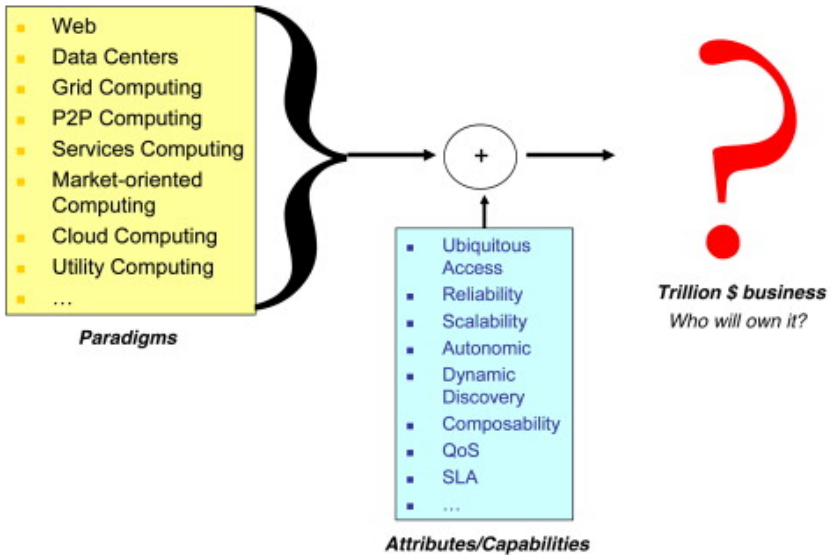

Fig. 1. Various paradigms promising to deliver IT as services

1.2   Characteristics of Cloud Computing

(1) The cloud computing provides users with the most reliable, most secure data storage center, the users need not worry about the loss of data and computer security.

(2) The cloud computing has low requirement on the user side of equipment and it is also very convenient.

(3) The cloud computing can share data between different devices and applications easily.

(4) The cloud computing provides almost infinite number of possibilities for us to use the network.

2   The Impact of Cloud Computing on the IT Curriculum

The subjects of IT have their own characteristics of the teaching methods compared with traditional disciplines, that is asynchronous features of the progress of teaching, teaching content and teaching requirements; students with full interactive features in the learning process [16-18]. The learning method with a typical interaction characteristic and teaching content and the results of the evaluation with the open features [19, 20]. Cloud computing aided instruction has provided appropriate and feasible solution from the characteristics of the IT curriculum as well as hardware and software in IT teaching. The impact of cloud computing on education of the IT is mainly in the following areas [21-24]:

2.1   Study Data in the Cloud, the Safety of the Learning Resources Are Ensured

In previous courses of information technology, students' learning data is stored in its own local computer and a FTP server provided by teacher, the learning data may most likely lose because of unsaved, crashes, viruses and other various factors [25-29]. This kind of situation has been changed after we have the cloud computing which offered students with the most secure and reliable data storage. The students and teachers can store and get their own data anytime, anywhere as long as they get access to the internet without worrying about the loss of data or virus. For example, you can save the document to the 100 office online office platform or save photos to the Baihui file center and so on.

2.2   Information Sharing is Helpful for Group Collaborative Learning

Users can put their own data on the cloud; they can use them anytime, anywhere as long as they have access to the Internet. Failure or collapse of the PC being used, users can continue their work only to change the PC. Users can take advantage of group collaboration, collaborating on the same file application or project real-time with other users around the world. Users can also set access privileges for their data; only people allowed can share the data. In traditional teaching, the students usually

share through FTP, or Network neighbor, prone to data loss with more restrictions and complicated operation. Baihui online office has realized resources sharing among the group with the support of cloud computing and has solved the problem of data sharing easily.

2.3   Promote Exchanges and Discussions among Students with the Record of Learning Activities

Students communicate through the network based on Baihui cloud computing instant chat software, Baihui chat instant messaging, e-mail systems at Baihui cloud mailbox for instant communication, all records and discussions can be saved in the Baihui cloud mailbox, facilitating discussion continuation. Exchange of records and the preservation of the learning data can better reflect the positive contribution of each student in the learning process. In addition, teachers can also observe at any time to understand students' learning, to make timely adjustments and grasp the learning process. Such as Baihui online office documents, forms, presentations, pictures, calendars, and other cloud services closely integrated together, provide a good learning platform for collaborative learning.

2.4   Reduce the Capital Investment, Save on Hardware Facilities

Teachers have to set up FTP server and web server with high-performance computer in the school to make the smooth development of the IT activities, sharing and exchanging learning materials. If you are using cloud computing services, these tasks can be handed over to the cloud, simply visiting the Internet you will be able to obtain these services, without considering upgrades of the hardware and software, schools can also save the capital investment of computer hardware and maintenance costs.

3   The Application of Baihui Office in the IT Education

Cloud environments are built mainly by the current widely used cloud computing platform Baihui (http://www.baihui.com) , as China's leading Software as a Service (SaaS) platform, providing many free cloud services and online applications software, without installing client software to meet the demand for IT courses. For example, Baihui Office which belongs to the SaaS model of cloud computing services, including Baihui write (online document editing), Baihui formatting (online sheets), Baihui Xiu Xiu (online presentation), not only have the basic Microsoft Office functions, but also have the ground-breaking seamless integration of data storage, collaboration, office, and cloud computing ; Baihui file as a multimedia electronic portfolio provides 1 GB of free online storage space; with baihui Wikipedia, you can create a site online which is a collaboration platform to create, edit and share documents and knowledge.

The following will be" learning supplies and health" teaching case, specific analysis based on cloud services at Baihui office in information education application and practice.

3.1   Active Target

(1) Cognitive objectives: Through the activities students can speak common and harmful learning supplies, to learn more about the harmful learning supplies.

(2) Emotional goals: to realize that some learning supplies physical and environmental hazards, students love their awareness of environmental protection.

(3) Behavioral objectives: the ability to use multimedia software todemonstrate the ability of the learning outcomes.

Active skills: the capacity and surveys of students planning activities, analysis of the problem-solving skills, ability to read the writing ability and the basic operation of the   Baihui   office of cloud services.

3.2   The Flow Of Activities

3.2.1   To Determine the Theme

The activity is semi-open theme selection. Prior to the event, teachers and students take baihui talk and baihui mailbox exchange students in the group while browsing side of a discussion of each student can express their views, you can also find relevant web pages link shared with team members, teachers can also be real-time to participate in the discussion of each group of students to discuss the efficiency. Finally, by summing up the outcome of the discussions, and a general understanding and analysis of off-campus curriculum resources, developed the five themes, so that students can make a choice.

3.2.2   Student Grouping, Draw Up a Plan of Action

By the grouping of students to freedom, in order to achieve heterogeneous grouping, teacher for group members to adjust. The team captain to introduce ourselves to produce or recommended by the group members and the group name. After discussion within the group baihui talk everyone should assume the task of determining the activity.

In order to group activities more effectively, the teachers under Table 1: Activity group composition and division of labor; Table 2: " Learning supplies and health" research program, the group leader is responsible for uploading Baihui Docs shared with group members and teachers, the use of the platform group collaboration to complete the form. Finally, the form of writing activities form to record the whole event, to help students experience the process of hardships and the joy of success.

"Learning supplies and health" activities of the group members and the division table

Table 1. Activity group composition and division of labor

Groups	Name	Post

Table 2. " Learning supplies and health" research program

Groups	Group leader	Activity time	Activity content	Activity mode
Ultraman team				
Red scarf team				

3.2.3   To Collect, Collate Information

This stage, the students identified "learning supplies and health" program of activities for the first refinement, the focus is the collaboration between the group learning activities.

In the case of " learning supplies and health"   , introduces the Ultraman team and red scarf two team group.

The Ultraman team focused primarily on the use of harmful learning supplies to students in the survey. Determined the tasks, the group members carried out a survey of our students and issued 100 copies of the student questionnaire, 91 valid questionnaires. Extracurricular time active discussion and active participation in the activities, the camera captured photos of students is the use of hazardous learning supplies, and from the Internet to collect different types of harmful learning supplies and photos.

Red scarf team activities focused on the study of the chemical composition of the harmful learning supplies, as well as the impact on human health. After programming, team member's oral survey school students to use correction fluid and other the harmful learning supplies, the body has no adverse reactions. A detailed analysis of the chemical composition of the the harmful learning supplies, from real life or in the network to collect the instance of several typical hazardous to their learning supplies.

The Baihui Docs Shared Documents feature, group members will collect all the data uploaded to the Baihui Docs, and shared. Each student can be seen in the search for information at the same time the other students of the group has collected information to understand the learning progress of the companion. Between peers can help each other, the ability of students can operate directly in the documentation of the group members, to help the same group of players. In this process, teachers can also use the teachers' accounts, real-time to participate in student learning, teacher-student interaction, student guidance, communicators and collaborators.

3.2.4   The Production of " Learning Supplies and Health" Program of Activities for Presentation

Through the Baihui Show online presentation function, leader of a new presentation, by Baihui Docs shared with members of each team, as long as the document is

opened, the leader coordination to complete a presentation. In the production process at any time can browse the group content, within the group between groups, between teachers and students to discuss the learning process problems, negotiate problem-solving strategies to draw on the advantages of other students play to their strengths, and further optimize their results. Use Baihui Chat within the group exchanges.

3.2.5   Work Exchange, Display

During the process of each group sent a representative to the podium to display the program, finally elected" best program"," the most creative activity plan"," the most cost-effective program" three awards.

The election process is used Internet voting, voting by Baihui Creator in the form of function realization. Teachers in the Baihui Docs set up a voting page, in all of the team after the lecture, all the students to vote, according to the votes to elect design awards, the whole process is well-ordered, no student protest.

3.2.6   The Evaluation

In the evaluation process, the evaluation scale (table four) and uploads to Baihui Docs and sharing to the all group, students completed the phase content can browse the evaluation form, carry out self-assessment and   peer assessment.

Table 3. Activities evaluation scale

Evaluation content	Score			Evaluators		
	3	2	1	Group members	Teacher	Own
Discussions and exchange	Express their views to give personal opinions, listen carefully to the companion to speak, and be able to suggest	Often express their views, listen carefully to the companion to speak	Published some of his own views or listen to other people to speak, but occasionally do not respect companion			

Table 3. (*continued*)

Division of labor	Willing to claim all kinds of tasks completed on time and initiative to help companion	Obey the division of labor within the group, willing to help the group members, but can not be completed on time	Obey to the division of labor within the group, but not completed on time and not good at taking the initiative to cooperation			
Achievements Exhibition	Content-rich, fully explain the problem clearly, but also the feelings of some of the activities	Content integrity to explain clearly the problem in detail	Incomplete, and explained the problem is not clear enough			

4   Teaching Reflection

Baihui office-based cloud services, IT education has achieved certain results, but there are also some shortcomings, The specific circumstances in the following table:

Table 4. Baihui office-based cloud services

Link	Advantage	Shortcoming
Determine the theme	1. One-to-one or group discussion, teachers can participate in student discussion to give guidance, discuss high efficiency. 2. Software can save your chats, easy to record and continuation of the discussion.3. Usually no show of hands of students can also active and teacher exchange.	1. Distracted students in the use of chat software, exchange unrelated to content. 2. Subject to students' typing speed, the lower efficiency of communication than face-to-face communication.

Table 4. (*continued*)

Collect,collate information	1. Data sorting is convenient, efficient, the same group of students can collaboratively edit. 2. Teachers can always browse the student's documentation to identify problems in time to point out.	1.The students operate more unfamiliar 2. The upload speed is slower.
Make a presentation	Clear division of labor, learning efficiency.	Baihui Docs making presentation simple functions, such as not animation effect.
Communication and presentation	Vote for the awards through the network, high efficiency, to ensure the fairness of the results.	Prone to repeat the vote affect the outcome.

References

1. Li, J.: Cloud computing aided instruction. Shanghai Education Press, Shanghai (2010)
2. Yang, Y., Yu, Z.: Cloud computing and cloud computing auxiliary teaching Review of the Development. Software Guide, 216–217 (2011)
3. Kleinrock, L.: A vision for the Internet. ST Journal of Research 2(1), 4–5 (2005)
4. Wikipedia. Cloud computing [EB/OL] (2010-2012), `http://zh.wikipedia.org/zh-cn/theE9%9B%B2%E7%AB%AF%theE9%81%8B%E7%AE%97`
5. Yanyan, Y., Chunguo, H., Chang, Z.: Analysis of cloud computing-based learning. Modern Educational Technology (9), 110 (2010)
6. Chen, Y.: Cloud computing-based collaborative learning in the junior high school ITcourse in practice research. Shanghai Normal University, Shanghai (2011)
7. He, X.: Cloud computing applications in education - Google Sites, for example. Software Guide (Educational Technology) (09), 72 (2009)
8. Liu, F.: Cloud computing-based secondary IT teaching materials development and application. Shanghai Normal University, Shanghai (2010)
9. Kleinrock, L.: A vision for the Internet. ST Journal of Research 2(1), 4–5 (2005)
10. London, S.: INSIDE TRACK: The high-tech rebels. Financial Times (September 06, 2002), `http://search.ft.com/nonFtArticle?id=020906000371` (July 18, 2008)
11. Foster, Kesselman, C. (eds.): The Grid: Blueprint for a Future Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
12. Chetty, M., Buyya, R.: Weaving Computational Grids: How Analogous Are They with Electrical Grids? Computing in Science and Engineering 4(4), 61–71 (2002)
13. Weiss, A.: Computing in the Clouds. Net Worker 11(4), 16–25 (2007)
14. Twenty Experts Define Cloud Computing, `http://cloudcomputing.sys-con.com/read/612375_p.htm` (July 18, 2008)
15. Buyya, R., Abramson, D., Venugopal, S.: The Grid Economy. Proceedings of the IEEE 93(3), 698–714 (2005)

16. Venugopal, S., Chu, X., Buyya, R.: A Negotiation Mechanism for Advance Resource Reservation using the Alternate Offers Protocol. In: Proceedings of the 16th International Workshop on Quality of Service (IWQoS 2008), Twente, The Netherlands (June 2008)
17. Hamilton, D.: Cloud computing seen as next wave for technology investors. Financial Post (June 04, 2008), `http://www.financialpost.com/money/story.html?id=562877` (July 18 , 2008)
18. Morgan Stanley. Technology Trends (June 12, 2008), `http://www.morganstanley.com/institutional/techresearch/pdfs/TechTrends062008.pdf` (July 18, 2008)
19. Keahey, K., Foster, I., Freeman, T., Zhang, X.: Virtual workspaces: Achieving quality of service and quality of life in the Grid. Scientific Programming 13(4), 265–275 (2005)
20. Allcock, B., Bester, J., Bresnahan, J., Chervenak, A.L., Foster, I., Kesselman, C., Meder, S., Nefedova, V., Quesnal, D., Tuecke, S.: Data Management and Transfer in High Performance Computational Grid Environments. Parallel Computing Journal 28(5), 749–771 (2002)
21. Bode, B., Halstead, D.M., Kendall, R., Lei, Z., Hall, W., Jackson, D.: The Portable Batch Scheduler and the Maui Scheduler on Linux Clusters. In: 4th Annual Linux Showcase & Conference, Usenix (2000)
22. Buyya, R., Abramson, D., Giddy, J.: Nimrod/G: An Architecture for a Resource Management and Scheduling System in a Global Computational Grid. In: IEEE Int. Conf. on High Performance Computing in Asia-Pacific Region, HPC ASIA (2000)
23. Buyya, R., Bubendorfer, K.: Market Oriented Grid and Utility Computing. Wiley Press, New York (2008)
24. Chebotko, A., Fei, X., Lin, C., Lu, S., Fotouhi, F.: Storing and Querying Scientific Workflow Provenance Metadata Using an RDBMS. In: Proc. of the 2nd International Workshop on Scientific Workflows and Business Workflow Standards in e-Science, in Conjunction with e-Science, Bangalore, India, pp. 611–618 (December 2007)
25. Chervenak, A., Foster, I., Kesselman, C., Salisbury, C., Tuecke, S.: The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Datasets. Jrnl. of Network and Computer Applications 23, 187–200 (2001)
26. Czajkowski, K., Ferguson, D., Foster, I., Frey, J., Graham, S., Maguire, T., Snelling, D., Tuecke, S.: From Open Grid Services Infrastructure to WS-Resource Framework: Refactoring & Evolution (March 5, 2004)
27. Davidson, S., Boulakia, S., Eyal, A., Ludäscher, B., McPhillips, T., Bowers, S., Anand, M., Freire, J.: Provenance in Scientific Workflow Systems. IEEE Data Eng. Bull. 30(4), 44–50 (2007)
28. Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. In: Sixth Symposium on Operating System Design and Implementation, OSDI 2004 (2004)
29. Dumitrescu, C., Raicu, I., Foster, I.: The Design, Usage, and Performance of GRUBER: A Grid uSLA-based Brokering Infrastructure. Intl. Journal of Grid Computing (2007)

Cloud Computing Security in Multi-processor Virtualization Environment

Chen Xi

Computer Department of Minjiang University, Fuzhou, Fujian, 350000, China

Abstract. With the development of Digital Library under the cloud computing, Virtualization security of the Digital Library of cloud computing is becoming the focus, but it also brings the new security threat. This paper firstly describes the Virtualization security issues of digital libraries under the environment of the virtual cloud computing. Then, it introduces the security and protection method of digital libraries under the environment of the virtual cloud computing, and puts forward security strategy and suggestions starting from the angle of the Virtualization security and the construction of "cloud" digital libraries. The security strategy can improve the overall security of Virtualization systems of digital libraries under the environment of cloud computing.

Keywords: Cloud computing, multiprocessor, Virtualization, computing applications.

1 Introduction

As the future trend of the IT industry, cloud computing is popularizing rapidly in the enterprise level applications [1-5]. It embodies the idea that "the network is the computer". Large scale computing resources, storage resources and software resources are linked together to form a great shared virtual IT resource pool. The highly efficient and flexible scheduling for large-scale resources provides remote user's dynamic expansion and contraction computing services, and also provides a huge space for resource scheduling in cloud service providers, and reduce costs with scale effect [6-9]. Cloud computing is to provide the resources available to users in the form of network services, to liberate the users from the complex maintenance of IT hardware and software, and also, the remote users can get services whenever and wherever possible, breaking the time and geographical constraints [10-13].

Cloud computing security, including identity and access management, data security, privacy protection, virtualization security [14-17]. Figure 1 describes the angle of virtualization cloud computing architectures. Physical hardware and network nodes through a multi-layered physical hardware virtualization logic simplifies the process of forming a flexible computing, storage and network bandwidth 3 by integrating virtual resource pool [18-21]. Provides on-demand and selected resource sharing, distribution, management and control platform, the user can form the upper layer data and the

different needs of the business, with a variety of mutual isolation applications. Way through virtual technology, formed a service-oriented, scalable lT infrastructure can provide cloud computing services. Such as Amazon EC2 (elastic compute cloud), which provides users with a large number of virtual resources, users simply according to their need to create a virtual machine instance. Thus the user through these resources to complete the tasks.

There is a buzz around cloud computing, as users of the cloud services only have to pay for what they use and the resources that they need to cope with demanding situations can be adjusted depending on the demand. This is recognized as the cloud delivery model (SPI – see Figure 1) which consists of three services known as Software-as-a-service (SaaS), Platform-as-a-service (PaaS) and Infrastructure-as-a-service (IaaS). Software-as-a-service allows the users to utilize various applications from the cloud rather than using applications on their own computer. The cloud service provider would usually provide some sort of software development environment to allow applications to be developed for use within the cloud. The application programming interface (API) which the users use to access and interact with the software allows the user to use the software without having to worry about how or where the data is being stored or how much disk space is available as the cloud service provider will manage this for them.
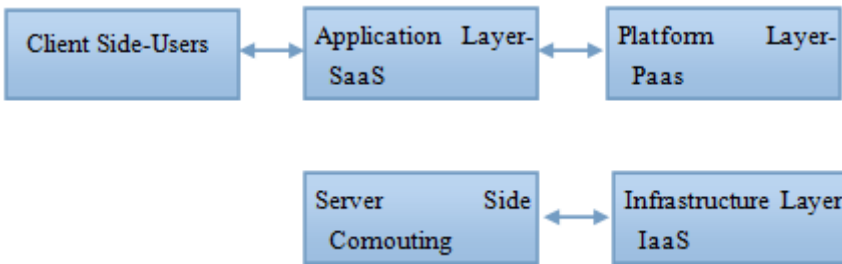

Fig. 1. Showing layers of the cloud delivery model

The current Xen security there is more security issues. For example, Domain 0 is a security bottleneck, its function than other domains strong, so easy to launch a rival worms, viruses, DoS and other attacks, if the adversary Domain 0 paralysis or is compromised, then the destruction of the whole virtual machine system. Xen covert channel problem is not resolved. It is impossible to run on Xen high security level of the operating system [22-25]. Virtual machines share the same set of hardware devices, some network security protocols may be more vulnerable to vandalism and malicious implemented. Xen provides a convenient mechanism for the preservation and restoration. Rollback enables the operating system and reproducing data is very easy, but these will affect the operation itself of the cryptographic properties. In addition, the Xen, because the security mechanisms to make the Guest OS, so can not guarantee the security of VMM. Xen can only restrict the memory page table an I / O address space. The interrupts and I / 0 port address space granularity smaller than the page table. If a different drive in the virtual machine is assigned to the same misfortune page table space, then they can access each other's memory address space, causing safety problems.

For Domain 0 issue. May impair its function. Its functional decomposition to other domains. It would be appropriate to reduce the bottleneck effect Domain 0. Specific strategy needs further study. For sensitive data to be repeatedly erased to prevent further recovery [26-30]. In addition, Xen's ACM module can not completely solve the problem of device isolation and resource isolation, combined with the Xen and LaGrande technology is a good choice. LaGrande is lnfel will be implemented as a hardware technology; it is a common set of hardware security enhancements. Sensitive information is used to prevent malicious software attacks. Its safety function will be integrated into the processor and chip set. Can effectively enhance the device isolation, to achieve I / O protection, cross-border protected memory, keyboard, display isolation protection. In short, after work, need to slowly resolve virtualization security issues, which is the key to cloud computing security.

For traditional security firewall technology can not effectively monitor virtual machine traffic problems, mtor Networks companies use VMware's API to develop virtual Security Analyzer to detect virtual switch traffic in a virtual layer above the network layer traffic. The company has also developed a Virtual Network Firewall, which is based on the Virtual Machine Manager can be certified stateful virtual firewall, check that all data packets through the virtual machine, the organization of all unauthorized connections and allow for a deeper level of data packets check to ensure the security of communications between virtual machines. Security issues for virtual environments, the current Resolution Enterprise Company also proposes to take deep virtualized environment protection strategy. This is a castle-like protection model, through the implementation of appropriate strategies to achieve cloud computing virtual resource pool protection. This new invention is worth further studying.

Virtualization, cloud computing opens the door, and the nature of cloud computing is virtualization services. As a new network computing, cloud computing faces many security problems. The virtualization security as a cloud -specific security issues, needs to focus on. Late research focused on the isolation of the virtual machine, the virtual machine monitor and virtual traffic solid on the trusted platform. Only problem to solve these problems, to be able to ensure that cloud computing platform virtualization security, thus safely use cloud computing.

2 Application Status of the Cloud Computing under the Virtualization Environment

For data security considerations, the cloud computing appears four seed models, i.e. the private cloud, public cloud, community cloud and hybrid cloud. Private cloud is the Virtualization and automation management to the enterprise internal IT resources, and is the integration of internal computing resources. For the internal users, cloud scale is determined by the total internal resources, and all the data are managed by the enterprises themselves. This cloud scale is limited, and can only enjoy part of the advantages of cloud computing. The public cloud is the final design model of cloud computing. The public cloud is also known as the third party cloud. There is greater flexibility and cost advantages. Enterprises can outsource the computation and storage

for cloud service providers, and do not need to buy their own equipment and investment professionals to do cloud system maintenance, but also flexibly add cloud resources rent in the changes of the calculation demand. However, outsourcing the computation and storage to cloud service providers means that the data is outsourced to the third party -- the cloud service providers. In this case the data confidentiality cannot be guaranteed, and this is the biggest difference with the private cloud, and is the reason why now the most cloud computing applications are limited to the private cloud. However, the public cloud is the final model of cloud computing. Community cloud is the semi public cloud formed by the organizations of several common application requirements. It has the larger space to optimize resources than the private cloud, and smaller risk than the public cloud. Hybrid cloud is a mixture of the private and public cloud, namely the users operate the privacy data and business in the private cloud, and operate the non-private data and business in the public cloud. Among these four types of clouds, public cloud is the design intention of cloud computing, and also has the maximum flexibility and cost advantages. The main goal of this paper is to study the security of the cloud computing application in the public cloud.

There are two kinds of user data stored in the cloud servers, the static storage and the dynamic calculation. Static data aims at the storage, which does not need to participate in the operation, but just use the memory function of the cloud, so static data can be protected by cryptography technology. Dynamic data may exist in memory or cache, network and other media, to participate in the operation. Dynamic data need to be reduced to the plain text in the operation, and there is no technology which can make data for computing in the encrypted state, so the private data in the cloud at least part of the time is in the form of plain text. In this section in plain text form the dynamic data may be revealed, because cloud computing service is provided to the users via the Internet, and the cloud service is always online, so the attack from the network may possibly intrude through the network interface of the user virtual machine, break through the operating system of the user virtual machine, and then visit the memory space of the user process.

The existing security mechanisms of cloud computing application are based on the virtual machine technology and implement the protection by adding a layer of software in a virtual machine monitor to calculate the application of cloud, such as the Chaos system which has been realized. However, this cloud computing application protection technology has affected the application of computational intensive efficiency to a certain extent. So the cloud platform needing protection and of very dense cloud computing applications will seriously affect the cloud computing application execution efficiency, and once the single cloud application protection module fails, cloud computing applications will not be protected.

3   The Security Management Strategy of Cloud Computing under the Multi Processor Virtualization Environment

3.1   Improve the Robustness of the Virtualization System

First of all, based on the cloud rental platform security features of the service provider, the security requirements of the cloud library, virtualization application types, safety

requirements and environmental characteristics, under the premise of ensuring the standard of the Virtualization system security and the virtualization application efficiency, the library should jointly formulate the corresponding safety management tactics with the cloud service providers, to achieve fine-grained management of the virtualized resources collection and distribution, and ensure that the virtualized application has a high visibility, correlation, usability and expansibility. Secondly, we should ensure that the virtual machine deployment activities are efficient, fine, controllable and safe, and the Virtualization application process meets the security standards of the cloud computing and the virtual machine application. Third, the cloud computing security monitoring and management platform can implement the large-scale monitoring management, can be able to recognize and predict risk and unknown threats, and ensure the Virtualization application security through the safety alarm, active defense, log review and intelligent prediction mode. Fourth, through concentrating, compressing and filtering the Virtualization application process and the amount of data, reduce the management cost, difficulty and the complexity of the application Virtualization, and realize the centralized control, fine-grained management, fast resource delivery and security policy based access control of the virtual resources and application of digital libraries.

3.2   Embedded Virtualization Security Based on Traditional IT Architecture

First of all, at the beginning of the Virtualization technology selection and Virtualization application construction, we should combine with Virtualization security requirements and cloud security, environmental characteristics, embed the security mechanism within the virtual environment, and ensure the fertilization with a high safety standard in the cloud computing environment. Secondly, in the virtual resource management and virtual machine software running process, on the basis of actual division of the virtual security isolation regions with different grades and safety standards, implement the safety control strategy of virtual resources and data exchange. Thirdly, we must ensure that the Virtualization security management and application have less cloud system resource consumption, and implement the delivery optimization of Virtualization security gains and cloud system resource consumption. Fourthly, in the formulation and implementation process of the virtual security policy, we should reduce the dependence on the physical infrastructure and safety management mode, and continuously improve the visibility and controllability of the virtual management of physical devices, server cluster, virtual pool of resources, virtual network, virtual data center automation configuration.

3.3   Security Protection of the Virtual Network

First of all, we should overall bind multiple virtual network equipment as a logical device for unified management, and according to the application characteristics and safety requirements of the virtual devices, divide different security regions and formulate corresponding security strategies. At the same time, using Virtualization

software protection, virtual firewall and intrusion detection methods, while reducing safety capital investment and the rational use of resources, improve the safety of the virtual network. Secondly, we should automatically and intelligently realize network intrusion prevention and network security vulnerabilities repair, and can predict, judge and prevent unknown viruses using cloud antivirus technology, and encrypt the important data in the network transmission. Thirdly, different users of the cloud library and Virtualization application may use the virtual machine with a physical device, and have a different safety standards and requirements. Therefore, while the traditional monitoring and management equipment and security policy cannot implement management on the data transference of the virtual machine, we shall establish a virtual switch efficiently and set the appropriate access control and data exchange strategy, to implement the network layer data transmission monitoring and traffic vulnerability attack behavior between different virtual machine detection, and use the virtual network firewall to implement the virtual data management and data flow separation.

3.4   Strengthen the Virtual Equipment Management and the Safety of Applications

Secure content of the virtual devices is mainly a physical device's virtual division, the reliability of the operating system, the efficiency of the virtual management program, the excitability of the virtual application process and the security of the relevant data and other aspects. First of all, we should guarantee the evaluation of the overall safety performance of the Virtualization system and the application process and the reliability testing of the operation of the virtual application system, to ensure that the virtual application process is secure, transparent, efficient and flexible. While improving the operational efficiency of a single virtual program, do not reduce the overall system performance for the price. Secondly, based on the safety requirements of the virtual application differentiate different safety classifications. For the Virtualization application related to cloud library management, operation, safety and services, we should create backup virtual machine located in different physical equipment, to ensure that when the host virtual machine fails we can quickly switch, to ensure the continuity of the Virtualization application. Thirdly, according to the virtual application priority and resource consumption, in the premise of reducing the complexity of the Virtualization application management, implement the dynamic allocation and intelligent load migration of the virtual resources, to ensure that the Virtualization application system is safe, efficient, balanced and economic. Fourthly, in order to improve the Virtualization application security, we should back up the virtual server configuration files and the virtual hard disk image. When the virtual application of digital libraries stops working, just restore the backup files to the new server to restore services. In addition, realize data communication encryption of the virtual machine with high security standards.

3.5   Improve the Efficiency and Implementation of the Virtual Safety Management Strategy

Establishing the efficient and executable Virtualization security management strategy is the premise and necessary condition to ensure secure Virtualization cloud library. First of all, based on the virtual system structure and function, we should clarify the security responsibilities of the cloud provider and cloud library and on the basis of cloud computing security standards, test and evaluate the Virtualization security configuration and application. At the same time, classify the data storage and the flow diversion, and strengthen the Virtualization system monitoring, security analysis and identity management of the data access. Secondly, we should strengthen the safety management of the user virtual reading terminal equipment. Through the lower safety investment such as the trust, identity management, log review, security patches, ensure that the virtual desktop has a strong operability and reliability of the data. Thirdly, we should improve the safety, economy, high efficiency and controllability of the Virtualization application management of cloud libraries, Virtualization administrators of cloud libraries create the user to fine-grained access control policies on the virtual system, and monitor, record and report the access to the activities. Fourthly, combined with the virtual safety management strategy, deploy intelligent and automatic system management, supervision and monitoring tools, and by minimizing the host installation reduce the interface number and attack success probability of hacker attacks.

3.6   Construct Open and Reliable Security Defense System

First of all, based on the virtual machine security level and task, for some virtual machines involving the security of the cloud library and the service quality of the cloud reading, through the allocation of independent CPU, hard disks and so on, as well as the virtual and physical isolation layers, protect the safety of the virtual equipment. At the same time, through setting the firewall, the isolation zone (DMZ) running Virtualization security monitoring, vulnerability patch mode in the database and the application layer, eliminate the virtual machine overflows. Secondly, while the cloud library develops and purchases virtual application products, we should strengthen the supervision of the open degree of API (Application Programming Interface), to ensure that the virtual product developed and purchased by the cloud library has high security and controllability. Thirdly, we should regularly comprehensively and systematically assess the robustness of cloud library system architecture and virtual application system, and pertinently reinforce and update the application system security. Fourthly, we must strengthen the deployment of the virtual network firewall and the intrusion detection. The position deployment of the virtual firewall should be especially strengthened between the core switches and the backbone server groups, between the physical network and the virtual network, and between different virtual networks, and impose forced isolation of suspicious or illegal virtual machine flow.

4   Conclusion

For traditional security firewall technology can not effectively monitor virtual machine traffic problems, mtor Networks companies use VMware's API to develop virtual Security Analyzer to detect virtual switch traffic in a virtual layer above the network layer traffic. The company has also developed a Virtual Network Firewall, which is based on the Virtual Machine Manager can be certified stateful virtual firewall, check that all data packets through the virtual machine, the organization of all unauthorized connections and allow for a deeper level of data packets check to ensure the security of communications between virtual machines. Security issues for virtual environments, the current Resolution Enterprise Company also proposes to take deep virtualized environment protection strategy. This is a castle-like protection model, through the implementation of appropriate strategies to achieve cloud computing virtual resource pool protection. This new invention is worth further studying.

Virtualization, cloud computing opens the door, and the nature of cloud computing is virtualization services. As a new network computing, cloud computing faces many security problems. The virtualization security as a cloud -specific security issues, needs to focus on. Late research focused on the isolation of the virtual machine, the virtual machine monitor and virtual traffic solid on the trusted platform. Only problem to solve these problems, to be able to ensure that cloud computing platform virtualization security, thus safely use cloud computing.

This paper presents cloud calculation application protection technology based on multi-processing Virtualization. Transplanting the cloud computing application protection module integrated in the virtual machine monitor and letting it integrate in the Linux kernel and become independent system collaborating with the VMM, to some extent meets the intensive cloud computing platform requirements of the cloud computing application protection. However, because the number of processors of the test machine is limited, it cannot test out the CAPPM number has been growing, and the average efficiency of cloud computing will change.

References

1. Carlin, S., Curran, K.: Cloud Computing Security. International Journal of Ambient Computing and Intelligence (IJACI) 3(1), 21–27 (2011)
2. Nan, K.: Cloud computing platform based business services model. Shanghai International Studies University (2010)
3. Tian, J.: Livelihood Information Resources Integration. Nanchang University (2010)
4. Liu, R.: P2P-based Semantic Web service discovery and composition model study. Nanchang University (2010)
5. He, P.: Cloud computing research and development to achieve. Electronic Science and Technology University (2011)
6. Wang, J.: Based on cloud computing technology cloud services platform architecture Digital Library Research. Jilin University (2011)

7. Xiao, T.: Cloud-based digital library research. Jilin University (2011)
8. Ding, Y.: Cloud-based material prices and labor and employment information platform for research. Wuhan University (2011)
9. Seedlings show: Cloud-based platform for mobile IPTV system design and load balancing technology research. Beijing University of Posts and Telecommunications (2011)
10. Tsang, P.: GridSphere-based cloud management platform implementation and application. Beijing University of Posts and Telecommunications (2011)
11. 10 package put hui; web-based configuration software SaaS architecture research. Hefei University (2011)
12. Tangdi, Lilin: Cloud computing research and information technology application. Office Automation 02 (2011)
13. Wei, X., Chen, K.: Application of GIS Application and Development. Beijing Surveying and Mapping 02 (2011)
14. Li, Z.: Risk control-oriented information resource sharing system performance evaluation coordination mechanism. National Library Science 03 (2010)
15. Surveying; Cloud computing profession of Library and Information Studies. National Library Science 03 (2010)
16. Liu, H.: Domestic mobile Library Research Status and Trends. National Library Science 02 (2012)
17. Li, J.: Cloud computing and its Development. Baoji University (Natural Science) 03 (2010)
18. Song, G., Wei, Z.: Cloud-based Library Construction and Service Development. Library and Information 01 (2011)
19. In particular flowers; cloud computing environment Library Information Resource Sharing Challenges and Countermeasures. Library and Information 04 (2011)
20. Deng, Z., Pan, Y., Li, Z., Luying, J.: Library of cloud services research participant relationship model. Library and Information 03 (2012)
21. Luying, J., Zheng, Y., Deng, Z.: American Library cloud services. Library and Information 03 (2012)
22. Cong, A., Wen, M.: Laboratory information management system (LIMS) development status and trend analysis. In: 2011 Third National Medical Science Research Management Forum and Jiangsu Province Medical Science Research Management Annual Conference Proceedings (2011)
23. Yan, D.: Based virtualization technology to build experimental teaching center environment. Beijing Institute of Higher Education Research Laboratory Symposium 2010 (2010)
24. Yan, D.: Based virtualization technology to build experimental teaching center environment. Beijing College Laboratory Research 2010 Annual Meeting Excellent Papers (2011)
25. Chang, C.-H., Huangfu, Liu, W.: Cloud-based substation maintenance consulting platform. Economic Development Mode Change and Innovation - The Twelfth Annual Meeting of China Association for Science and Technology, vol. IV (2010)
26. Wang: Cloud IT management and library services. North China Association of Universities Figure Twenty-fourth Annual Conference Papers (article) Compilation (2010)
27. Meng, Z.: Cloud computing and its application in the mobile learning mode Exploration. Journal of Computer Research Progress (2010) - Henan Computer Society 2010 Annual Conference Proceedings (2010)
28. China, C., Wu, G.: Cloud-based website security group structure and design and practice. In: 26th National Computer Security Symposium (2011)
29. Jie, Aimin, King, Y.: Smart grid cloud computing technology research. In: China Smart Grid Conference Proceedings (2011)
30. Zhao, B., Xu, G., Liu, X.: Hadoop-based platform software system testing research. In: 17th National Youth Communication Annual Conference Proceedings (2012)

Multimedia Video Information Retrieval Based on MapReduce under Cloud Computing

Zhang Jing-zhai[1], Qiao Xiang-Dong[1], and Zhang Peng-zhou[2]

[1] Xuzhou College of Industrial Technology, Xuzhou, 221140
[2] Communication University of China, Beijing, 100000

Abstract. With the development of network and digital multimedia technology, in the Internet, the application of a variety of videos popularizes increasingly, and the number of online videos explodes rapidly. How to discover and retrieve the vast data of online videos effectively? It has been a problem in emergency to be studied and solved in the field of research and digital media sector. In this paper, from the selection and algorithm analysis of massive data-oriented video features, stream processing can be conducted in multimedia video data by Hadoop's MapReduce distributed computing framework, then users could assign implementation of tasks just by compiling Map function and Reduce function, and the underlying run-time system will conducts parallel computing by scheduling large-scale cluster automatically, so as to retrieve multimedia video information effectively.

Keywords: Cloud Computing, MapReduce, Distributed, Video information retrieval.

1 Introduction

With the rapid development of the Internet and new media technologies, people prefer to use electronic means to store multimedia texts, graphics, images and video data. For example [1-6], New York Stock Exchange generates a daily trading data of 1TB; photos stored in the Facebook reach up to 10 billion, about 1PB data storage; video numbers generated in such video sites as Youku also in the PB level. The relevant statistic pointed out that, the annual global total amount of data has come up to ZB level, which is equivalent to 1 billion TB. With such a vast amount of data, we have walked into the era of big data.

How to find the multimedia video information we're interested in has become an indispensable part in people's lives. However, the existing multimedia video retrieval systems mostly use B / S single-node architecture, by which the retrieval time is not very ideal. Especially, in the cases of multi-user concurrent operation and rapid growth of multimedia data, the real-time of system will drastically reduce, which has failed to meet the demands of people for high-definition video, images and other multimedia information retrieval [7-12]. Therefore, how to timely and efficiently retrieve the video data we need from the vast Internet multimedia resource library is also the current research focus [13-17].

2   Status Analysis of Video Retrieval

Traditional search engines, like Google, Baidu and so on, all conduct retrieval based on text index [18-22]. This method completes indexing and retrieval of video content through the indexing of the surrounding texts of videos, which could solve the basic retrieval needs of users. However, due to the differences between video content and text content, and the semantic gap between video data and its semantic information, the method purely depending on the surrounding texts also exists its own disadvantages. On one hand, it cannot describe video content accurately; on the other hand, it ignores the visual perception of videos and other multimedia features and information. On the basis of text retrieval methods, Internet video retrieval mainly studies the following two problems [23-27]. Firstly, how extract the descriptive texts of videos accurately for the index of texts, so as to improve the quality of retrieval. Secondly, how effectively use visual content of videos to aid the retrieval of videos, including the retrieval and cluster based on content similarity. This two problems, not only influence whether users could effectively browse and retrieve Internet video clips in the vast Internet, but also, for the publishers and regulators of video content, relate to effectively monitoring video content and protecting video content from illegal duplication and dissemination.

The existing Internet video retrieval issues mainly including [28-30]: (1) the process of retrieval depends on network video titles, description text, and other information. Incomplete indexing of video information and inaccurate information of text descriptive will affect retrieved results; (2) the process of retrieval fails to analyze and utilize the visual information of videos adequately, so as to result in the repetition of retrieved results, which has deviation with people's expectations; (3) In the process of handling massive network video data, the traditional serial method cannot meet the practical requirements in performance.

Advantages Of Multimedia Video Retrieval Based On Cloud Computing

2.1   Cloud Computing

In order to solve above-mentioned problems of video retrieval in the era of big data, cloud computing came into being. Cloud computing, a super-computing model, is a resource pool concentrating many computers. It is not only the distributed processing, but also is a whole of intelligent processing functions with self-management and self-coordination on the distributed basis.

Retrieval of video data is concentrated in the "cloud". It packages data processing platforms as a whole, publishes service and shields the details of data processing. What it provides users or the users of the application layer is nothing but a simple access interface. For various data types and the large total amount of video data in the Internet, cloud computing is undoubtedly an appropriate choice, of which the parallel processing features is to meet the demand on the processing time of large data .

2.2   Hadoop Distributed Infrastructure

Hadoop is an open-source implementation of the cloud computing model. Through the distributed processing technology, it makes full use of idle computer resources and

constitutes a Hadoop cluster to improve the utilization of system resources; In the Hadoop cluster, through MapReduce parallel computing framework, it allocates users' demands on multimedia video retrieval to 'idle' nodes for handling, which can effectively solve the real-time problems caused by multi-user concurrent access and retrieval of the target resources from vast database.

In 2004, Google first put forward MapReduce technology. This technology, as a parallel computing model for data analysis and processing, gave rise to extensive attention of industrial and academic circles. At the beginning of design of MapReduce the designers committed to achieve parallel processing of large data through mass cheap server clusters, and gave the priority to scalability and system availability

MapReduce technology framework consists of three levels: (1) distributed file system; (2) parallel programming model; (3) parallel execution engine.

Distributed File System (Google file system) run on basis of the large-scale clusters, and the clusters are built by cheap machines. Data will be stored in the modes of key / value pairs. The entire file system, with modes of centralized management of metadata and distributed storage of data blocks, achieves fault-tolerance in a high degree through replication of data (at least three backups each data). Large data storage (64MB or 128MB each block) will be adopted, so as to compress the data easily, and save storage space and transmission bandwidth.

MapReduce parallel programming model divides the calculation process into two main phases, that is, Map phase and Reduce phase. When Map function handles Key / Value pairs, a series of intermediate Key / Value pairs will be produces. And Reduce function is used to merge all the intermediate key-value pairs with the same Key value to calculate the final results. The specific process of implementation of MapReduce program is shown in Figure 1:
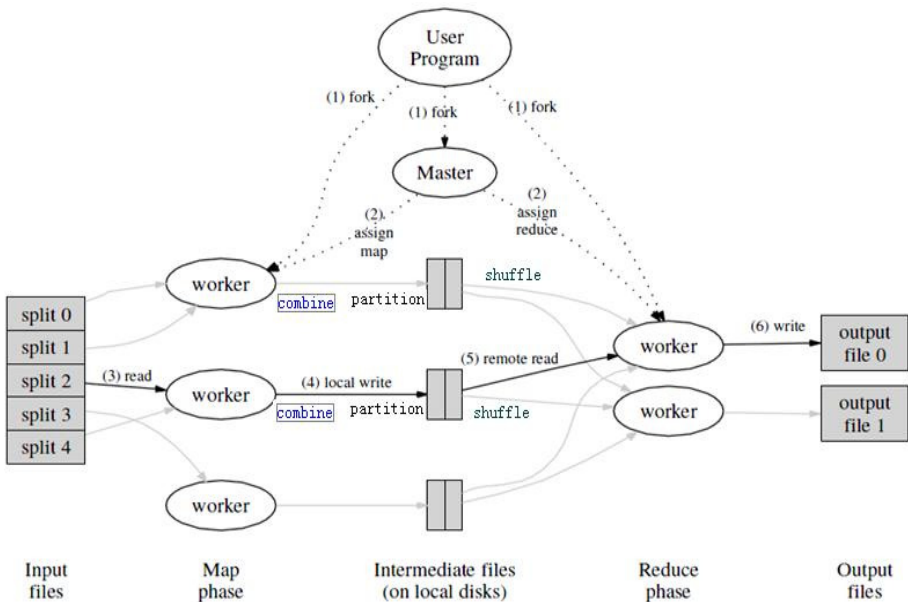

Fig. 1. The Specific Process of Implementation of MapReduce Program

At first, MapReduce block data sources and then hands over multiple Map tasks to perform. Map tasks execute Map functions, classify data according to certain rules, and write data into the local disk. Behind Map phase is Reduce phase. Reduce tasks execute Reduce function, then the intermediate results with the same Key value will be collected together (shuffle) from the nodes multiple Map tasks located in and conducted co-treatment, finally, the output results are written into the local hard disk (distributed File System). The final results of the program can be output and obtained by merging all Reduce tasks.

3 The Retrieval Methods of Massive Videos Based on Mapreduce

3.1 The Selection of Video Features Facing Massive Data

We let the video in the video database be V_i, and let key frame collection be ring $\{ f_i^1, f_i^2, \ldots\ldots f_i^n \}$, then extract feature $s(f_i^n)$ independently from any frame f_i^n. Feature $s(f_i^n)$ is a high-dimensional data vector with such common features as color mean, histogram characteristics and color moment features.

Here, we take the color mean as an example: frame image will be divided into h × l areas (in the calculation, let h = 5 and l = 4, so that the frame image is divided into 20 sub-regions). For the area j, we will respectively calculate the mean of its HSV component $m_j(f_k)=\{ h_j(f_k), s_j(f_k), v_j(f_k)\}$, and combine the mean of all regions into a vector, so the frame image f_k can be expressed as a H * L * 3-dimensional vectors:

$$M(f_k)=\{ m_j(f_k)\}, j=1\ldots h*l \tag{1}$$

The feature vector taking $M(f_k)$ as frame i, for the video v_i containing n-frame, the feature space is represented by its feature vector of all key frames.

$$V_i \rightarrow \{ M(f_1), M(f_2), \ldots\ldots M(f_n)\} \tag{2}$$

Generally, we simplify the feature representation of video v_i on basis of feature set of key frames as:

$$V_i \rightarrow \{ f_i^1, f_i^2, \ldots\ldots f_i^n\} \tag{3}$$

Each key frame is represented by a quantitative high-dimensional feature vector. We call this method of feature representation as the first-order feature representation of videos. Correspondingly, if each key frame isn't just represented by one vector but a set (k) of high-dimensional vectors, we call this method of representation as the k-order feature representation of videos. If color moments are used as the features of frame images, each frame image is made up of 3 sets of three-dimensional vectors, that is, three-order feature representation. If SIFT are used as feature description, frame images

are made up of 7 SIFT feature points, so we call this method as k-order feature representation (k is an arbitrary value).

Through hash quantitative methods, we index feature vectors, and choose LSH as the method of hash indexing.

We index all the feature vectors of video v_i through LSH, and then combine index keywords of hash bucket projected by feature vectors, so as to obtain LSH visual term collection $W(V_i)$ of v_i and use $W(V_i)$ to characterize v_i. This is called as 'package of video terms' method, and the feature vectors representation of videos are converted into document vector space representation of videos.

Computing model of text similarity is further adopted to compute the similarity of videos. Through this method, the performance costs caused by the method of feature vectors will be reduced greatly.

3.2   Distributed Computing Model of Video similarity on Basis of MapReduce

We mainly adopt distributed computing method to improve the execution efficiency of the algorithm and enhance the speed of computing the video similarity, so as to make the algorithm be applied to the vast video database.

Distributed processing method splits the massive calculation into independent tasks, which will be calculated by multiple computers respectively, and then merges the results of each block to get the final result. The process framework of the entire system is shown in Figure 2, which contains five major steps:

At first, we extract the features of videos in video database D to get the feature vector representation of videos, and then store feature vectors in the corresponding hash buckets through LSH index. Each hash bucket can be regarded as an inverted list of feature vectors, and hash index value can be index terms (visual terms) of inverted lists. The specific process is shown as algorithm 1:

Algorithm 1. LSH Index on Videos

LSH(D)
 //D: the video database
 //output: the LSH indexes of all video features in D
 for each video v_i in D:
      for each frame feature vector f in v_i
          LSH(f)

For the inverted lists of feature vectors, we first conduct Reduce TF (Algorithm 2) to carry on the statistics of the inverted lists, so as to get visual term frequency of each visual term in the videos.

Algorithm 2. Reduce TF Conducting Statistical Analysis of Visual Term Frequency on Visual Terms

Reduce_TF (Index key, Iterator values)

//key: the visual word

//values: a list of feature vectors according to the visual

//word

for each frame feature vector f in values:

    if f$\in v_i$,

      then tf(v_i)+=1

After visual terms via Reduce TF, the intermediate result generated is the visual term frequency of index visual terms in each video. And then Map Correlation (Algorithm 3) is conducted in the intermediate result and records the same visual term's frequency in different videos, so as to obtain the correlation of the visual term between videos. Specific process of algorithm is as follow:

Algorithm 3. Map Correlation Computing Correlation through Visual term Frequency

Map Correlation (Index key,   Iterator values)

//key: the visual word

//values: the term frequency of all videos according to

//the visual word

for each term frequency pair<tf(vi),   tf(vj)>in values:

  set correlation(vi,vj)=tf(vi)* tf(vj).

Then, according to the correlation between videos, the same set of videos will be shuffled in the fixed reducer and merged. The process of Reduce Correlation is shown in algorithm 4:

Algorithm   4. Reduce Correlation Merging Correlation

Reduce Correlation (Index key, Iterator values)

//key: the video pair(vi, vj)

//values: the correlation values of video pair(vi, vj)

sim (vi, vj)=0

for each value V in values:

set total_corr(vi,vj)+=V.

The final result: All video <vi,vj> relevant visual dictionary W is merged into video database.

3.3   Experiment and Analysis

We use MapReduce computing model mentioned in this section to compute similarity. While in distributed computing framework, MapReduce environment contains three PC, of which one is used as NameNode and JobTracker (Master) under the Hadoop environment, and another two is worked as DataNode and TaskTracker host (Slave).The dataset adopted in the experiment is MUSCLE VCD dataset. Under such an environment, the total time used by the calculation of video similarity is 1.4h, while the time used by serial method is three times more than parallel method.

In order to verify the influence made in the efficiency of similarity calculation by the numbers of DataNode / TaskTracker hosts, in this paper, we also adopt the above-mentioned collection of video clips as dataset for the experiment, choose 1,2,3,4 DataNode / TaskTracker hosts respectively and run the computing program of similarity under the Hadoop system. Experimental results show that, as the umbers of Slave hosts increase, the time used in the calculation of similarity will decrease gradually. Therefore, facing the massive video data, we could increase the numbers of Slave hosts to enhance the efficiency of calculation, so that the system has a better scalability.

4   Conclusions

This paper mainly studies, in the era of big data, under cloud computing environment, the retrieval methods of multimedia video information based on MapReduce. Its core problem is the calculation of similarity between video clips. For rapid calculation of the similarity between videos, the paper refers to the vector space model of text retrieval, and proposes that LSH can be used to calculate vector space model of videos. The experiment results show that, with Hadoop distributed platform as basis, and MapReduce parallel computing model as the method of handling massive video data, this platform can be built in a large amount of cheap normal computers, so as to achieve the efficient retrieval of multimedia video data.

With digital technology and the rapid development of the Internet, especially with the development of Web2.0, the amount of data on the internet rapid growth of the Internet has also led to the relative lack of data processing capabilities. As more and more data to be processed, and more difficult to one or a limited number of storage servers to accommodate and more not by a one or a limited number of compute servers can handle such volumes of data. Cloud computing is an Internet-based network application platform, cloud computing, "cloud" is that there is a cluster of servers on the Internet resources on , which includes hardware resources (servers, storage , CPU, etc.) and software resources (applications, integrated development environment, etc.) , the local computer only needs to send a demand information through the Internet , there are thousands of remote computers will provide you with the necessary resources and returns the results to the local computer , so that almost no local computer what to do , all processing in the cloud computing providers of computer cluster to complete. Cloud computing proposed by Google , followed by the Internet community wind "cloud" Chung , the ensuing cloud computing services and success stories emerging technology

platforms , such as Google's GFS, MapReduce, Bigtable, Chubby and AppEngine, Amazon's Dynamo, EC2 , S3, SQS, SimpleDB and CloudFront, Microsoft's Azure, SQL, ". Net" and Live services , open source cloud computing platform HDFS, HBase and Eucalyptus, VMware virtualization platforms. Cloud computing as an emerging computing model is changing the way the world works .

Cloud computing is a parallel computing, distributed computing and grid computing evolved , MapReduce is one of the core technology of cloud computing , which is a parallel data processing system provides a simple , elegant solution , and its main purpose is to large clusters of systems on large data sets in parallel , and the data for large-scale parallel computing . First introduced MapReduce cloud computing and related knowledge and research on the current status of MapReduce and abroad were introduced and assessment, and summarizes the current research issues related MapReduce model , and finally summarize and looks to the future development trend.

MapReduce together as a new programming model for cloud computing , large-scale data -intensive applications, such as web search, scientific computing and data mining , has been widely used. MapReduce library provides designers with a transparent hardware development environment , making it easy to store and use parallel computing architecture to simplify the previous operational difficulties of the underlying issues such as MapReduce-based system itself has a separate storage , high scalability and good fault-tolerant processing mechanism and so on. Although MapReduce itself has more advantages, but because the mechanism does not mature at this stage , resource scheduling has been a limiting its efficiency in the implementation of the biggest problems . Equipment for heterogeneous environments and applications , the use of open source implementation Hadoop , MapReduce paper focuses irrational mechanisms exist and belongs resource scheduling issues and put forward corresponding improvement ideas . The main contents are as follows: (a) in a homogenous environment with MapReduce original resource scheduling method is proposed based on the proportion of dynamic resource allocation algorithm (DPRS), dynamic monitoring node's load circumstances, a reasonable allocation of tasks corresponding resources, improve the original mechanism deal with heterogeneous environments Reduce task scheduling problem of imbalance . (2) In order to localize the input data as much as possible the implementation , using the optimal allocation model of local capacity (LCPO), eliminating the need for the original backup redundancy caused by overhead reduces network traffic flow , improve the efficiency of the implementation of the Map task . (3) In order to improve heterogeneous environment Reduce backward node's backup execution efficiency and address the mechanism behind the original node miscarriage of justice issue, a fast algorithm for long backup tasks (FLTB) o (4) The data for dividing the input file blocks are distributed to the various tasks arising Map imbalance , pretreatment method to achieve this heuristic data partitioning , to some extent, improve the original data in a heterogeneous environment, task allocation imbalance .

References

1. Du, X., Wang, J., Jiang, C.: Heterogeneous environment backbone relaxation labeling method for task scheduling. AAS, 06 (2007)
2. Zhu, Y., Gu, X.: Double auction-based grid resource scheduling model and Bidding Strategy. Library and Information Technology, 12 (2008)
3. Gang: Heterogeneous computing grid system communications and network topology discovery mechanism. Computer Learning, 01 (2004)
4. Liu, L., Luo, D.: Utilize SQL SERVER Data Import and Export. Changsha Aeronautical Vocational and Technical College, 01 (2004)
5. Shang, M.: Heterogeneous bus network can be divided into load optimization scheduling algorithm. Computer Engineering, 20 (2005)
6. Zhou, L., Lu, J.: Web services in a distributed component integration application. Microcomputer Information, 17 (2005)
7. Chang, Z., Zhang, Y.: Heterogeneous environment adaptability perceived real-time collaborative study. Computer Age, 04 (2008)
8. Asia, L., Ecliptic, Hai, S.: Heterogeneous Sensor Networks Energy Efficient Routing Algorithm. Computer Science, 05 (2008)
9. Wu, X.-C., Chao, Y., Guo, Y., Huang, X., Wang, Q., Huang: Heterogeneous environment-oriented teaching team building in GIS. Surveying and Mapping, 09 (2008)
10. Von Tin, W., Wang, Y.-Q., Xue, F., Zhang, W., Lu, H.: Heterogeneous data exchange technology research and implementation of the initiative to find. Liaoning Shihua University, 02 (2009)
11. Tao, T.: Cloud computing environment MapReduce-based resource scheduling model and algorithm. Dalian Maritime University (2012)
12. Chao, C.: Emergency Resource Scheduling Model and Algorithm. Nanjing University of Aeronautics and Astronautics (2010)
13. Xin, W.: Cloud computing's MapReduce parallel programming model. Henan Polytechnic University (2010)
14. Jianping: Cloud computing model based on MapReduce cluster scheduling optimization and research. Nanjing University of Posts and Telecommunications (2013)
15. Li, W.: Heterogeneous data integration model of CAPP Research and implementation. Dalian University of Technology (2005)
16. Yang, C.: Enterprise data access platform for research and design. Electronic Science and Technology (2005)
17. Zhu, S.: MapReduce-based scientific computing application performance analysis and optimization. Fudan University (2010)
18. Zhuohui: Heterogeneous environments in real-time collaborative Research and Implementation of Adaptive Perception. Zhejiang University (2007)
19. Zhang, H.: Rsync heterogeneous environments based data synchronization mechanism. Electronic Science and Technology University (2009)
20. Hong, L.: Dynamic load balancing based on feedback scheduling algorithm in heterogeneous environments Hadoop Design and Implementation. Nanjing University (2012)
21. Cong, H., Wen-sheng, Xie: Agricultural Information Grid resource scheduling data transmission method. In: Chinese Agricultural Information Technology Innovation and Development Conference Proceedings Disciplines (2007)
22. Song, N., Zhao, Z., Dai, Y., Bo: Trust based on a dual-level feedback Grid Resource Scheduling. In: 16th National Youth Communication Conference Proceedings (2011)

23. Dong defense, division Enpei: Grid based on trust mechanism for resource scheduling. In: Tenth National Enterprise Information and Industrial Engineering Conference Proceedings (2006)
24. Rui, Kun, W., Wei, Z., Wang, Weiping: MapReduce framework based on an approximate copy text detection. In: NDBC 2010 27th China Conference Proceedings Database, B Series (2010)
25. Wei, Z., Chen, Li, L.: MapReduce cloud computing model based collision detection algorithm. In: 2010 System Simulation Technology and Application Conference Proceedings (2010)
26. Sun, G.-I., Xiao, F., Xi, X.: MapReduce model for scheduling and fault tolerance mechanism. In: 2007 National Open Distributed and Parallel Computer Conference Proceedings (2007)
27. Zhao, K., Linkui, Yang, D., Yang, J.: Environmental monitoring network for heterogeneous components management system. In: 2010 China Environmental Science Society Annual Meeting Proceedings (2010)
28. Zheng, Q., Fang, M., Wang, S., Wu, X., Wang, H.: MapReduce model based on parallel scientific computing. In: 2009 National Open Distributed and Parallel Computer Conference Proceedings (2009)
29. Zheng, Q., Wang, H., Wu, X., Fang, M.: HPMR: multi-core high-performance computing cluster support platform. In: 2008 National Open Distributed and Parallel Computer Conference Proceedings (2008)
30. Zhang, J., Zhang, Y.: Hierarchical heterogeneous components of the control system design and interaction semantic description. In: Twenty-sixth Chinese Control Conference (2007)

Study of Cloud Computing in Cellphone Terminal

Zhao Xi

Information Management Dept, Shanghai Finance University,
Shanghai, 201209, China

Abstract. With the continuous progress and development of our economy and society, computer science and technology has also achieved substantial progress. In recent years, cloud computing has been extensively applied in the IT industry and witnesses an increasing range of application in cellphone terminal. By virtue of the application of cloud computing in cellphone terminal, functions of cellphone can satisfy human being's demands, thus thoroughly meeting our demands of cellphone at all levels. Based on some related experience of my own, some opinions about the application of cloud computing in cellphone terminal will be discussed to facilitate mutual exchange between us, thus being conducive to all of us.

Keywords: cloud computing, cellphone terminal, application study.

1   Introduction

Cloud computing is a kind of technology which doesn't come into being until recent years [1-4]. Its emergence has caused huge influence on the development of IT industry. As our economy and society develops, cellphone becomes one of the important parts in human being's life [5-10]. Since then, cloud computing has been applied in cellphone. The core idea of cloud computing is to utilize resources connected via the internet to conduct unified management and scheduling, thus constructing a resource library and better serving its users [11-14]. As the application of cloud computing in cellphone terminal brings better development, cloud computing will become an inevitable choice of various cellphone manufacturers.

2   The Connotation and Advantages of Cloud Computing in Cellphone

As a mode of network service based on shared architecture, cloud computing provides users with corresponding services through network resource pool [15-20]. Cloud computing is not a special technology, but a computing philosophy or mode. Under the mode of cloud computing, application programs users use are in the network resource sharing cluster, rather than manufacturers of cellphone terminal device. The application of cloud computing allows smoother operation of users' cellphone terminal [21-25].

The application of cloud computing in cellphone terminal has many advantages, which can be currently reflected in several basic aspects. First of all, cloud computing enjoys extremely superior advantages in computing and storing capacity. Server equipped with cloud computing has extremely strong expansibility, which makes it possible to bring highly expansive storage space and extremely strong computing capacity to users; secondly, the cost performance of cloud computing is relatively high [26-30]. The application of cloud computing enjoys a high cost performance, which can be attributed to its extremely low requirements toward hardware device and high level of resource utilization, which greatly enhances the cost performance of cloud computing. Thirdly, cloud computing has such features as safe, stable, convenient and flexible storage. The server-side of cloud computing enables storage to be more safe and reliable, thus free users from any concerns that their storage might be attacked by viruses and relevant data damage followed by that, which can make users feel more relieved in the process of being served. In addition, cloud computing allows platforms to be customized correspondingly in accordance with users' services, which allows more flexibility and convenience to cloud computing and makes it possible for cloud computing to better serve the extensive range of users. Finally, cloud computing is more convenient in the aspect of data sharing. By applying cloud computing, users can synchronize data in different devices. Cloud computing also offers much convenience to data sharing between different users, thus allowing the application to be smoother than before. As shown in figure 1.
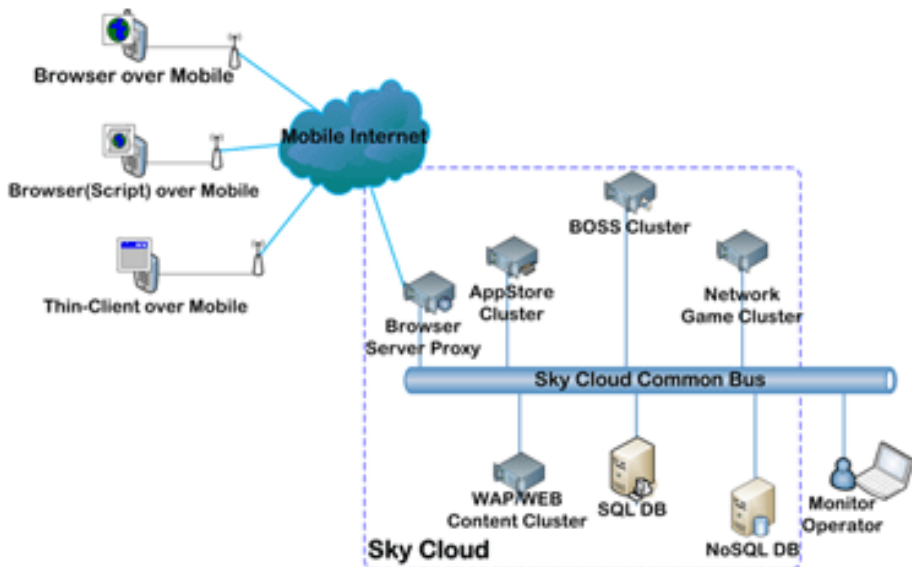

Fig. 1. Cloud Computing model in Cellphon

3   The Extensive Application of Cloud Computing

Cloud computing has been applied in a highly extensive range, and its advantages can be reflected in many fields in better serving the huge group of users. The application of cloud computing can be mainly reflected in several basic aspects as follows.

3.1   The Application of Cloud Computing in Software as a Service (SaaS)

Cloud computing can offer software service to the huge group of users via the internet explorer. In other words, it is unnecessary for users to buy software, service providers complete the purchasing work of software and upload them to the internet explorer, so that all the users can have access to software. In this way, not only can users utilize software more conveniently and rapidly with the help of cloud computer, but also they can greatly reduce the expense on purchasing software. Cloud computing can bring so many advantages to its vast users. One primary example of the application of cloud computing in software as a service is CRM in Salesforce.Com, which has already brought substantial benefits to its users. [2]

3.2   The Application of Cloud Computing in Utility Computing

In the aspect of utility computing, cloud computing has also been applied in an extensive way. By establishing a virtual data center, cloud computing endows this center with a more expansive storage capacity and stronger computing capacity and centralizes various kinds of device, thus making it possible for this virtual data center to better serve its vast users and thoroughly utilize its resources.

3.3   The Application of Cloud Computing in Web Platform

The extensive range of application of cloud computing also spreads to web platform. Through Application Program Interface (API), web service providers dispatch relevant developers to design software, thus shaking off the responsibility of providing software with comprehensive functions. In this way, web service providers can obtain more economic interest from this process. Moreover, with the help of cloud computing, the range of their services becomes more extensive. GoogleMaps is a very typical example of this application. [3]

3.4   The Application of Cloud Computing in Platform as a Service

As another form of Software as a Service (SaaS), the application of cloud computing in platform as a service can better serve the vast users. By receiving services of environment development provided by service providers, users can further develop on this basis and pass on the already developed software to other users via the internet.

3.5   The Application of Cloud Computing in Managing Service Supply

Cloud computing has been applied in the field of managing service supply as soon as it emerges. This kind of cloud computing application is not targeted to the vast users, but an application field serving IT managers. By virtue of the application of cloud computing in managing service supply, relevant emails can be scanned and software can be monitored, thus enabling these fields to serve the vast users more smoothly, efficiently and safely.

3.6   The Application of Cloud Computing in Commercial Service Platform

The application of cloud computing also covers commercial service platform, but it is not realized merely by cloud computing. As a matter of fact, it depends on a perfect combination of cloud computing, Software as a Service (SaaS) and Management Service Provider (MSP), thus providing its users with a stable and interactive service platform, in order to offer corresponding assistance to users with commercial activities on the internet. It has also been extensively applied in our daily trades. In our daily life, we are served by many similar commercial service platforms, which guarantee smooth progress of our trades. Alibaba is the most typical example of the application of cloud computing in commercial service platform.

4   The Application of Cloud Computing in Cellphone Terminal

4.1   The Application of Cloud Storage in Cellphone Terminal

Cloud storage means to integrate application software in different types of storage device on the internet through some relevant technologies, thus setting up a system which can better serve the vast users. As to users, device in cloud storage is completely transparent, which allows them to better utilize relevant resources and receive superior services. The cloud computing function of cellphone mainly involves two aspects of content as follows: one is in the software market of cellphone. Cellphone users can download application software through corresponding ports, thus making the download service of cellphone application software more convenient and rapid, which further offers much convenience to the vast cellphone users; the other lies in the client-side of cellphone information. Thanks to cloud storage, this information window won't occupy too much space in the cellphone, which can save huge storage space for users. [4]

4.2   The Application of Cloud Security in Cellphone Terminal

As an important branch of cloud computing technology, cloud security has been extensively applied and generated very obvious effect in the field of anti-virus. Cloud security technology can monitor abnormal behaviors of users' software, which prevents application software from being attacked by some viruses or Trojans and avoids any

damage to application software. Cloud security pushes the latest information about relevant Trojans and viruses to the server for further analysis and processing, and then sends back solutions to each client-side accordingly, which allows perfect processing of Trojans and viruses and protects users from being affected by them.

4.2.1   Questions Confronting Cloud Security System

The establishment of a scientific and perfect cloud security system makes it possible for cloud security to better serve the vast users. However, there are still some relevant questions that need to be handled to enable a smoother operation of cloud security system. At present, questions need to be solved can be categorized into several aspects as below. Firstly, massive cloud security probe needs to be solved, because it is the root of perfect processing of relevant viruses and Trojans. Only by possessing massive cloud security probe can viruses and Trojans on the internet be responded to in the shortest period of time, thus processing relevant viruses and Trojans in the first time; secondly, professional anti-virus technologies and experience must be achieved, because it is the root guaranteeing functions of cloud security system. With rich experience and professional anti-virus technologies, which can keep pace and be well integrated with the latest technologies in time, cloud security system can be equipped with powerful functions and cloud security technologies can be given full play; thirdly, ample financial and technological resources must be invested. Cloud security technology is expected to keep pace with the continuous progress of development of our science and technology to meet the requirements of scientific and technological development. The application of new technologies in cloud security must be guaranteed by ample funds, which can allow better development for cloud security technology; finally, open system is needed. Cloud security system needs to be operated together by more partners, thus enabling cloud security system to be more stable and enjoy more powerful functions, which can better serve the vast users.

4.2.2   The Application of Cloud Security in Cellphone

As science and technology continuously develop, cloud security technology enjoys a wide range of application in cellphone and plays a vital role in the normal function of cellphone. In recent years, as android system has been introduced to our life, cellphone viruses are in an increasing tendency. Cellphone viruses can lead to many adverse influences to our cellphones and get in the way of normal function of cellphone, and that's why we must adopt reasonable technological means to prevent and eliminate cellphone viruses to guarantee normal operation of cellphone. Currently, cloud security technology witnesses an increasing application in cellphone, which can be reflected in only two aspects though. The first one is the application of cloud security technology in remote control, which is primarily applied in cellphone security by positioning lost cellphone and encrypting important data in cellphone through certain technologies, thus preventing users from adverse influences caused by losing important cellphone data. Cloud security technology also allows transfer of some important data from terminal to terminal, which brings much convenience to users; the second one is data backup function. Data backup is one of the main reflections of the application of cloud security technology in cellphone. Cloud computing in cellphone allows relevant data in

users' cellphone to be uploaded to the server equipped with cloud computing through personal computer, thus realizing data backup. Users can have access to backup data at any time by inputting accounts. In this way, even if data in the cellphone has been damaged, users can download corresponding data from the backup data, thus recovering users' cellphone to the original mode.

5   Case

5.1   BlackBerry Enterprise Application Server Solutions

As a research and development leader in enterprise application solutions for many of Canada's RIM's enterprise users with a set of BlackBerry Enterprise application server solutions, this solution can be said to be a characteristic of cloud computing with mobile Internet applications.

In this solution, the BlackBerry Enterprise mail server enterprise applications, wireless communications network operators around the world, and BlackBerry mobile terminals connected together, users do not need to own a mobile terminal has a huge storage space, through the application of push (Push) technology BlackBerry terminal remote access server to access their e-mail accounts, so you can easily synchronize remote mail and calendar, view attachments and address this Blackberry operating model illustration.

Although sales of RIM BlackBerry devices, and also authorize other mobile device platforms (eg iPhone, Android and Windows Phone platform) access BlackBerry servers, enjoy Blackberry services. Currently, the BlackBerry wireless platform is extended through its own applications, such as online CRM, etc.. In the cloud computing model to the user's application to become a core RIM business model, and achieved great success.

5.2   Apple's Siri Voice Assistant Service

Apple is today the world's largest mobile phone maker, is also the largest PC maker, its latest core technology products Sift With iPhone4S should be shipped out. When the user to speak against the iPhone, the voice will immediately encoding contains all useful information into a digital file. This information will be sent via Internet to Apple's cloud computing servers, and server module will identify the user saying.

Meanwhile, the user's voice will also be identified in the mobile terminal, when the cloud computing server understands the request fit the user's instructions are processed locally (such as play some local music, find phone contacts, open an application, etc.), then cloud computing server will inform the mobile client self- execution. If the user's instruction can not be done locally (such as Find local weather for the latest news, etc.), cloud computing servers will replace the phone for information retrieval and the result is fed back to the mobile client.

As part of the speech recognition and speech synthesis relatively simple and can be done locally, but in the Siri semantic understanding of this technique requires a lot of

knowledge, which is calculated using the juice of cloud powerful processing capabilities, instead of handling the phone semantic recognition, knowledge computing (information retrieval) that require high-performance computing, only the calculation results back to the mobile client presented to the user (Figure 3). This is what gives mobile users a powerful user experience.

6   Conclusion

In conclusion, the application of cloud computing in cellphone terminal is an inevitable trend of the continuous progress and development of science and technology. However, several questions still exist in the application of cloud computing in cellphone terminal. We must handle these questions in a scientific and reasonable way, thus facilitating better application of cloud computing in cellphone terminal, and in turn, better function of cellphone for the vast users.

As a pioneer in cloud computing, Google the company actively developed for the Android platform mobile environment and terminals, has introduced mobile devices and cloud computing based on a new application, which in addition to the common integrated mobile search and voice search service, but also includes : fixed search, Google mobile maps and Google Street on Android.

Location Search and Google Maps for mobile services, through the base station location, Wi-Fi positioning and GPS positioning 3 ways to identify the user 's location information and in accordance with change of place to provide the appropriate search results. Achieve precise positioning via GPS, which provides driving directions, etc. services. And in the Google cloud computing servers, "cloud" is mainly responsible for the user submits a search to calculate car navigation route calculation, a large number of map data storage, and mobile client will only receive the results of these calculations and map data and presents to the user. And Googl ~ street applications, cloud computing server storage space with every single point coordinates and image data, and the use of single-point -based spatial hemisphere of spatial mapping images to reconstruct the real image data for each angle, finally sent to the mobile terminal 's just an image data.

As China's telecom industry giant China Telecom, but also when the World Expo launched based on cloud computing technology and 3G mobile Internet cloud computing concept mobile terminal "e cloud phone", the main provider of public services and the Expo mobile Internet applications. Achieved through the phone's built-client mobile Internet to provide users with "Expo Services cloud", "Tianyi application cloud," the two applications, user- Expo Garden exhibition concept with the use of mobile Internet applications, and Internet access via mobile cloud computing platform to achieve the item content, applications and venue information automatically added and updated.

By using the Android operating system that supports multi-touch, in addition to built-in support Tianyi Star, Tianyi Express, surfing video, such as China Telecom Tianyi space characteristics of the application, but also built a "Tianyi application cloud" and other special services, the user can always experience characteristics of the "

cloud treasure box ", "e cloud storage " and other cloud-based technology to provide value-added services.

Cloud Box is a cloud-based client, so that the low-end mobile phone via the Internet through Tianyi cloud to achieve certain value-added services. It elevates the customer service area, through the deployment in the cloud mobile search, mobile reading, mobile communities, mobile games, mobile securities market, mobile storage and other value-added services, making low-end high-end mobile phone has enjoyed some business functions.

e cloud storage is a cloud-based application, it will extend the phone's storage space to the cloud, via mobile phone able photography, video, office documents, contacts and other data transmitted in real time to be saved to the cloud, enabling large amplitude is reduced footprint mobile client and the client can always be synchronized and Pc, so that data backup becomes faster and easier.

Now mobile cloud computing has been achieved in the mobile terminal such as the enterprise application solutions, voice assistant, mobile maps and Tianyi cloud applications and other services, in fact, cloud computing applications on a mobile phone is not only just that, I'll pick a few simple an idea about the mobile cloud computing.

The traditional arcade games for mobile phones, all of the user 's role of action instruction to go through two-way transmission to the other side of the mobile terminal, through a series of calculations will be fed back to the other phone, this is bringing huge traffic costs and terminal computing capabilities. The mobile cloud computing applications on the mobile phone network, the user need only unilaterally instruction unidirectional transmission to the cloud, the cloud holds all the data the user role, through the comparison of both the strength of battle after battle results will be returned their phones, such as a mobile client need only show the game interface and thus the cost of traffic and terminal computing capacity requirements will be greatly reduced.

The traditional mobile advertising is advertising information uniformly distributed to all advertising accept client, regardless of whether the user is interested in the advertising rush. However, mobile advertising and cloud computing integration through the cloud complex algorithms and stored in the cloud mass user information for different times, different places, different sex and age to distinguish between professional and other information, to achieve a targeted and practical ad sent.

Although mobile cloud computing in the development of mobile terminal is still a lot of space, but the application of these phones is just the tip of the iceberg mobile cloud computing, far from achieving it should produce results. Now in the mobile terminal, including MP3, PSP, NDS and other mobile entertainment devices and cloud computing if you can combine together, then we will work to provide more convenient life.

References

1. Xuan, Z., Pu, H.: OMS platform cloud-based mobile service test system. In: China Institute of Communications Sixth Annual Meeting Proceedings (on) (2009)
2. Zhao, N., Wei, F., Shi, B.L., Wang, S., Wang, J.-J.: Cloud-based digital home audio and video service platform. In: 2009 National Conference Proceedings Virtual Instruments (two) (2009)

3. Fen, Fen, H.: Mobile phone terminals introduction and development of meteorological services Thinking. In: Chinese Meteorological Society 2005 Annual Conference Proceedings (2005)
4. Dan, D., Tang, J.: Mobile Device Security Research. In: 2005 China Institute of Communications Wireless and Mobile Communications Commission Annual Conference Proceedings (2005)
5. Fei, Z., Wang, Q.: Cloud computing network operation and maintenance management model. In: Guangdong Communications 2010 Outstanding Youth Forum Proceedings (2010)
6. He, X., Jie: Radio spectrum monitoring and management system for cloud computing trend. In: Spectrum Management and Monitoring System Conference Proceedings (2011)
7. Yu, C.Q.: Talking about cloud computing technology in customer service System. In: 2011 Wireless and Mobile Communications National Academic Conference Proceedings (2011)
8. Xu, Y.: IOT cloud computing platform. In: 2011 Annual Communications and Information Technologies and New Advances - Eighth China Institute of Communications Annual Conference Proceedings (2011)
9. Suhan, C., Li, H., Takasugi, Liu, Q.: PTLR: cloud computing platform to handle large data confidence moving domain logic regression algorithm. NDBC 2010 27th China Conference Proceedings Database (B Series) (2010)
10. Belts, Shao, W.: "Cloud computing" and the production and broadcasting of multimedia integrated service platform. In: Chinese News Technical Workers Union 2011 Annual Meeting Proceedings (Part II) 2011
11. Lei, X.: Based on the BREW platform for WCDMA mobile terminal software design, Northwestern Polytechnical University (2006)
12. Chiu, D.: Cloud-based telecommunication network relationships analysis technology research and application. Beijing University of Posts and Telecommunications (2012)
13. Yang, S.: Based on complex network analysis of large-scale telecom data. Beijing University of Posts and Telecommunications (2010)
14. Fei, L.F.: Cloud computing environment can prove that the data held by technology research. Shanghai Jiaotong University (2011)
15. Wei, Z.: Cloud computing user data transmission and storage security scheme. Beijing University of Posts and Telecommunications (2011)
16. Song: Cloud computing in the telecommunications business platform application. Beijing University of Posts and Telecommunications (2011)
17. Xiaoyang: Mobile cloud security authentication key technology research. Henan University (2012)
18. Channing: Cloud computing platform based telecom business support system scheduling technology research. Nanjing University of Posts and Telecommunications (2011)
19. Lei, G.Y.: Cloud-based GPS navigation terminals Research and Design. Yanbian University (2011)
20. Ning: Cloud computing platform based telecom business support system scheduling algorithm research. Nanjing University of Posts and Telecommunications (2012)
21. Yan, K., Aijun, Y.: Mobile Phone mobile communication terminal market analysis and performance testing. Information and Communication, 03 (2011)
22. Chen, Y., Dong, Z.C.: Talking Business Support System of cloud computing and SOA. Telecommunications Technology, 08 (2011)
23. Xiaohui: Mobile Internet Overview. Information and Computer (Theoretical Edition), 06 (2011)

24. Sheng Tong, D.B., Jun, S.F.: Triple play under a cloud-based real-time transcoding technology research and application. TV Engineering, 02 (2011)
25. Zhang, S.-B.: Cloud computing in the exploration and use of radio and television in Xiamen. TV Engineering, 02 (2011)
26. Tang, X.-H., Ji, J., Wu, H.: On IOT Technology and Application. Guangxi Communication Technology, 02 (2011)
27. Wang, W.: Cloud computing and broadcasting video content operators. Audiovisual Sector (Radio and Television Technology), 04 (2011)
28. Zheng, W., Feng, L.: Cloud computing applications in the IDC research and practice. Telecommunications Technology, 07 (2011)
29. Wang, X., Liu, J., Sun, H.C., Zhang, H.: Telecom IT support system integration strategy research. Telecom Express, 07 (2011)
30. Xing, L., Hui, S., Hu, B.: HU Bo; C-RAN for mobile communications opportunities and challenges. Telecommunications Network Technology, 07 (2011)

Evolution Analysis of Online Knowledge Transfer Network

Jing Wei[1,*] , Ruixiao Song[2], Jianjun Miao[2], and Zan Xu[2]

[1] College of Economic and Management,
Nanjing University of Post and Telecommunication, Nanjing 210046, China
[2] College of Economic and Management,
Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
`mirror820909@163.com`

Abstract. Based on the study of online knowledge transfer network's topology and dynamic statistical characteristics, this paper found the characteristic differences between the evolution of online knowledge transfer network and the traditional BA network model. Through the empirical data analysis of a BBS forum, it gives out an evolution model of the online knowledge transfer network based on the PageRank algorithm. Meanwhile, in the network analysis it also found that the network growth model has a power-law distribution of degree. Through the control of the attenuation coefficient and the node degree growth factor, it can ultimately realize growth control of the online knowledge transfer network.

Keywords: knowledge management, knowledge transfer, complex network, BBS, PageRank algorithm.

1   Introduction

According to the degree of power law distribution characteristics, Barabasi and Alert BA proposed the classical BA network model [1]. But BA model which use degree as the deciding factor of connection probability can't describe the actual network connectivity exact. Many scholars expand the BA model. LiJi (2006) took Moore's law as node growth rule to put forward the node number accelerated growth network model [2]. Aimed at weighting network, Su Kai (2009) bring forward the multi-infrastructures variable weighting network model [3]. Consider the node growth situation in a period of time, QinSen (2007) gave out the network evolution model with edge growth [4].

In the reference of formal BA model study, this paper improves BA model again and proposes the online knowledge transfer network evolution model based on Page Rank algorithm.

This paper is organized as follows. In section 2 we will introduce the variation trend of the online knowledge transfer network. In section 3 we will show out the

---

[*] Corresponding author.

evolution pattern and statistics quality followed by the effect deduction in section 4. The conclusions are given in section 5.

2   Variation Trend of Online Knowledge Transfer Network

The organization knowledge forming process is actually based on social network, and the social network is eventually formed by interaction of individual behavior（Bai Shuying，2003）[5] . Internet presence sets up individuals' interaction method. The knowledge transfer process based on social network takes network interactive as carrier to show out online characteristic. Therefore, this paper takes BBS forum which is the representative of online knowledge transfer network as the discuss object.

2.1   Network Building

BBS forum is a certain kind of virtual community whose full name is Bulletin Board System. Precisely speaking, BBS is a discussion system with more participants, and composed by many bulletins. The topic in every bulletin is called post. Individuals build the connected relationship by post reply to form the network structure. Along with the time, the network relationship becomes more and more complicated through posting and forum building.

The Online knowledge transfer network takes the posting and reply individuals as nodes and takes replay way as edge. Different posters and different repliers form a complex network. Every post which forms an isolate sub network crossover linked each other through the nodes and facilitates the knowledge circulation.

Suppose that BBS network is composed by N nodes and K edges, so the weighting network can be denote as $G = (N, \Phi, W)$ .The node set is $N = \{n_1, n_2, ... n_N\}$, the edge set is $\Phi = \{l_1, l_2, ... l_K\}$,and the weighting set is $W = \{w_1, w_2, ... w_K\}$ .If node i and node j are not connected $w_{ij} = 0$ .If node i and node j are connected, w_{ij} is the edge weighing which points from i to j.

2.2   Evolution Trend

Zhang Zhongzhi (2006) had said that the evolution process of network were all constitute by the events of add nodes, add edges, remove nodes or remove edges. Every local variation of network is all complicated by the above four events. [6]Online knowledge transfer network in the course of evolution usually shows out the following characteristics（Zhang Li，2008） [7]: the joint of new node and the birth of new edge.

We take a sub bulletin in a certain BBS employee forum as the data accumulate object. Through data analysis we get the degree distribution curve (as shown in figure 1), which matches power law distribution and power law index is $\gamma = 1.7897 \pm 0.7736$. The numerical result is consistent with the literature [8] demonstration which the BBS degree distribution is1.6~1.9. But in BA network the

degree distribution index is $\gamma = 3$, which deviates from the empirical data. It means that online knowledge transfer network topology is quite different from the BA scale-free network. In the BA model, preferred choice means that the elder node is much more attractive. But in online knowledge transfer network, the state is not so. In figure 2, the attractive node appeared in every stage. The reason is that just wonderful posts attract much more replies but not elder posts.
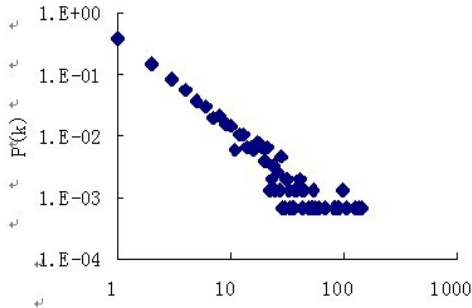

Fig. 1. The degree distribution of online transfer network
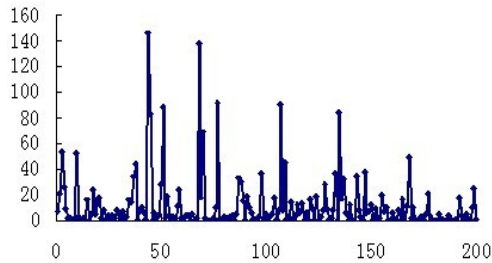
Fig. 2. The degree of online transfer network' nodes

3   Evolution Pattern and Statistics Quality of Online Knowledge Transfer Network

Through the statistical characteristic analysis of online knowledge transfer network growth, we can see that BA model isn't a good simulation for online knowledge transfer network evolution. This paper does some alternations to the BA model, and puts forward a new suitable algorithm for online knowledge transfer network evolution process.

3.1   BA Model

BA model summarizes the reason of power law distribution as the increasing nodes and priority connection properties. Barabasi and Alert deduced the index of power law distribution as $r = 0.3$ [9] theoretically. But Dorogovtsev (2000) found that most

the reality of complex networks' r scope is [2, 3][10]. Therefore, BA model's edge connection mode still needs further improvement, in order to well simulate the real network.

3.2   Constructing Algorithm Based on Page Rank

Many scholars have ever improved the BA model's above shortcomings. This paper references Zhang Liang (2008)'s [11] research conclusions to improve the BA model again. Page Rank algorithm was first applied to Internet search engine to rank search results which brought out by Lawrence Page, Serge Brin（1999）[12]:

$$PR(A) = (1-d) + d\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)} \tag{1}$$

$PR(A)$ is A's Page Rank value. $PR(T_i)$ is T_i's Page Rank value, and T_i is the web which links to A. $C(T_i)$ is the out degree of T_i. d is attenuation coefficient which can be adjusted, and $d \in (0,1)$.

Besides degree, Page Rank algorithm also takes other impact factors into consideration. The more times the web is linked to, the more important the web is. In another aspect, maybe a web has been repeatedly cited, but compared to the out degree, the numerical is not important.

Through the former analysis, we know the growth of online knowledge transfer network is realized by new nodes joint and new postings. Page Rank algorism can be used to balance the abnormal increase of the nodes' out degree. So we can avoid the BA model's shortcomings which just consider the node judge priority in the process of growth.

In addition, Zhang Liang (2008) was not considered the node's degree increase speed when improved BA model. But in fact, the increase of the node's degree is not linear, and we have to describe it. In the improve process, we add node degree increase coefficient α, which can also offset Page Rank algorithm's congenital deficiency.

Considering the above factors, we form the online knowledge transfer network:

(1) Network Initialization:

Set the initial number of network nodes as m_0. The network attenuation parameter is d （$0 \le d \prec 1$）and every nodes' initial Page Rank value is $1-d$.

(2)Joint new nodes:

We put l nodes into the network, whose Page Rank value is $1-d$. The new nodes joint with the old nodes by $m(m \le m_0)$ edges. The probability π_i of chosen the old nodes is not only relate to the degree of i, but also related to the degree increase coefficient $\alpha \in (0,1)$.

$$\pi_i = \frac{PR(i) + \alpha}{\sum_j PR(j)} \tag{2}$$

(3)Updated the node's Page Rank value:

$$\begin{pmatrix} PR(1) \\ \vdots \\ PR(l) \end{pmatrix} = (1-d) \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + d \frac{A_{il}}{m} \begin{pmatrix} PR(1) \\ \vdots \\ PR(l) \end{pmatrix} \tag{3}$$

A_{i*l} is the network's link matrix.

(4)Repeat nodes join until meet the requirements.

3.3   Network Statistics Quality Based on Page Rank

The exits of node degree increase coefficient α are used to reduce the Excessive attention caused by Page Rank algorism. As in figure 3 to figure 5, when $d = 0.9$,different α causes different degree distribution. Along with the increase of α , the nodes with low degree increase which lead to polarization mode. Through the improving online knowledge transfer network evolution algorithm, we get the following degree distribution images in figure 5. According to the definition of Page Rank, d is attenuation coefficient which effects the growth choose probability π_i .

When d approaches 0, the network doesn't show out power law distribution (as shown in fiugre 6). It means the connection of new nodes and form nodes are random, and not under the control of preferred condition. Along with the network growth, the old posts will sink gradually and the probability which every node be pointed to will likely decrease. So even in the state of small d , and no preferentialness probability under the network growth state, the network also can appear "fat tails" phenomenon.
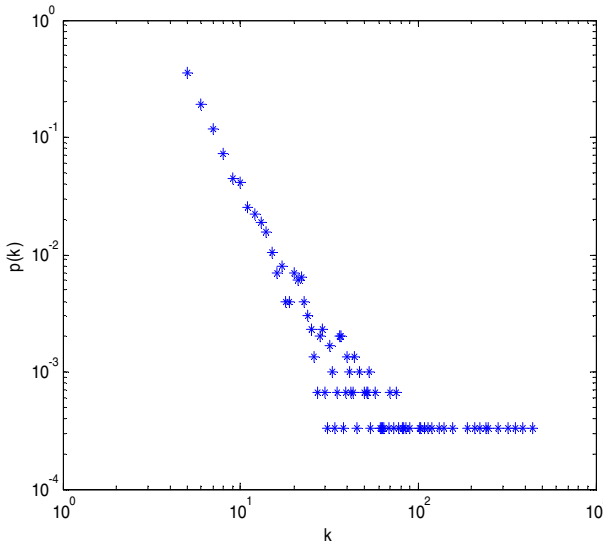
Fig. 3. The affection of node degree increase coefficient $\alpha = 0.1$ to degree distribution
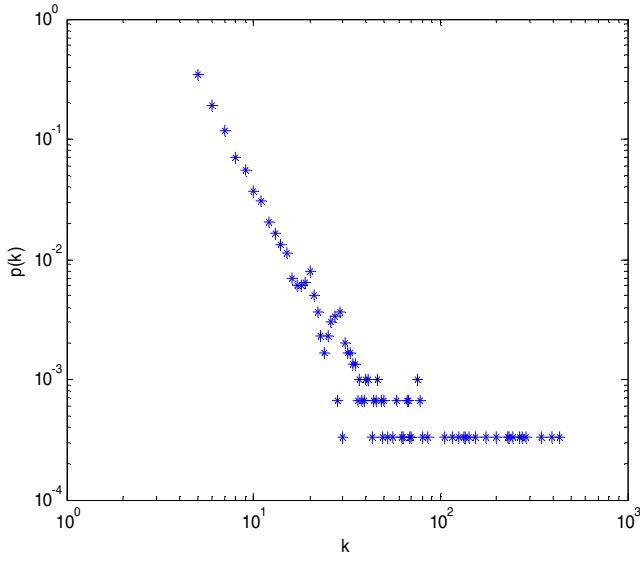
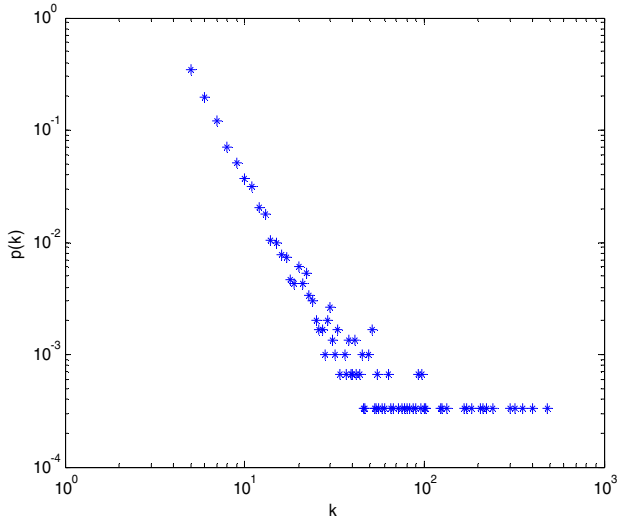**Fig. 4.** The affection of node degree increase coefficient $\alpha = 0.5$ to degree distribution

Fig. 5. The affection of node degree increase coefficient $\alpha = 0.8$ to degree distribution
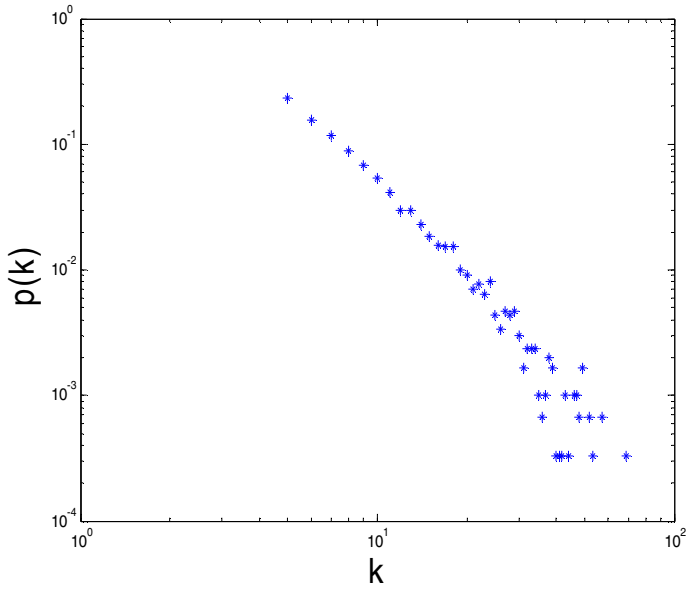
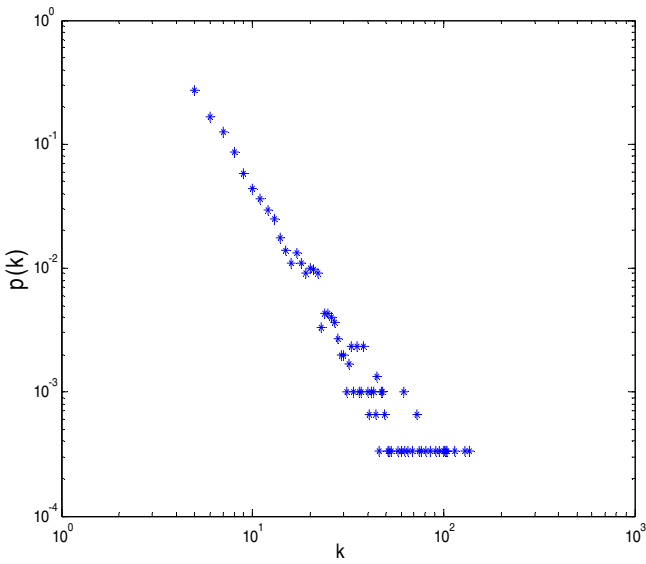**Fig. 6.** The network growth degree distribution curve based on Page Rank algorithm, d=0.1

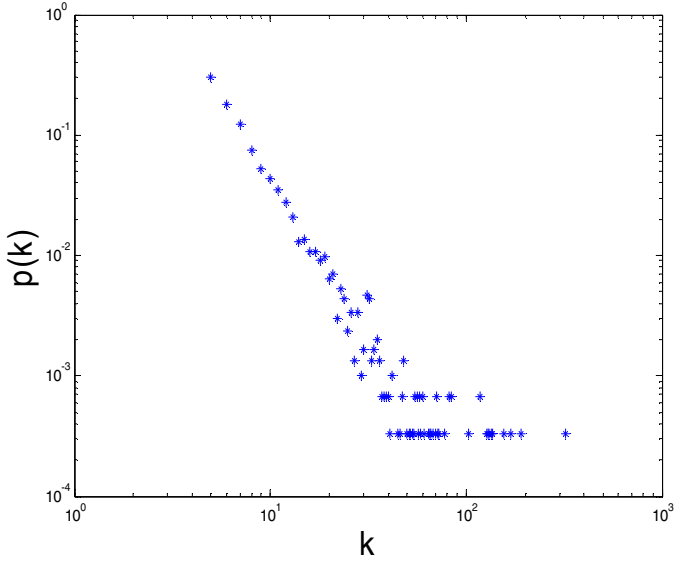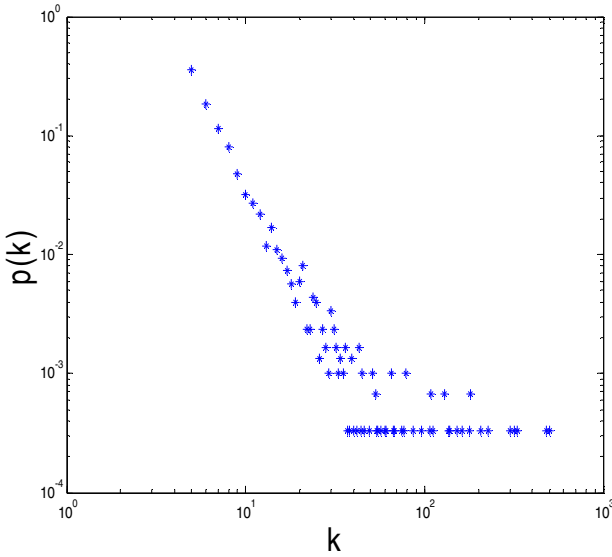**Fig. 7.** The network growth degree distribution curve based on Page Rank algorithm, d=0.4

Fig. 8. The network growth degree distribution curve based on Page Rank algorithm, d=0.6

Fig. 9. The network growth degree distribution curve based on Page Rank algorithm, d=0.9

Along with the increase of attenuation coefficient d , the network distribution degrees' "fat tails" began to grow fat as shown in figure 6 to figure 9, which means that the number of nodes with low degree increases all the time. The most

magnanimous degree distribution is getting higher, and the differentiation between the nodes' degree distribution appeared gradually more obvious. Meanwhile the power rate of degree distribution increases, namely the tilt degree of degree distribution image increases.

4   Effect Deduction of Online Knowledge Transfer Network

We suppose that there are m_0 nodes in the online knowledge transfer network, using Page Rank algorithm to do evolution. According to the algorithm, we add l nodes to the network every stage. The new node and the former node will lead to m edges with the probability of $\pi_i = \dfrac{PR(i)+\alpha}{\sum_j PR(j)}$.

Using the PR value calculation formula, we get:

$$PR(i) = 1 + d + \frac{d}{m}k \tag{4}$$

$$\sum_j PR(j) = (1+d)t + \frac{d}{m}mt = 1 + 2dt \tag{5}$$

d is the attenuation coefficient of Page Rank algorithm, and k is the degree of i. So, the degree added rate of online knowledge transfer network is:

$$\frac{\partial K}{\partial t} = m\frac{PR(i)+\alpha}{\sum_j PR(j)} = m\frac{1+d+\dfrac{d}{m}k+\alpha}{1+2dt} = \frac{m(1+d+\alpha)+dk}{1+2dt} \tag{6}$$

The network degree increase speed has the positive relationship with node degree increase coefficient α, which agrees with figure 4.

To solve the differential equation:

$$K(t) = -\frac{m(1+d+\alpha)}{d} + \frac{m(1+d+\alpha)}{d}(\frac{t}{t_i})^{2d^2} \tag{7}$$

Assuming the time interval is equal, time distribution density is $P(t_i) = 1/t$ .So the probability of which the node degree less than k is:

$$P(K \prec k) = P(t_i \succ [\frac{\dfrac{m(1+d+\alpha)}{d}}{k+\dfrac{m(1+d+\alpha)}{d}}]^{\frac{1}{2d^2}} t) = 1-[\frac{\dfrac{m(1+d+\alpha)}{d}}{k+\dfrac{m(1+d+\alpha)}{d}}]^{\frac{1}{2d^2}} \tag{8}$$

Derive equation (8), we get degree distribution probability function $P(K)$:

$$P(K) = P(K(t)) = \frac{\partial P(K \prec k)}{\partial k} = \frac{1}{2d^2} \frac{(\frac{m(1+d+\alpha)}{d})^{\frac{1}{2d^2}}}{(k + \frac{m(1+d+\alpha)}{d})^{1+\frac{1}{2d^2}}} \tag{9}$$

Predictably, the node degree distribution meets power rate distribution $P(K) \sim k^{-r}$,and $r = 1 + \frac{1}{2d^2}$. The scope of d is $(0,1)$,so the result of $\gamma \in [1.5, +\infty)$ is correspond with this paper's empirical result.

5   Conclusion

This paper researches the online knowledge transfer network, and analyses the differences between actual online knowledge transfer network and BA scale-free network. On the base of BA model, this paper imports Page Rank algorithm to construct a new growth model. We expanded the BA model to form a network growth model which corresponded with the characteristics of online knowledge transfer network. In the analysis of network growth rate, the network growth model shows out power law characteristic. Through the adjustment of the node degree increase coefficient α and attenuation coefficient d, we can realize the control of network connection strength and create the nodes of high degree. Therefore through the algorithm control, we can control the network growth, and realize the control of network's information delivery direction and strength.

References

1. Barabasi, A.: Emergence of scaling in random networks. Science 286, 509–512 (1999)
2. Ji, L.: Growing complex network model with accelerating increasing number of nodes. Acta Physica Sinica 55(8), 4051–4057 (2006)
3. Su, K., Wang, L.: Flexible Weighted Complex Network Evolving Model and Simulation. Journal of System Simulation 22(1), 266–271 (2009)
4. Qin, S., Dai, G.: Dgree Distribution of Evolution Networks with the Acceleration of Edge Attachment. Systems Engineering Theory & Practice 11, 159–163 (2007)
5. Bai, S., He, M.: The Interactive Structure and Process of BBS. Sociological Research 5, 8–18 (2003)
6. Zhang, Z.: Evolving Models of Complex Networks. Dalian University of Technology (2006)

7. Li, Z.: Research on the Evolution Process of Virtual Community Networks. Journal of Physics 57(9), 5419–5424 (2008)
8. Lind, P.G., Gallas, J.A.C., Herrmann, H.J.: Coherence in scale-free networks of chaotic maps. Physics Review E 70(5), 20–24 (2004)
9. Barabasi, A., Jeong, H.: Mean-field theory of scale-free random networks. Physical A 272, 173–187 (1999)
10. Dorogovtsev, S.N., Mendes, J.F.F., Samukhin, A.N.: Structure of growing networks with preferential linking. Physical Review Letters 85, 4633 (2000)
11. Liang, Z.: The Growing Models of Complex Networks and the Partitional Method for Community Structure. Dalian University of Technology (2008)
12. Page, L., Brin, S.: The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab (1999)

An Improved Symbol Detection Algorithm
Based on Expectation-Maximum

Ge Wang, Hong-Yi Yu, and Zhi-Xiang Shen

Institute of Information Science and Technology, Zheng Zhou, China
`suhperwangge@sohu.com`

Abstract .n this paper, the problem of symbol detection of linearly modulated signals in the presence of unknown carrier phase is studied. We proposed a signal detection algorithm based on EM, which is different from conventional solution that not need estimation carrier phase，the symbol information can be obtained directly, thus reducing the performance loss caused by hierarchical processing of the traditional receiver. And resolved the problems of SER degradation that caused by phase ambiguity and the integral interval. The deterioration of the symbol detection performance was improved by sub-regional integration and adjusts the integration interval. Simulation results show that the symbol detection performance improved significantly.

Keywords: Expectation, Maximization Algorithm, Symbol Detection, Phase Ambiguity, Integration Interval, Ser.

1 Introduction

In packet-based communications frames arrive at the receiver with an unknown carrier phase. For the problem of symbol detection of MPSK or QAM signals with initial unknown phase offset, conventional solution is estimating the phase offset first and then detecting symbols. Another method is taking symbol detection and phase offset estimation jointly with maximum likelihood algorithm, such as maximum likelihood sequence detection [1-2].

Expectation-Maximization (EM) algorithm is a broadly applicable approach to the iterative computation of maximum likelihood (ML) estimates, useful in variety incomplete-data problems, and has been proposed in. References proposed a general formulation of iterative ML estimation of unknown parameters in the presence of nuisance parameters by means of the EM algorithm. The channel gain, the symbol timing, the carrier frequency offset and the phase offset has proposed in. The maximum-likelihood sequence detection and estimation algorithm is considered in general framework based on maximum-likelihood detection/estimation theory a joint sequence detection and phase estimation algorithm based on decision-directed [3-4].

When phase estimation is performed by means of conventional non-data-aided (NDA) algorithm the resulting estimate exhibits a phase ambiguity, due to the

rotational symmetries of the signaling constellation. Phase ambiguity resolution can be accomplished by a data-aide (DA) algorithm that exploits the presence of a known pilot sequence in the transmitted data stream and by differential encoding. In an iterative ML algorithm for joint phase estimation and phase ambiguity resolution for turbo-coded system is reported [5].

In the present paper we will focus on the problem of symbol detection of linearly modulated signals in the presence of unknown carrier phase. We proposed a signal detection algorithm based on EM, which is different from conventional solution that not need estimation carrier phase. And resolved the problems of SER degradation that caused by phase ambiguity and the integral interval. The rest of the paper is organized as follows; Section 2 introduced the signal detection algorithm based on EM. The influence of the phase ambiguity and the integration interval to signal detection is developed in Section 3 [6]. In Section 4, simulation results and analysis are given. Finally, conclusions are drawn in Section 5.

2   The Proposed Algorithm

2.1   Signal Model

We consider the transmission of data symbols over an AWGN channel. It is well known that matched filter followed by symbol rate sampling can produce sufficient statistics for data detection in the AWGN channel. The phase offset between the transmitter and receiver clocks is unknown and thus it has to be estimated by the receiver [7]. It is assumed that the fixed block of L data symbols are transmitted over the channel and then subsequently processed by the receiver. The received signal model can be described as

$$y_k = c_k e^{j\theta} + n_k, \quad k = 0, 1, 2, ..., N-1 \tag{1}$$

Where y_k is sampled signal after the matched filter, θ is an unknown phase offset and is a zero-mean white Gaussian noise sample with variance σ^2. The likelihood function for data symbols and phase offset can be written as?

$$p(y_k \mid c_k, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left|y_k - c_k e^{j\theta}\right|^2}{2\sigma^2}\right) \tag{2}$$

Let $y = [y_0, y_1, y_2, ..., y_{N-1}]^T$ is the received signal vector, $c = [c_0, c_1, c_2, ..., c_{N-1}]^T$ is the transmitted PSK or QAM signal vector. The conditional probability density of y, c and θ is:

$$p(y \mid c, \theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^N \exp\left(-\frac{\left| y - ce^{j\theta} \right|^2}{2\sigma^2} \right)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^N \exp\left(-\frac{\left(y - ce^{j\theta} \right)^H \left(y - ce^{j\theta} \right)}{2\sigma^2} \right) \tag{3}$$

2.2   The Proposed Algorithm

Let $x = \{ y, \theta \}$ be the complete data, using EM algorithm to realize symbol detection.

A) E (Expectation) Step

$$Q(c, \hat{c}^i) = E_\theta \left[\ln p(y, \theta \mid c) \mid y, \hat{c}^i \right] = E_\theta \left\{ \ln \left[p(y \mid \theta, c) p(\theta \mid c) \right] \mid y, \hat{c}^i \right\}$$

$$= E_\theta \left[\ln p(y \mid \theta, c) \mid y, \hat{c}^i \right] + E_\theta \left[\ln p(\theta \mid c) \mid y, \hat{c}^i \right] \tag{4}$$

Since the phase vector θ is independent of the transmitted signal $\mathbf{c}$, there will be

$$Q(c, \hat{c}^i) = E_\theta \left[\ln p(y \mid \theta, c) \mid y, \hat{c}^i \right] + E_\theta \left[\ln p(\theta) \mid y, \hat{c}^i \right] \tag{5}$$

Removing the items unrelated with c^i,

$$Q(c, \hat{c}^i) = E_\theta \left[\ln p(y \mid \theta, c) \mid y, \hat{c}^i \right] = \int_\theta \left[\ln p(y \mid \theta, c) \right] p(\theta \mid y, \hat{c}^i) d\theta \tag{6}$$

With the Bayed formula and total probability formula

$$p(\theta \mid y, \hat{c}^i) = \frac{p(y \mid \theta, \hat{c}^i) p(\theta \mid \hat{c}^i)}{\int_{\theta'} p(y \mid \theta', \hat{c}^i) p(\theta' \mid \hat{c}^i) d\theta'} \tag{7}$$

Since the phase vector θ is independent of the transmitted signal $\mathbf{c}$,

$$p(\theta \mid y, \hat{c}^i) = \frac{p(y \mid \theta, \hat{c}^i)}{\int_{\theta'} p(y \mid \theta', \hat{c}^i) d\theta'} \tag{8}$$

With (6) and (8), and removing the items unrelated with $\mathbf{c}^i$, therefore,

$$Q(c, \hat{c}^i) = \int_\theta \left[\ln p(y \mid \theta, c) \right] p(y \mid \theta, \hat{c}^i) d\theta \tag{9}$$

B) M (Maximum) Step

$$\hat{c}^i = \arg\max_c Q(c, \hat{c}^i) = \arg\max_c \int_\theta \left[\ln p(y \mid \theta, c) \right] p(y \mid \theta, \hat{c}^i) d\theta \tag{10}$$

So (10) can be iterative solved by $p(y \mid \theta, \hat{c}^i)$ and log-likelihood $\ln p(y \mid \theta, c)$.

The $Q(c, \hat{c}^i)$ function is symmetrical on the extreme points, so we can instead of the real extreme point by the discrete values that recent the extreme points. Therefore the symbol $\mathbf{c}$ can be seen as continuous variables [8].

And the estimation of $\hat{\mathbf{c}}$ can be completed by partial derivative and equal the result to zero of $Q(c, \hat{c}^i)$, then we get

$$\hat{c} = \frac{y \int_\theta e^{-j\theta} p(y \mid \theta, \hat{c}^i) d\theta}{\int_\theta p(y \mid \theta, \hat{c}^i) d\theta} = \frac{y \int_\theta e^{-j\theta} \exp\left(\frac{1}{\sigma^2} \mathrm{Re}\left(c^H y e^{-j\theta}\right)\right) d\theta}{\int_\theta \exp\left(\frac{1}{\sigma^2} \mathrm{Re}\left(c^H y e^{-j\theta}\right)\right) d\theta} \tag{11}$$

From (11) we know that $\mathbf{c}$ is the soft information of the signals, while $\hat{c}^i$ in (10) is decision point of signal constellation set. Therefore, according to the minimum Euclidean distance criterion, the decision symbol $\hat{\mathbf{c}}^i$ is

$$\hat{c}^i = \arg\min_{\alpha_m}\left\{\left\|\hat{c}^i - \alpha_m\right\|\right\} \tag{12}$$

So, the algorithm of signal detection based on EM can be described as follows

Step (1): Set the iterative time $i = 0$, initialization $\mathbf{c}^0 = 0$;

Step (2): According to (11) compute $\mathbf{c}^{i+1}$;

Step (3): Set $i = i + 1$, and repeat step (2) until algorithm convergence ($\mathbf{c}^{i+1} = \mathbf{c}^i$).

3   The Improved Algorithm Describer

3.1   Phase Ambiguity Description

When phase estimation is performed by means of conventional non-data-aided (NDA) algorithm, the resulting estimate exhibits a phase ambiguity, due to the rotational symmetries of the signaling constellation. Phase ambiguity resolution (PAR) can be accomplished by a data-aided (DA) algorithm that exploits the presence of a known pilot sequence in the transmitted data stream. The need for PAR can be removed by using differential encoding, which however results in BER degradation, and requires significant changes to the decoder in case of iterative demodulation/decoding. Since a phase ambiguity resolution failure gives rise to the loss of an entire packet, its probability of occurrence should be made sufficiently small [9]. At the same time, the pilot sequence must not be too long as it reduces the spectral efficiency of the system.

In section 2, we described the signal model, where the unknown carrier phase θ is in the interval $(-\pi, \pi)$. Detection of the data symbols $\mathbf{c}$ is based upon the rotated vector $\mathbf{y}e^{-j\hat{\theta}}$, with $\hat{\theta}$ denoting an estimate of the carrier phase θ.

We introduce the integer part k_θ and the fractional part ε_θ, of the phase θ, defined by

$$\theta = k_\theta \frac{2\pi}{M} + \varepsilon_\theta \tag{13}$$

Where $|\varepsilon_\theta| < \pi / M$ sand $k_\theta \in \{0,1,...,M-1\}$ .The phase estimation algorithm involves the estimation of continuous parameter ε_θ or θ, whereas phase ambiguity resolution refers to the estimation of the discrete parameter k_θ .Estimation of ε_θ and θ will be denoted by fractional phase estimation and total phase estimation, respectively [10].

3.2   The Effect of Phase Ambiguity for the Symbol Detection Performance

In part 2, we introduced the symbol detection algorithm, average the unknown carrier phase, so in the formula (11), we need to numerical integration for the carrier phase. Due to the effect of phase ambiguity problem, making the integral operation in the interval of $(-\pi, \pi)$ will result in the symbol detection performance deterioration. Therefore, we proposed that integrating within the region may cause the phase ambiguity, respectively. Take the minimum symbol error rate as the ultimate symbol detection results after convergence [11-13].

3.3   The Effect of Integration Interval for the Symbol Detection Performance

In part II, we proposed the symbol detection algorithm, the estimated value of carrier phase is obtained in ith iteration is

$$\hat{\theta}^i = \arg\left\{ \int_0^{\pi/2} e^{-j\theta} \exp\left(-\frac{\left(y - ce^{j\theta}\right)^H \left(y - ce^{j\theta}\right)}{2\sigma^2} \right) d\theta \right\} \tag{14}$$

With the convergence of the algorithm, not only obtained the symbol information, also can calculate the ML estimation of carrier phase, so realized the joint symbol detection and phase estimation. In formula (11), when the carrier phase offset at the edge of the integration interval, that is, the peak of $p(r|\theta)$ at the edge of the integration interval, caused the symbol detection performance deterioration. Therefore, we proposed the improved methods that adjusting the integration interval. That is, when jointly estimated carrier phase value is located at the edge of the integration interval, to adjust the integration interval in order to the peak of $p(r|\theta)$ is located in the middle of the integration interval [14-16].

4   Simulation Results

The performance of the proposed symbol detection algorithm and improved schemes based on EM is evaluated by computer simulations [17-19]. In simulations, the transmitted signal was selected to be MPSK modulated signal with $M = 2, 4$ .In particular, the symbol error rate (SER) as a function of E_s / N_0 are measured with simulations using two schemes:

4.1   Resolve the Effect of Phase Ambiguity for the Symbol Detection Performance

Based on the improved symbol detection algorithm was described in part 3.2, Fig. 1 and Fig. 2 indicates that the SER performances improved significantly due to resolve the phase ambiguity problem.
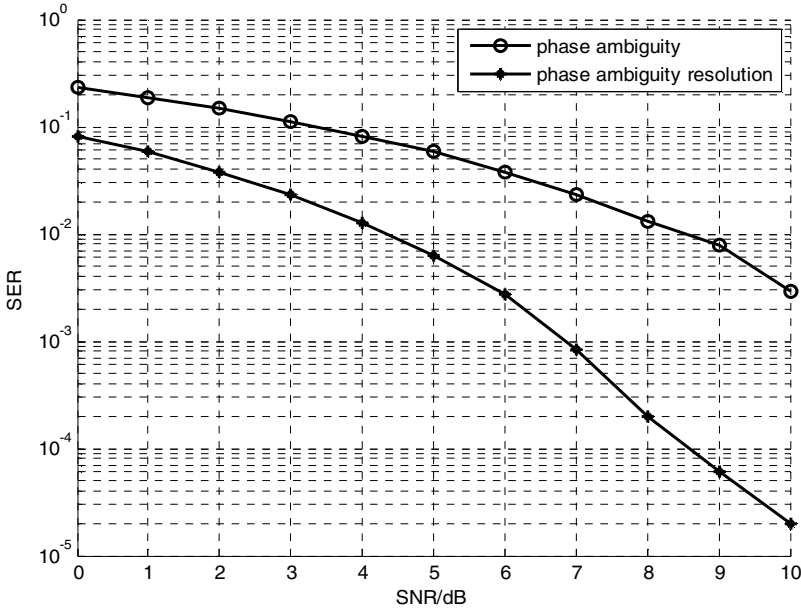
Fig. 1 The performance of BPSK symbol detectionfor the the effect of phase ambiguity
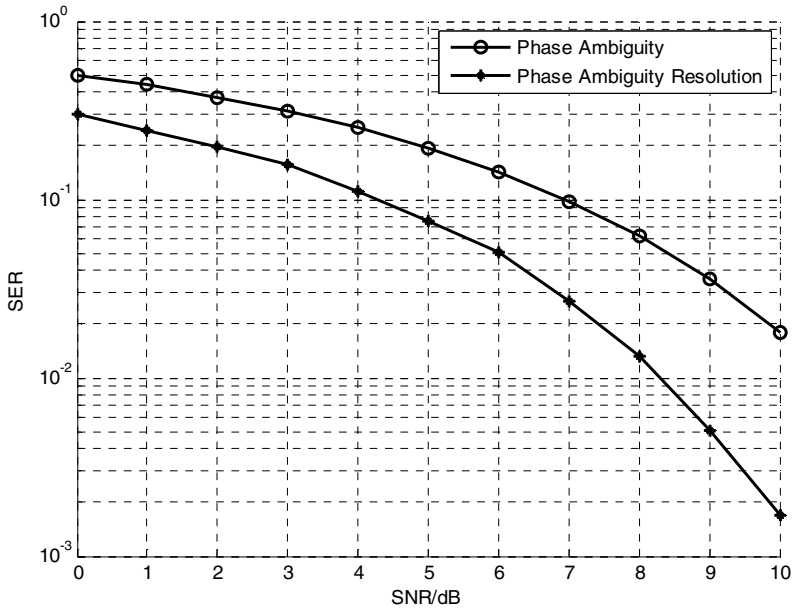
Fig. 2. The performance of QPSK symbol detectionfor the the effect of phase ambiguitt

4.2   Resolve the Effect of Integral Interval for the Symbol Detection Performance

Based on the improved symbol detection algorithm was described in part3.3, Fig. 3 and Fig. 4 respectively indicate that the symbol error rate performance curves when the carrier phase offset $\theta = \dfrac{4}{9}\pi$, the BPSK modulated signal and the carrier phase offset $\theta = \dfrac{2}{9}\pi$,the QPSK modulated signal. The simulation result shows that the algorithm performance has improved significantly".
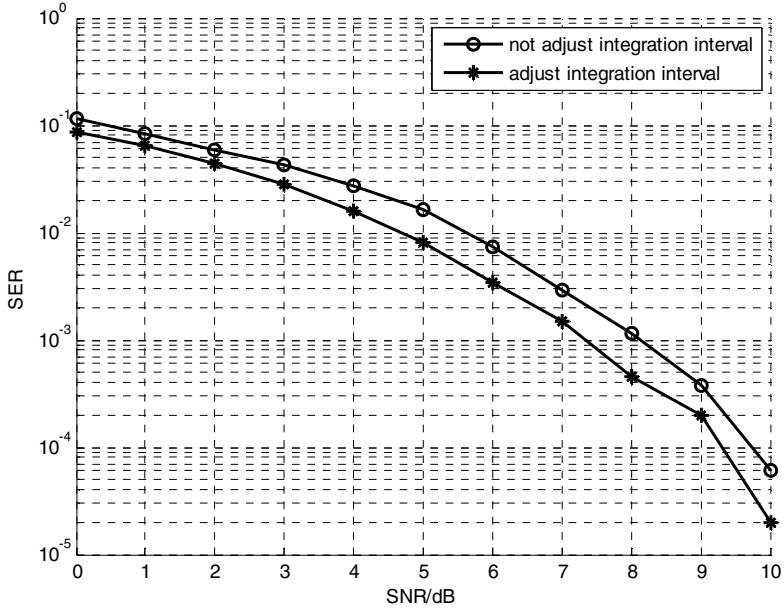
Fig. 3. The performance of BPSK symbol detection for the effect of integration interval
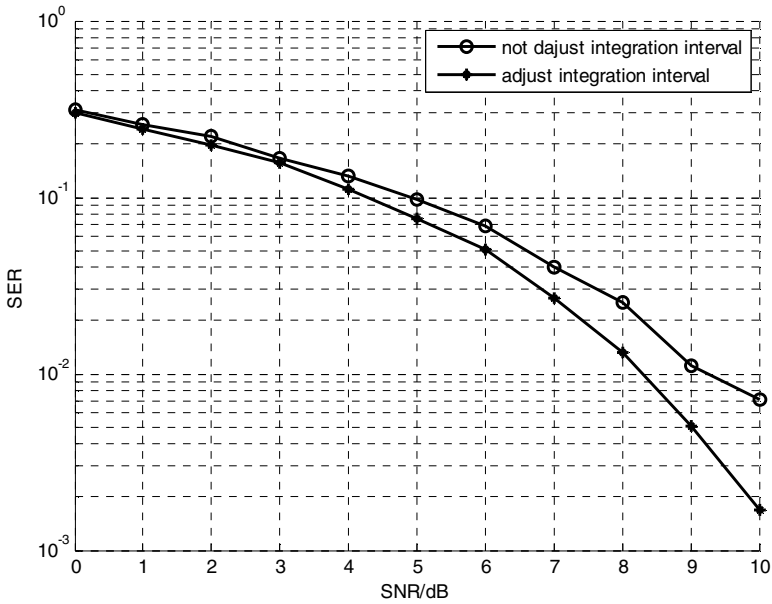

Fig. 4. The performance of QPSK symbol detectionfor the effect of integration interval

5   Conclusion

The paper against the phase ambiguity and integration interval caused symbol detection performance deterioration, proposed an improved algorithm based on expectation-maximum. The algorithm does not require carrier phase estimate, the symbol information can be obtained directly, thus reducing the performance loss caused by hierarchical processing of the traditional receiver. The deterioration of the symbol detection performance was improved by sub-regional integration and adjusts the integration interval. Simulation results show that the symbol detection performance improved significantly.

References

1. Moeneclaey, M., Jonghe, G.: ML-oriented NDA Carrier Synchronization for General Rotationally Symmetric Signal Constellations. IEEE Transactions on Communications 1(5), 2531–2533 (1994)
2. Gworghiades, C.N.: Blind Carrier Phase Acquisition for QAM Communication. IEEE Transactions on Communications 2(3), 1477–1486 (1997)
3. Dempster, Laird, N.M., Rubin, D.B.: Maximum-likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. 3(6), 189–197 (1977)
4. McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions, New York. Wiley Series in Probable. Statist., vol. 4(6), pp. 363–376 (1977)
5. Noels, N., Herzet, C., Dejonghe, A., Lottici, V.: Turbo synchronization: an EM algorithm interpretation. In: IEEE International Conference on Communication, vol. 5(8), pp. 2933–2937 (2003)
6. Nissila, M.J., Pasupathy, S., Mammela, A.: An EM Approach to Carrier Phase Recovery in AWGN Channel. In: IEEE International Conference on Communication, ICC, vol. 6(7), pp. 2199–2203 (2001)
7. Zamiri-Jafarian, H., Pasupathy, S.: Adaptive MLSDE Using the EM Algorithm. IEEE Transactions on Communication 7(11), 1181–1193 (1997)
8. Georghiades, C.N., Han, J.C.: Sequence estimation in the presence of random parameters via the EM algorithm. IEEE Transactions on Communication 8(4), 300–308 (1997)
9. Nassar, C.R., Reza Soleymani, M.: Joint Sequence Detection and Phase Estimation Using the EM Algorithm. In: Conference on Electrical and Computer Engineering, vol. 9(5), pp. 296–299 (1994)
10. Eergali, U., Sandri, A., Salivary, A.: Phase Ambiguity Resolution in Trellis-coded Modulations. IEEE Transactions on Communications 10(5), 2087–2088 (1990)
11. Wymeersch, Moeneclaey, M.: Iterative Code-Aided ML Phase Estimation and Phase Ambiguity Resolution. EURASIP Journal on Applied Signal Processing 11(5), 981–988 (2005)
12. Wiegand, T.: Study of final committee draft of joint video specification (2002)

13. Choi, W., Jeon, B., Jeong, J.: Fast motion estimation with modified diamond search for variable motion block sizes. In: Proc. IEEE Int. Conf. Image Processing, vol. 3, pp. II-371–II-374 (2003)
14. Yin, P., Tourapis, H.Y.C., Tourapis, A.M., Boyce, J.: Fast mode estimation within the H.264 codec. In: Proc. Int. Conf. on Multimedia and Expo, vol. 3, pp. III-517–III-520 (2003)
15. Ahmad, A., Khan, N., Masud, S., Maud, M.A.: Efficient block size selection in H.264 video coding standard. Electron. Lett. 40(1), 19–21 (2004)
16. Ma, S.Y., Shen, C.F., Chen, L.G.: Analysis and reduction of reference frames for motion estimation in MPEG-4 AVC JVT/H.264. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, vol. 3, pp. III-145–III-148 (2003)
17. Xuan, Z., Zhenghua, Y., Songyu, Y.: Method for detection all-zero DCT coefficients ahead of discrete cosine transformation and quantization. Electron. Lett. 34(19), 1839–1840 (1998)
18. Sousa, L.A.: General method for eliminating redundant computations in video coding. Electron. Lett. 36(4), 306–307 (2000)
19. Malvar, H., Hallapuro, A., Karczewicz, M., Kerofsky, L.: Low-complexity transform and quantization with 16-bit arithmetic for H.26L. In: Proc. IEEE Int. Conf. Image Processing, vol. 2, pp. 489–492 (2002)

A High-Efficient Distance Spectrum Algorithm for Turbo Codes

Sun-Ting Lin[1,*] and Shou-Sheu Lin[2]

[1] Department of Electronic Engineering, National Kaohsiung Normal University,
Yanchao, Kaohsiung 824, Taiwan
[2] Department of Computer and Communication Engineering, National Kaohsiung First
University of Science and Technology, Nanzih, Kaohsiung 811, Taiwan
`stl@nknu.edu.tw`

Abstract. A high-efficient distance spectrum calculation approach for turbo code is proposed in this paper. It uses periodic characteristic of recursive systematic code in turbo codes to define a fundamental period. The distance spectrum can be calculated with leadfrog method by utilizing the fundamental period look-up table instead of using bit-by-bit basis calculation. The error patterns can be classified into groups of turbo codewords in term of the different fundamental periods. Thus, the computational time of distance spectrum can be reduced dramatically. Simulation results show that the proposed algorithm is applicable to turbo codes with the block length as large as eight thousand bits with a reasonable computational time.

Keywords: Turbo code, Distance spectrum, Interleaver.

1 Introduction

Turbo code [1] is a capacity-approaching error correction code and its bit error rate (BER) performance is close to the theoretical Shannon bound. Turbo decoding exchanges the extrinsic information from one component decoder to the other and repeats the decoding process several times to achieve better decisions. Due to its superior performance and reasonable decoding complexity, it has been widely used in many communication systems, such as the 3G WCDMA, WiMAX and 4G LTE-A. The turbo code is constructed by parallel concatenating two identical recursive convolutional codes via an interleaver. To solve the high complexity problem caused by using optimal scheme and maintain high performance at same time, an efficient suboptimal scheme is employed by decoding each component code alternatively and iteratively.

The turbo code has better performance than other traditional error correction code especially in low signal to noise ratio (SNR) condition. However, it is a time-consuming process to evaluate its performance due to the iterative decoding scheme. The ultimate bit error rate (BER) performance of a turbo code is bounded by an error floor for high SNR regime. For a given component code, the level of error floor

---

mainly depends upon the interleaver [2]-[3]. Due to lack of analytical approach to evaluate the BER performance, there are two alternatives to evaluate the BER performance of turbo codes. The first is the Monte-Carlo simulation approach. The other is the semi-analytical approach named as distance spectrum [4]-[12] analysis. The Monte-Carlo method can simulate the BER performance very accurately, however, it takes very long time to perform the simulations especially when BER is less than 10^{-8}. For a well-design turbo code, it may spend 2,400 hours to for BER=10^{-10} using a very high speed software simulator [4].

Leveraging the theory of the Hamming weight distribution of turbo codewords, distance spectrum method can evaluate the BER performance much faster than the Monte-Carlo method. Researchers prefer to use distance spectrum method to study the turbo codes. The distance spectrum method is a powerful tool to design the component code and interleaver for turbo codes. It characterizes the performance of turbo codes at moderate to high SNR [13]-[15]. The method in [7] statistically evaluates the distance spectrum but its efficiency is not always guaranteed. The algorithm in [8] is proposed for the computation of the free distance of turbo codes. The symbol interleaver design using distance spectrum is studied in [9]. The method in [10] presents a new evaluation method of the distance spectrum of turbo codes when the encoder structure is given. Because the distance spectrum needs to use simulation approach to obtain the distance spectrum, how to reduce the simulation time is still a problem.

The rest of this paper is constructed as follows. Section 2 shows the system model of turbo codes. A new algorithm of the distance spectrum evaluation method is proposed in Section 3. To demonstrate the effectiveness of the proposed method, a numerical example is presented in Section 4. Section 5 concludes this paper.

2   System Model

As shown in Fig. 1, the turbo encoder consists of two recursive system code (RSC) encoders and an interleaver π.
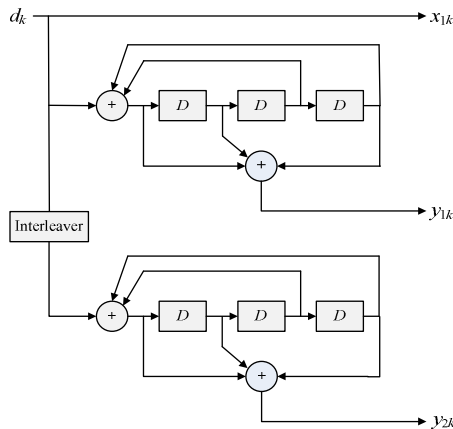

Fig. 1. Functional block diagram of a turbo encoder

The performance of a turbo code depends on the polynomials of the RSC encoders and the permutation sequence of the interleaver. The connection structures of the RSC encoders depend on the applications of turbo codes. For example, the polynomial of RSC in Fig. 1 is $(1, 1+D^2+D^3 / 1+D+D^3)$, which is defined as the standard polynomial of the LTE-A system. Therefore, designing a proper interleaver with respect to its corresponding RSC encoders can improve the performance of a turbo code.

The interleaver can lower correlation between information bits and prevent the burst error caused by noise and decoding process. There are various types of interleavers such as block interleaver and random interleaver. The basic principle of a block interleaver is to write the information bits into a matrix buffer in row and then read them out in column. Its operation is quite simple in principle but it break up information bits effectively. Due to its regular behavior, its performance is not good for turbo codes. Random interleaver writes the information bits into a matrix buffer in random order and then read them out in inverse order. Its performance is better than the block interleaver due to random property.

The BER performance of different interleaving sequences can be evaluated via the calculation of their corresponding distance spectrum. Thus, the distance spectrum can be used as an index to select the permutation sequence of the interleavers in turbo codes to improve the performance.

3   Proposed Distance Spectrum Algorithm

In this section, we present the methods to calculate distance spectrum and then propose a high-efficient algorithm to evaluate the distance spectrum of a turbo code. In the distance spectrum calculation process, an effective generation of error patterns is important because the generation process is a time-consuming procedure. In the proposed method, the utilization of the he fundamental period in the component codes can effectively skip unnecessary error patterns with leadfrog method and then shorten the computation time.

3.1   Distance Spectrum

The free distance is determined by calculating the Hamming distance between all possible codewords and the all-zero code. It can be used to estimate the performance of a turbo code, especially under high SNR condition that it is hard to estimate the system performance by BER simulation. The method to calculate free distance can be described as follows. Given a parallel concatenated turbo code, the input weight of an information block is IW, the output weight of the first RSC encoder is P_1, and the output weight of the second RSC encoder is P_2, then the total Hamming distance from all-zero code is defined as $IW + P_1 + P_2$. The distance spectrum is the spectrum of all possible codewords and the all-zero code and it can be described as the following polynomial with dummy variable X.

$$A(X) = A_1 X + A_2 X^2 + A_3 X^3 + \cdots A_N X^N \tag{1}$$

where A_i means the number of code at distance i, and N represents the length of information block. The free distance d_{free} is defined as the minimum of Hamming distance of the turbo code, i.e.,

$$d_{free} = \min\{A_i\} . \tag{2}$$

The free distance usually dominates the BER performance of a turbo code. Also, the BER performance P_b is bounded by [4]

$$\frac{N_{d_{free}} \overline{w}_{d_{free}}}{N} Q\left(\sqrt{d_{free} \frac{2RE_b}{N_0}}\right) \leq P_b \leq \sum_{d=d_{free}}^{2(N+v)} \frac{N_d \overline{w}_d}{N} Q\left(\sqrt{d \frac{2RE_b}{N_0}}\right). \tag{3}$$

where $N_{d_{free}}$ is the number of code at free distance, $\overline{w}_{d_{free}}$ is the average input weight, R is code rate, E_b / N_0 is the energy per bit, v is the length of tail bits, $Q(\bullet)$ is the Q-function.

3.2   Fundamental Period and Error Pattern

Shown in Fig. 2 is a state machine description of a RSC. It is observed that the states of the RSC change cyclically and periodically from a given initial state. For example, if the initial state is at 2, it will return to the state 2 after input three 1s. Thus, we can define the fundamental period (FP) as the shortest path to return the initial state. In this example, the FP is equal to 3. Later, we will use this property to skip unnecessary input bits to reduce the calculation time of distance spectrum.
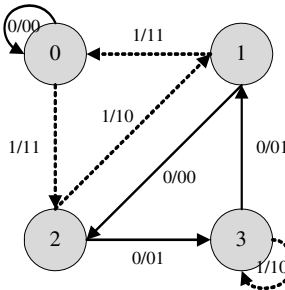

Fig. 2. State machine description of a RSC

The error pattern depends on the ratio and locations of '1' and '0' of the input information bits. The various combinations of signal '0' and '1' generate the error pattern. The information weights IW is defined as the number of 1 in the information block. For example, if the number of information block size N equals 1024 and IW equals 2, it means that there are two '1's and 1022 '0's in the information block and the number of error pattern is equal to C(1024,2)=523,776. If n times FP zeros between these two bits '1's are extracted, the error patterns can be classified into three fundamental error patterns (FEP) as shown in Fig. 3. The number of FEP is

proportional to the number of information block length and information weight. Therefore, the time required to calculate the distance spectrum depends on the above two parameters n and FEP.
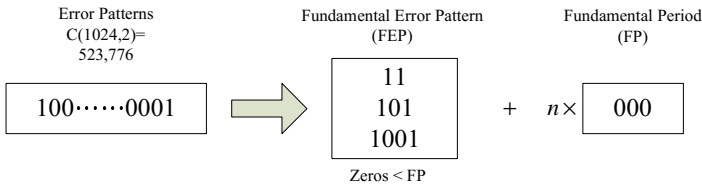

Fig. 3. The error pattern classification of a RSC for $IW=2$

Given two ones at ith and jth of codeword block respectively, the distance can be calculated by

$$d = w_{FEP} + \left\lfloor \frac{(j-i)}{FP} \right\rfloor \times w_{FP}, \tag{4}$$

where w_{FEP} is the Hamming weight of FEP and w_{FP} is the Hamming weight of FP from a given initial state and then return to the same state and $\lfloor \bullet \rfloor$ is the smallest integer function. If a look-up table of w_{FP} for all states is built in advance, the distance spectrum can be calculated with leadfrog method by utilizing the fundamental period look-up table instead of using bit-by-bit basis calculation.

3.3   Proposed Distance Spectrum Detail

Since the distance spectrum is calculated by the property of the error patterns, the proposed error pattern evaluation algorithm is presented as follow and its flow chart is shown in Fig.4.

Initialize the parameters where IW equals 2; array i, array j, and array *temp* are set to zeroes; *spread* is 0.

If $IW \geq IWmax$, skip the following steps and output DS, while *IWmax* is the maximum input weight of IW.

Initialize the locations of the '1's and record them to the array i and *temp*.

Use FP to simplify the error patterns to FEP and calculate the weight of the first level encoder P_1

If $IW + P_1 \geq IWmax$ or the last memory state is not zero, then go to step 8 else

Input the error pattern into the interleaver and obtain the interleaved error pattern (IEP). Record the location of '1's to array j.

Use FP to simplify the interleaved error patterns to FEP and calculate the weight of the second level encoder P_2.
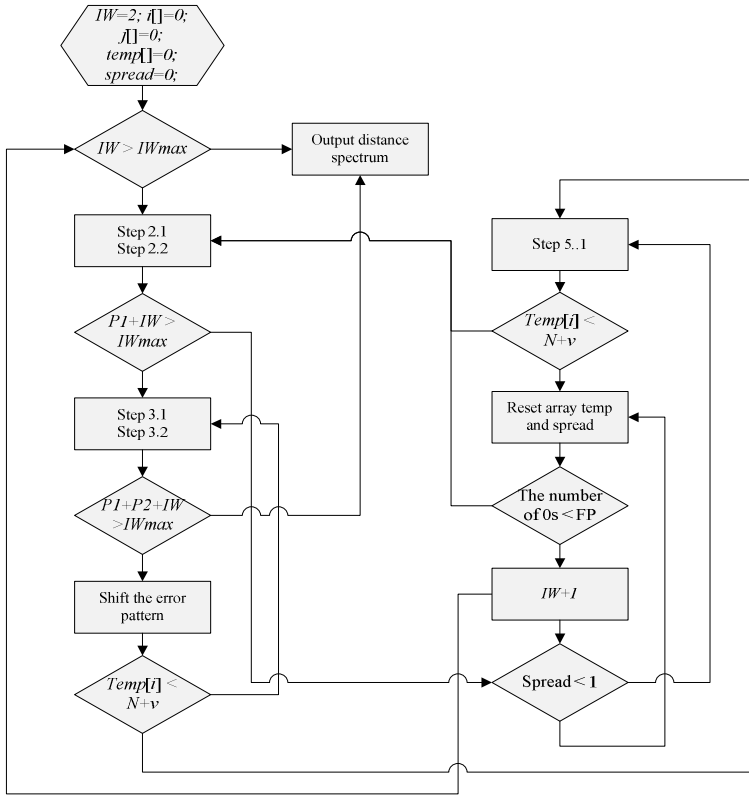
Fig. 4. The flow chart of the proposed distance spectrum algorithm

If $IW + P_1 + P_2 < IWmax$ then calculate the DS, else shift the error pattern 1 bit.

If the location of the last '1' in the shift error pattern is smaller than the block size $(N+v)$, then go back to 3.1, else

Set the parameter *spread* to 1, and add the number of FP '0's between two 1's to the error pattern.

If the location of the last '1' of the spread error pattern is smaller than the block size then go back to 2.2, else

reset the parameter *spread* to 0 and array *temp* to the original value.

Input one '0' to the error pattern and make sure that the number of '0's between two '1' signal is smaller than FP, record the location of '1' to the array *temp* and return to step 2.2 else increase the parameter *IW* and return to step 2.

If the parameter *spread* equals 1 then go to step 5.1, else go to step 6.1.

4   Simulations

In order to demonstrate the efficiency of the proposed distance spectrum algorithm, simulations were conducted to verify the performance. The simulation parameters are

as follows. The turbo code constructed by parallel concatenating two (7,5) and (15,13) RSC codes with zero-state trellis termination via a quadratic polynomial permutation (QPP) and random interleaver are used. The information block length is 6144 and the code rate is 1/3. For simplicity, the variable *IWmax2* and *IWmax4* represent the case of *IWmax* = 2 and *IWmax* = 4, respectively.
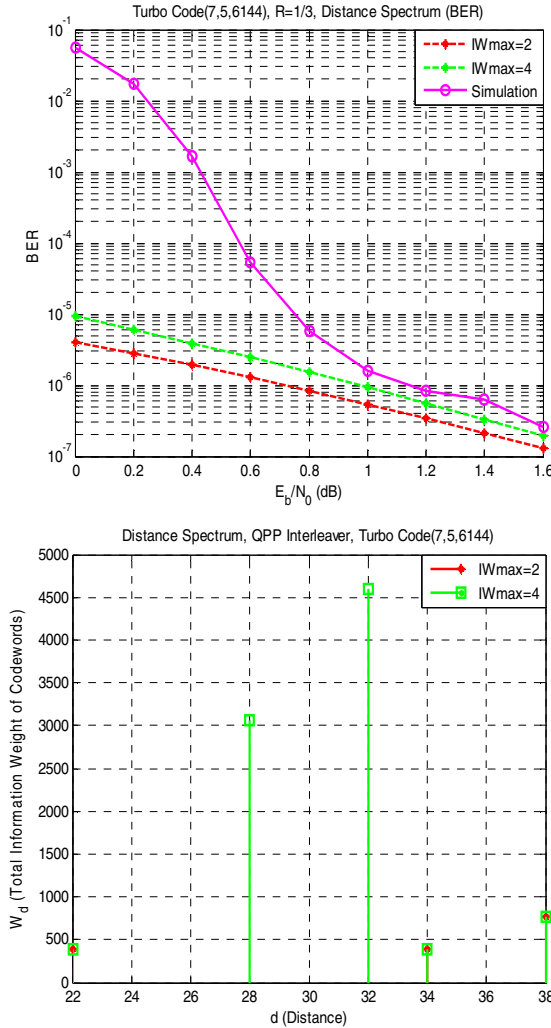

Fig. 5. (a) BER performance and (b) distance spectrum of turbo code (7, 5, 6144) with QPP interleaver

The simulation result of BER performance curve of (7,5) RSC with QPP interleaver and the predicted curve with *IWmax2* and *IWmax4* are shown in Fig. 5(a). As expected, the BER performance predicted by the proposed distance spectrum algorithm is closed

to the Monte-Carlo simulation result. Note that Monte-Carlo simulation used as the ideal reference BER curve. The distance spectrum both cases are also shown in Fig. 5(b). The curve predicted by *IWmax*4 is more accurate, but it is a more time-consuming process than *IWmax*2 because more input pattern are considered.

We also apply the proposed algorithm to search a random interleaver with better performance. Different random seeds are use to generate random interleavers. The BER simulation curve and distance spectrums are shown in Fig. 6(a) and 6(b), respectively. The seed of the random interleaver are set to 1, 10, 100, 1000, and 10000, respectively. The performance of seed=1000 is better than the seed=10 which is consistent with the results by the BER simulation. Though this is just a try with few interleavers, however, it demonstrates the feasibility of the proposed algorithm to search a better interleaver.
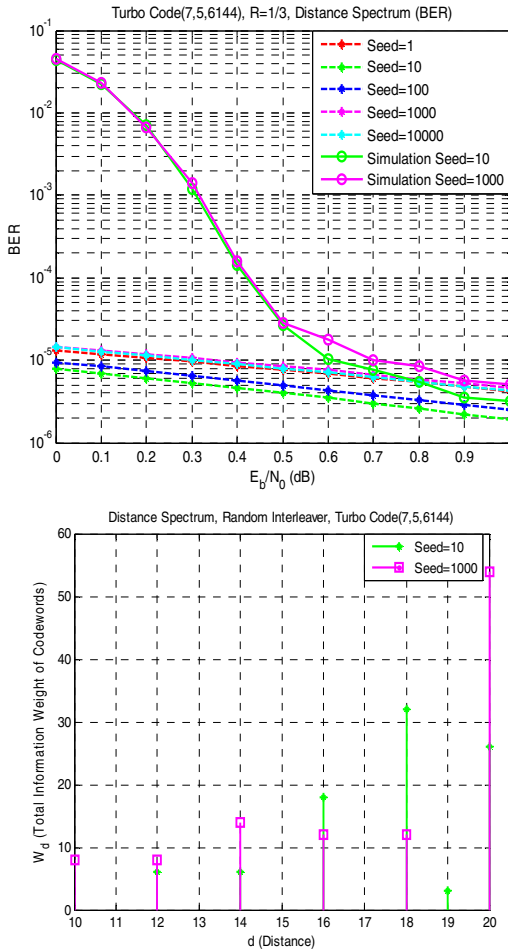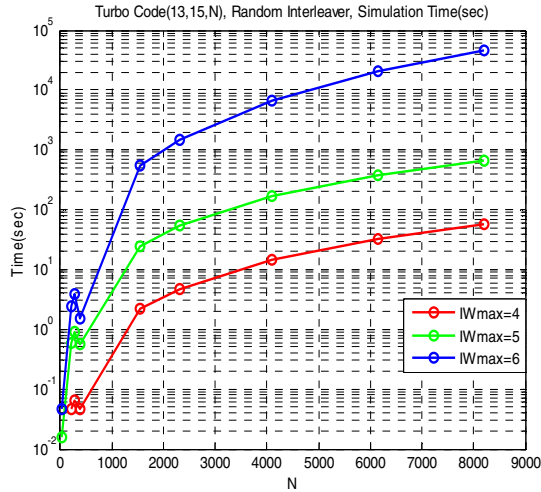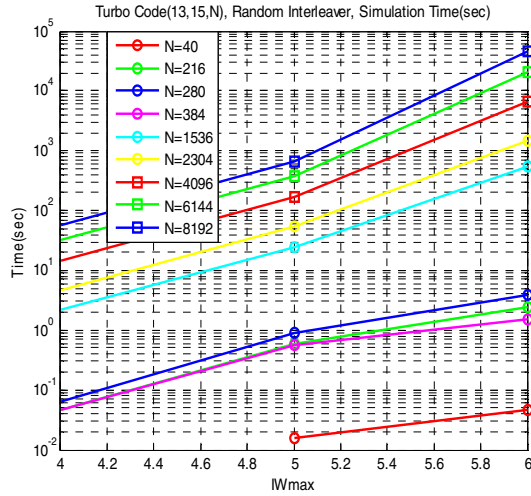

Fig. 6. (a) BER performance (b) distance spectrum analysis of turbo code (7, 5, 6144) with various random interleavers

(a)

(b)

Fig. 7. Calculation time of the distance spectrum with parameters (a) information block length *N* and (b) maximum input weight *IWmax*

The computation time of distance spectrum is an importance factor of its application. To investigate the range of the proposed algorithm, we calculate the distance spectrum for information block length up to 8,192 for the turbo code with RSC (13,15) in LTE-A standard and random interleaver. As shown in Fig. 7(a), it takes 5×10^4 seconds to complete the calculation for *IWmax*=6, which is large enough for searching the free distance. For *IWmax*=4 and block length=8,192, it takes only 60 seconds to complete the calculation. Also the computation time is evaluated with

different maximum input weight up to 6 shown in Fig. 7(b). It is observed from these figures that the computation time is monotonically proportional to information block length and maximum input weight. The proposed algorithm is feasible to a wide range of turbo codes application.

5   Conclusions

The improved distance spectrum calculation algorithm is proposed. It can significantly reduce the computation time required to calculate the distance spectrum. The computation time of the proposed algorithm is much less than the simulation time of traditional Monte-Carlo method. Simulation results show that the proposed algorithm can closely estimate the BER performance at the high SNR region. Also, it can apply to a wide range of application for turbo codes performance simulation and design problem.

References

1. Berrou, C., Glavieux, A., Thitimajshima, P.: Near Shannon Limit Error-Correcting Coding and Decoding: Turbo-codes. In: Proc. IEEE Int. Commun. Conf. Geneva, Switzerland, pp. 1064–1070 (1993)
2. Trifina, L.: Two Methods to Increase the Minimum Distance for Turbo Codes with QPP Interleavers. In: IEEE International Symposium on Signals, Circuits and Systems, pp. 1–4. IEEE Press, New York (2009)
3. Trifina, L., Gherca, L., Lupu, B., Rotopanescu, A.-M.: Modified Ω' Metric for QPP Interleavers Depending on SNR. In: IEEE International Symposium on Signals, Circuits and Systems, pp. 1–4. IEEE Press, New York (2009)
4. Garello, R., Pierleoni, P., Benetto, S.: Computing the Free Distance of Turbo Codes and Serially Concatenated Codes with Interleavers: Algorithms and Applications. IEEE J. Sel. Areas Commun. 19, 800–812 (2001)
5. Rouanne, M., Costello, D.J.: An Algorithm for Computing the Distance Spectrum of Trellis Codes. IEEE J. Sel. Areas Commun. 7(6), 929–940 (1989)
6. Perez, L.C., Seghers, J., Costello, D.J.: A Distance Spectrum Interpretation of Turbo Codes. IEEE Trans. Inf. Theory 42(6), 1698–1709 (1996)
7. Breiling, M., Huber, J.: A Method for Determining the Distance Profile of Turbo Codes. In: Proc. Third ITG Conf. on Source and Channel Coding, pp. 219–224 (2000)
8. Scalise, S., Bae, Y.K., Ernst, H.: A Fast Algorithm to Estimate the Distance Spectrum of Turbo Codes. In: Int. Conf. Telecomm. pp. 90–95 (2003)
9. Wu, Y., Ogiwara, H.: Symbol-interleaver Design for Turbo Trellis-Coded Modulation. IEEE Commun. Lett. 8, 632–634 (2004)
10. Sakai, T., Shibata, K.: A Distance Spectrum Estimation Method of Turbo Codes. In: Int. Symp. on Inf. Theory and its Applic., Auckland, New Zealand (2008)
11. Benedetto, S., Montorsi, G.: Unveiling Turbo Codes: Some Results on Parallel Concatenated Coding Schemes. IEEE Trans. Inf. Theory, 409–428 (1996)

12. Cedervall, M.L., Johannesson, R.: A Fast Algorithm for Computing Distance Spectrum of Convolutional Codes. IEEE Trans. Inf. Theory 35, 1146–1159 (1989)
13. Matuz, B.: On the Distance Spectrum of Tail-Biting Repeat-Accumulate Codes. IEEE Commun. Lett. 16, 518–521 (2012)
14. Zhou, H., Mitchell, D.G.M., Goertz, N., Costello, D.J.: Distance Spectrum Estimation of LDPC Convolutional Codes. In: IEEE International Symposium on Information Theory, pp. 1–4. IEEE Press, New York (2012)
15. Ould-Cheikh-Mouhamedou, Y.: A Simple and Efficient Method for Lowering the Error Floors of Turbo Codes that Use Structured Interleavers, vol. 16, pp. 392–395 (2012)

Granular Sketch Based Uncertain Data Streams Pattern Mining

Jingyu Chen[1,2], Ping Chen[2], and Xian'gang Sheng[3]

[1] School of Computer Science and Technology, Xidian University, 710071, Xi'an, China
[2] Software Engineering Institute, Xidian University,710071, Xi'an, China
`{Jychen,chenping}@mail.xidian.edu.cn`
[3] College of Information Engineering, Qingdao University, 266071, Qingdao, Shandong, China
`stzhibuwu@qdu.edu.cn`

Abstract. Uncertainty is inherent in data streams, and presents new challenges to data streams mining. For continuous arriving and large size of data streams, representations of uncertain time series data streams require significantly more space. Therefore, it is important to construct compressed representation for storing uncertain time series data. A granular sketch is designed to create hash-compressed storage and store granules. As the granular sketch may be saturated with the increasing of data streams, this paper presents an optimization strategy to delete the absolute sparse patterns. Based on the granular sketch, a sequential pattern mining algorithm is proposed for mining uncertain data streams. The experimental results illustrate the effectiveness of the pattern mining algorithm.

Keywords: granulation, sketch, pattern mining, uncertain time series data, data streams.

1 Introduction

With the developing of information technique and wireless sensor networks, a great variety application need to analysis and mining large-scale, continuous and rapid arriving data streams. As a common data form in data streams, time series data streams contain much more useful pattern and information for many applications fields.

Time series is a series of observation data according to a certain time sequence, and aggregates with time and event [9]. Time-series data mining is an important way which mining some useful and potential knowledge from a great deal of time-series. Time series data are created by many applications such as RFID, traffic, mobile, economic and finance applications. In these applications, uncertainty often happens because of network failure, noise and sampling error, etc. Probability time series are common uncertain time series data. The probabilistic information should be handled in sequential pattern mining uncertain time series data streams.

To address the problem of uncertain sequential pattern mining we suggest a hash based granular sketch approach to reduce storage spaces and computational complexity. The approach is based on granulation mechanism, which granulates

uncertain time series data to a set of possible sequences. And we use sketch technique to compress all possible sequences to a granular sketch. As the granular sketch may be saturated with the increasing of uncertain data streams, an optimization strategy is designed to delete the absolute sparse patterns. Based on the granular sketch, a sequential pattern mining algorithm is proposed for mining uncertain time series data streams. Final, we verify the accuracy and efficiency of the proposed scheme via experiments.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 presents a granular sketch model for uncertain time series data. Section 4 outlines the granular sketch based pattern mining methods. Simulation methodology and performance evaluation result and analysis are presented in section 5, and we conclude the work in section 6.

2   Related Work

Time series data mining has been attracting much attention in research and practice. As a hot research area in time series data mining, frequent pattern mining can be mainly classified into two broad categories: Apriori-based [3,7,8] and Tree-based [2,9,10] algorithms. As an inherent property of data streams, uncertainty brings great challenge to data streams mining.

Uncertain time series often have same characteristics as uncertain data streams, such as uncertain, large-scale, continuous and rapid arriving. To handle the continuous arriving uncertain time series streams, we often expand some streams mining algorithms to mine frequent patterns uncertain streams.

In recent years there have been a plethora of methods for managing and mining uncertain data streams. Several important stream pattern mining algorithms have been introduced in recent years. Chui et al. present an uncertain data model and propose a U-Apriori to mine frequent itemsets from uncertain data [4]. Based on FP-growth [2], Leung et al. propose two tree-based mining algorithms UF-Streaming and SUF-growth to efficiently find frequent itemsets from streams of uncertain data, where each item in the transactions in the streams is associated with an existential probability [1]. Kaneiwa et al. propose a method for mining such local patterns from sequences by using rough set theory [16]. Ackermann et al. present a new corsets trees based clustering algorithm to improve quality of stream clustering [5]. Tran et al. present the PODS model for processing uncertain data using continuous random variables [6].

Sketch is a popular method for handling huge and fast data streams. Sketch techniques use a sketch vector as a data structure to store the streaming data compactly in a small-memory footprint. The main advantage of using these sketch techniques [11, 12] is that they require a storage which is significantly smaller than the input stream length. Sketch techniques are used in stream data frequent items mining [13, 14], clustering [15] and anomaly detection [17, 18] recently.

Our work is closely related to mine frequent patterns of uncertain time series data. In this paper, we model the uncertain time series data based on granular sketch. We use a granular sketch approach to create hash-compressed representations. We also

design an optimization strategy to avoid the granular sketch saturated. Then, we propose a sequential pattern mining algorithm to mine frequent sequential pattern of the uncertain time series data streams.

3   Granular Sketch

One of the most effective ways to deal with imprecise and uncertain data is to employ probabilistic approaches. Since we may get several probabilistic points at a certain time t, the probabilistic time-series data may require much more size for storage. The probability also increases the complexity of modeling and analysis of time series data. The way of modeling the probability of data is a key point of uncertain data stream mining.

Probability time series is a kind of uncertain time series data, which has time, event and event's probability. We use the following definition to represent and store the probability time series data,

Definition 1. (Element) An element e= <x,p,t> is a basic element of probability time series data, where x represents an observation value , p represents the probability of x value,   and t represents observation time.

Definition 2. (Item) An item I_t of a sequence s_i at time t includes all possible elements at time t. $I_t = \{(<x_1,p_1,t>, <x_2,p_2,t>,…)\}$

Definition 3. (Uncertain sequence) A uncertain sequence us_i is an ordered list of items that include n consecutive items of uncertain time series, $us_i=\{ I_1, I_2,…I_n\}$. Examples of Uncertain time series data are showed in Table 1.

Table 1. Examples of Uncertain time series data

Time	us_1		us_2		us_3	
t	x	p	x	p	x	p
1	a	0.6	c	0.5	f	0.4
1	b	0.4	d	0.5	e	0.6
2	a	0.5	c	0.8	d	0.7
2	b	0.5	d	0.2	f	0.3
3	a	0.3	c	0.2	g	0.2
3	b	0.4	d	0.4	b	0.5
3	e	0.3	e	0.4	d	0.3

In Table 1, us_1 has two possible items $I_1=\{<a, 0.6, 1>, <b, 0.4, 1>\}$ at the time t=1.

And, the uncertain sequence of us_1 is $\{I_1,I_2,I_3\}=\left\{ \binom{a:0.6}{b:0.4} \binom{a:0.5}{b:0.5} \begin{pmatrix} a:0.3 \\ b:0.4 \\ e:0.3 \end{pmatrix} \right\}$.

The probability increases difficulty for analysis uncertain time series, because it generates much more possible combination of sequences and makes sequences model much complex.

3.1   Granularities of Uncertain Sequences

Granular cognition plays an important role for complex data modeling. We adopt granulation mechanism to model and represent the complex uncertain time series data.

To deal with the diversity of uncertain sequential data, we describe sequences of time points using a set of granules. We consider the different length of subsequences time points as different granules. We define the following concepts for uncertain time series granulating.

Definition 4. (Granule) A granule g_i is a set of possible element values at a time interval and the possibility value of each value should larger than $\gamma=0.3$, where γ is the probability threshold that can be defined by users.

We can directly get a granule from time at a time and we denoted this granule by basic granule. For instance, in Table 1 at time t=1, us_1 has two possible values a and b, and the possibilities values of a and b are p(a)=0.6 and p(b)=0.4 , and both are larger than $\gamma=0.3$. Let p(a) be the possibility value of a. Therefore we can use a granule $g_1=$ {a, b}, and g_1 is a basic granule.

A granule also can be generated within a time interval, which means we can combine two or more granules which are consecutive in time into a new granule. For example, for the uncertain sequence us_1, at time 1 and 2, we can get two same basic granules $g_1=$ {a, b}. We combine the values of two granules into 2-length sequences {aa, ab, ba, bb}, and the possibilities of each sequence are

    p(aa)= 0.6*0.5=0.3;
    p(ab)= 0.6*0.5=0.3;
    p(ba)= 0.4*0.5=0.2;
    p(bb)= 0.4*0.5=0.2.

The possibility values of sequences ba and bb are smaller than $\gamma=0.3$.So, we can generate a new 2-length granule g2={g1g1}={aa, ab}.

Definition 5. (K-granule) A k-granule Gk(usi) is a sequences set that includes k-length sequences (patterns). We define the k-granules $G_k(us_i)$ of a uncertain sequence us_i to be the set of all unique k-length granules of us_i.

In Table 1, for the uncertain sequence us1, let g3={a, b, e} is the granules at time t=3. From time t=1 to t=3, the 2-granules G2(us1) of the uncertain sequence us1 is the set of 2-length granules, G2(us1)={ g1g1, g1g3}. For g1g3, the 2-length sequences are {aa, ab, ae, ba, bb, be}, and the corresponding cumulative possibilities of each sequence are:

    p(aa)= 0.5*0.6+0.5*0.3=0.45;
    p(ab)= 0.6*0.5+ 0.5*0.4=0. 5;
    p(ae)= 0.5*0.3=0.15;
    p(bb)= 0.4*0.5+ 0.4*0.5=0.4;
    p(ba)= 0.4*0.5+0.5*0.3=0.35;
    p(be)= 0.5*0.3=0.15.

With the probability threshold $\gamma=0.3$, we can detail the 2-granule $G_2(us_1)$={aa, ab, ba, bb}. We obtain the following sets of 2-granules.

 G2(us2)={cc, dc, cd, ce}
 G2(us3)={ed, db}

We can also construct 3-granules and even k-granules for different uncertain time series. The k is the possible largest length of sequential patterns and can be defined by user. For an uncertain time series, we can build a set of k-1 granules from 2-granules to k-granules for all uncertain time series, and denote the set as k-granules set. Based on the k-granules set, we construct a granular sketch to store the sequences patterns in the k-granules set.

3.2   Granular Sketch

Sketch based approaches [10] were designed for enumeration of different kinds of frequency statistics of data sets. A commonly-used sketch is the count-min method [10]. The count-min sketch use w = $\lceil \ln(1/\delta) \rceil$ pairwise independent hash functions, each hash function maps data into uniformly random integer in the range h = [0, e/ϵ], where e is the base of the natural logarithm. The data structure itself consists of a two dimensional array with w·h cells with a length of h and width of w. Each hash function corresponds to one of w 1-dimensional arrays with h cells each. In standard applications of the count-min sketch, the hash functions are used in order to update the counts of the different cells in this 2-dimensional data structure.

Assume that the number of uncertain sequences is N in uncertain time series data. As the continuous arriving of uncertain time series data, we construct a granular sketch for k-granules sets to reduce the scale of uncertain time series data. We compress various possible sequences of uncertain time series to a granular sketch. Rather than count all possible sequences, we calculate the possible sequential patterns on the granular sketch.

Definition 6. (Granular Sketch)A granular sketch includes a two-dimensional matrix SK[w, c] (w<<N, c), c= e/ϵ is the maximum value of hash value range and w is the number of hash functions hr[w]. The sequential patterns of the k-granules set are denoted by SJ, such as {aa, ab, …}. Let hr_i be the i-th hash function in hr[w]: SJ→{0,…,c} be a hash function that hash a sequential pattern and its possibility value to the i row number and store the possibility value at the hr_i column. The initial value of each element in the granular sketch is 0. For each sequential pattern sj in the k-granules set, we add possibility value P(sj) of sj in the granules to SK[i, hr[i](sj)].

$$SK[i, hr_i(sj)]= SK[i, hr_i(sj)]+ P(sj). \qquad (1)$$

In the granular sketch, each row represents a k-length sequential pattern. Using granular sketch, we can calculate the support and minimal support for each pattern. The support Sup for a pattern sj is the sum of the row of h(sj),

$$Sup(sj)=avg(\Sigma(SK[h(sj)])) \qquad (2)$$

Where avg is a average function.

The minimal support minSup for a pattern sj is the minimum value of the row of h(sj),

minSup(sj)=min(SK[h(sj)]).

To store the pattern information, we also construct a granular array GA[c] to store a m-granule (the value range of m is [2..k]). The initial value of each element in granular array GA[c] is null. We construct the granular array GA in then following strategy.

When a m-granule g_j is added to the granular sketch, for each pattern sj in g_j, we check the granular sketch and find the column c_{m1} that stores the most minimal minSup(sj) in m-granule g_j (sj∈g_j).

We check the PA[c_m], if PA[c_{m1}] is null, we store k-granule g_j into PA[c_m];

If the PA[c_{m1}] is not null, we find the column c_{m2} that stores the second minimal minSup(sj);

If PA[c_{m2}] is null, we store m-granule g_j into PA[c_{m2}],else find the third column c_{m3};

If half of patterns in g_j are checked and still no null column, then we will not store g_j .

The pattern array PA will store the more frequency m-granules.

Since the size of the granular sketch is much less than the size of primitive uncertain data streams, as data continuous arriving, the granular sketch may be filled full and saturated. Therefore, we adopt optimization technique to delete the absolute sparse pattern in the granular sketch.

4   Sequential Pattern Mining Algorithm

According the characteristics of the uncertain data stream, we define uncertain frequent pattern for time series pattern mining.

Definition 7. (Uncertain frequent pattern). A pattern sj is an uncertain frequent pattern, if its support Sup(sj) is larger than MinS, where MinS (0<MinS<1) is   a user-defined parameter.

Definition 8. (sparse pattern). A pattern sj is a sparse pattern, if Sup(sj)/N is smaller than MinS, where MinS (0<MinS<1) is a user-defined parameter.

As the uncertain data streams continuous arrival, the number of sequential patterns, which are stored in the granular sketch, increases rapidly and makes the sketch saturated. Therefore, it is important to delete those sparse sequential patterns.

4.1   Absolute Sparse Pattern

Due to the fast-changing characteristics of the data streams, some sparse patterns that were deleted quickly appeared in large numbers, even grown into frequent patterns. Therefore, some sparse patterns which have potential growth should be retained. We analyze the potential growth of sparse patterns and divide them into four categories:

Some patterns appear in small number and always belong to the sparse mode;

Some patterns used to be non-sparse patterns. Later due to small number of appearance, those patterns degraded to sparse patterns, and keep to be sparse patterns for a long time;

Some patterns used to be non-sparse patterns and degraded to sparse patterns. Then, those patterns appear frequently and have potential growth;

Some patterns are new patterns, but appear frequently and have potential growth.

From the above four categories, we can see the potential growth of categories 1) and 2) are very small. We should recognize those patents that belong to categories 1) and 2) as absolute sparse patterns and delete it form the granular sketch.

We identify absolute sparse patterns according the changes of patterns over time. For each pattern we calculate its shortest time interval δ_0 that degraded to sparse patterns.

$$\delta_0 = \left\lfloor \frac{1}{\lambda} \log_2 \left(\frac{\min Sup - \varepsilon}{(\min Sup - \varepsilon) - 1} \right) \right\rfloor \qquad (4)$$

where $\varepsilon (0 < \varepsilon < \min Sup)$ is a user defined parameter and λ is a attenuation coefficient $(\lambda = 0.0001)$. For each row in granular sketch, we add two columns to store the most recent creation time t_o and the most recent modification time t_m. Let $T_d = \delta_0$ be the degraded period of a pattern. For each pattern sj, from its t_o to t_m, if its support is Sup(sj)/N $< \varepsilon$, then it will be a sparse pattern. For a sparse pattern sj, if its support is still Sup(sj)/N $< \varepsilon$ during future time period $[t_c, t_c + T_d]$, the sparse pattern sj will become a absolute sparse pattern. And the all value of row h(sj) in the granular sketch will be deleted.

4.2   Sequential Pattern Mining Algorithm

We adopt basic time windows technique to process the continuous arriving time series data. The size of each basic window is L. Let B[i] be the i-th basic windows. In figure 1, B[1] is the first window, and B[n] be the most recent window.
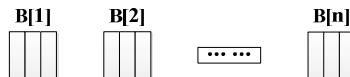

Fig. 1. Basic time windows

According the granular sketches and optimization strategy constructed above, we proposed a new sequential pattern mining algorithm UG-Miner based on basic time windows for uncertain time series data streams.

The algorithm of UG-Miner is shown as following.

Algorithm of sequential pattern mining algorithm
Input:
granular sketche SK[w,c] and granular array GA[w]
basic windows B[1], B[2],…
user-given parameters: γ,ε
user-given the minimal support threshold MinS
Output:
A set of sequential pattern: fpset
PrCount=0;   //initial a pattern counter
for the first basic window B[j]
store pattern in SK and GA
for each granule g_i in GA
for each pattern p_i in g_i
check p_i in SK;
if Sup(p_i)> MinS
add p_i to the fpset;
PrCount=PrCount+1;
if Sup(p_i)/N <ε
delete all   in SK;
next pattern;
if no pattern in g_i is add to fpset
delete g_i in GA;
if(PrCount>w)
delete the absolute sparse pattern in SK;
next granule;
if half of granules in GA are not add to fpset
increase MinS= MinS+ MinS*ε/j;
if all granules in GA are add to fpset
decrease MinS= MinS+ MinS*ε/j;
next basic window;

The key steps of the UG-Miner algorithm are designed to check and count the possibility value in the granular sketch to approximately statistics sequential patterns. And the UG-Miner algorithm updates the granular sketch and the granular array, according the number of sequential patterns and optimization strategy.

5   Experiment

This section shows the results from our experiment to validate the accuracy and efficiency of our proposed sequential pattern mining algorithm.

We will show the accuracy of our algorithm based on the evaluation of coverage. Cover is the coverage of the algorithm results,

$$Cover = \frac{|C_M \cap C_T|}{C_T} \tag{5}$$

where C_M is the results frequent patterns set of algorithm, and C_t is the real frequent patterns set.

We compare our UG-Miner algorithm to SUF-Growth algorithms by using the datasets generated by the program developed at IBM Almaden Research Center [3].

Figure 2 shows the comparison of coverage between UG-Miner and SUF-Growth with increasing size of data set. In figure 2, the coverage of SUF-Growth is higher than that of UG-Miner, because the UG-Miner is an approximate algorithm and use hash compressed storage to save memory space and improve computational efficiency. The efficiency of UG-Miner does not change very much with the increase in the size of data set.
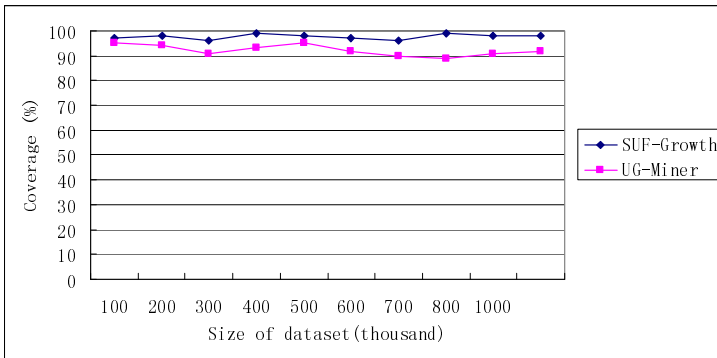

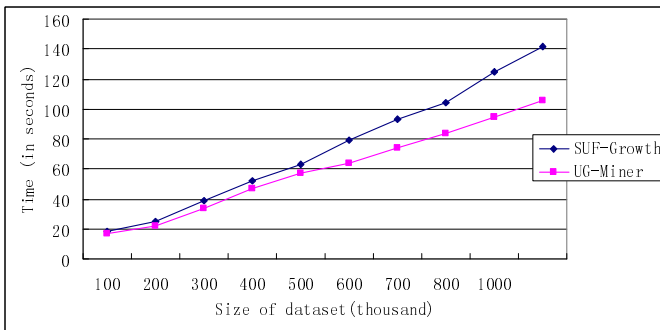
Fig. 2. Coverage comparison

Fig. 3. Time consuming comparison

Figure 3 shows the comparison of time consuming between UG-Miner and SUF-Growth with the increase in the size of testing data set. In figure 3, the time consuming of UG-Miner is smaller than that of SUF-Growth, because the hash compressed sketch can reduce the statistics time.

Figure 4 shows the comparison of memory space consuming between UG-Miner and SUF-Growth with the increase in the size of objects set. In figure 4, the memory space consuming of UG-Miner is smaller than that of SUF-Growth, because the UG-Miner adopt granular sketch and hash method to compress storage space.
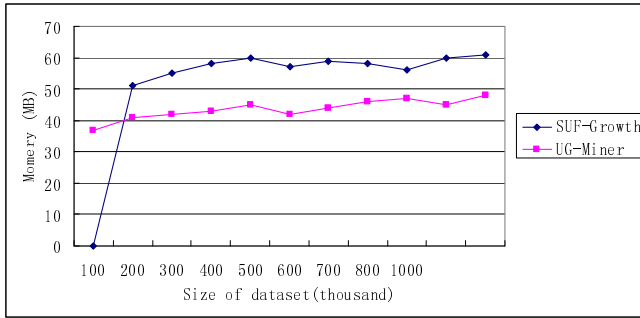
Fig. 4. Memory space comparison

Experiment results show that the UG-Miner algorithm can save memory space and improve efficiency. The UG-Miner benefits from using compressing storage space and deleting the absolute sparse patterns.

6   Conclusion

This paper has addressed a new efficient sequential pattern mining algorithm for uncertain time series streams. First, for reducing memory consumption, we construct a granular sketch and a granular array to store continuous arriving uncertain time series data and granules. For the continuous averring and massive size of the uncertain data streams, the granular sketch may be saturated. Therefore, we design an optimization strategy to delete the absolute sparse patterns. And then, based on the granular sketch, the UG-Miner algorithm for sequential pattern mining uncertain time series streams is given. Experiment results show that the UG-Miner algorithm can effectively mine frequent patterns in uncertain data streams. In future work, we plan to design distributed granular sketches and pattern mining algorithm for distributed uncertain streams.

References

1. Leung, C., Hao, B.: Mining of frequent items from streams of uncertain data. In: Proc. IEEE Computer Society, pp. 1663–1670 (2009)
2. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: Proceedings of the 2000 ACM SIGMOD, pp. 1–12 (2000)
3. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proc. 20th Int. Conf. VLDB, pp. 487–499 (1994)
4. Chui, C.-K., Kao, B., Hung, E.: Mining frequent itemsets from uncertain data. In: Zhou, Z.-H., Li, H., Yang, Q. (eds.) PAKDD 2007. LNCS (LNAI), vol. 4426, pp. 47–58. Springer, Heidelberg (2007)

5. Ackermann, M.R., Lammersen, C., Martens, M., Raupach, C., Swierkot, K., Sohler, C.: StreamKM++: A Clustering Algorithm for Data Streams. Journal of Experimental Algorithmics (JEA) 17(1) (July 2012)

6. Tran, T.T.-L., Peng, P., Li, B.D., Diao, Y., Liu, A.N.: PODS: a new model and processing algorithms for uncertain data streams. In: Proceedings of the 2010 International Conference on Management of Data, Indiana, USA, pp. 159–170 (2010)

7. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proc. of 20th ICDE, pp. 487–499 (1997)

8. Bonchi, F., Lucchese, C.: Pushing tougher constraints in frequent pattern mining. In: Ho, T.-B., Cheung, D., Liu, H. (eds.) PAKDD 2005. LNCS (LNAI), vol. 3518, pp. 114–124. Springer, Heidelberg (2005)

9. Cheung, W., Zaiane, R.: Incremental mining of frequent patterns without candidate generation or support constraint. In: Proc. IDEAS, pp. 111–116 (2003)

10. Leung, C.K.-S., Mateo, M.A.F., Brajczuk, D.A.: A tree-based approach for frequent pattern mining from uncertain data. In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD 2008. LNCS (LNAI), vol. 5012, pp. 653–661. Springer, Heidelberg (2008)

11. Alon, N., Matias, Y., Szegedy, M.: The Space Complexity of Approximating the Frequency Moments. In: ACM Symposium on Theory of Computing, pp. 20–29 (1996)

12. Cormode, G., Muthukrishnan, S.: An Improved Data-Stream Summary: The Count-min Sketch and its Applications. Journal of Algorithms 55(1), 58–75 (2005)

13. Cormode, G., Muthukrishnan, S.: What's hot and what's not: Tracking most frequent items dynamically. In: Proceedings of the 22nd ACM Symposium on Principles of Database Systems, pp. 296–306 (2003)

14. Manerikar, N., Palpanas, T.: Frequent items in streaming data: An experimental evaluation of the state-of-the-art. Technical Report DISI-08-017, University of Trento (March 2008)

15. Aggarwal, C.: A Framework for Clustering Massive-Domain Data Streams. In: IEEE 25th International Conference on Data Engineering (ICDE 2009), pp. 102–113 (2009)

16. Kaneiwa, K., Kudo, Y.: A sequential pattern mining algorithm using rough set theory. International Journal of Approximate Reasoning 52(6), 881–893 (2011)

17. Liu, Y., Zhang, L., Guan, Y.: Sketch-based Streaming PCA Algorithm for Network-wide Traffic Anomaly Detection. In: Proc. of IEEE 30th International Conference on Distributed Computing Systems (ICDCS), pp. 807–816 (2010)

18. Kanda, Y., Fukuda, K., Sugawara, T.: Evaluation of Anomaly Detection Based on Sketch and PCA. In: Proceedings of IEEE Global Telecommunications Conference (GLOBECOM), pp. 1–5 (2010)

A Quick Attribute Reduction Algorithm Based on Incomplete Decision Table

Zhangyan Xu, Jianhua Zhou, and Chenguang Zhang

College of Computer Science & Information Technology Guangxi Normal University, 541004, Guilin, China
xyzwlx72@yahoo.com.cn, zjh1988100@163.com, zhangchenguanga@sohu.com

Abstract. As the object of study incomplete decision table, with the study of the notion of conflict region, the definition of attribution reduction based on conflict region in incomplete decision table is provided. it is proved that the attribute reduction is equivalent to the attribute reduction based on positive region, at the same time ,a new attribute reduction algorithm which is in incomplete decision table is designed, whose time complexity is $O(|K||C|^2|U|)\left(|K| = \max\{|T_C(x_i)|, x_i \in U\}\right)$ .Finally, an example is used to illustrate the efficiency of the new algorithm.

Keywords: rough set, incomplete decision, conflict region, attribution reduction.

1   Introduction

Rough set theory is put forward by Z. Pawlak in 1980s. And it is a mathematical tool which is used to study and deal with the incomplete data and imprecise knowledge. It is used in artificial intelligence and cognitive science, especially in the field of intelligent information processing; it has been widely used [1-4]. The attribution reduction delete irrelevant or redundant attributes in the same classification or decision-making capacity of the Knowledge Base case [5-8].

Some scholars have proved that seeking decision table all attribute reduction and minimum attribute reduction is an NP-Hard problem [9-11]. Many scholars design various of algorithms for attribution reduction in complete decision table. In practical applications due to the measurement error of the data, restrictions of knowledge acquisition and other various reasons, there will be some default values in the decision table. So the decision table that we often have to deal with is incomplete. In general, the tolerance class algorithm of object concentrated two comparison, compare them in each attribute property set whether or not to meet the definition, if meet the tolerance class, belong to the same tolerance class; or to the object of every object, according to whether the value judgment of its property belongs to the tolerance existing class. Reference [4] designs an algorithm for attribution reduction based on positive region in incomplete decision table. Reference [5, 6] gives a method

based on information entropy. The time complexity of both algorithms is $O(|C|^3|U|^2)$ .Reference [5] gives a method based on discernibility matrix, and its time complexity is $O(|C|^2|U|^2)$.

This paper combines conflict region with tolerance relation by reference [7-10], and gives a new algorithm for attribution reduction. Finally this algorithm is proved to be correct and efficient.

2   The Basic Concepts of Rough Set

The basic concepts, notations and results of rough sets as well as their extensions are briefly reviewed [12-15].

Definition 1. A decision table could be defined as $S = (U, A, V, f)$ where U is a finite non-empty which is represented the set of objects, $A = C \cup D$ is the attribute set, subset C is called condition attribute and subset D is called decision attribute, $V = \bigcup_{r \in A} V_r$ is the set of attribute values, $f : U \times R \to V$ is an information function, which offers an attribute value to each attribute of each object, if $\forall r \in C \bigcup D, \ x \in U$, then $f(x, r) \in V_r$.

If there exists at least $a \in C$ and V_a contains uncertain value (marked $f(x, a) = *$), then the decision table is called incomplete decision table.

Definition 2.[3, 16]. Suppose T is a binary relation of Incomplete Information System $S = (U, A, \{V_a, a \in A\}, \{f_a, a \in A\})$, and T is defined as follow: $\forall x, y \in U$, $\forall B \in R$, $T(x, y) \Leftrightarrow (\forall b \in B)((b(x) = b(y)) \vee (b(x) = *) \vee (b(y) = *))$.

Definition 3 [3]. $T_B(x)$ is the tolerance class of object x which denotes that object set meets the tolerance relation with x on condition attribute set B: $T_B(x) = \{y \mid y \in U \wedge T(x, y)\}$ .The upper and lower approximations of X with regard to B under the characteristic relation are $X^B = \{x \mid x \in U \wedge I_B(x) \cap X \neq \varnothing\}$, $X_B = \{x \mid x \in U \wedge I_B(x) \subseteq X\}$.

Definition 4. Let $S = (U, A, V, f)$ be an incomplete decision table, $A = C \bigcup D$. If $Q \subseteq U, P \subseteq C$, the positive region of Q with respect to P is defined as follow: $POS_P(Q) = \bigcup\{x \mid T_P(x) \subseteq Y\}$, $Y \in Q/D$ .if $P = C$, $Q = U$,then $POS_C(D) = \bigcup\{x \mid T_P(x) \subseteq D_i\}$ $D_i \in U / D$.

Definition 5. Let $S = (U, A, V, f)$ be an incomplete decision table, $A = C \bigcup D$, $P \subseteq Q \subset C$, $U / P = \{T_P(x_1), T_P(x_2), \cdots T_P(x_{|U|})\}, U / Q = \{T_Q(x_1), T_Q(x_2), \ldots T_Q(x_{|U|})\}$, $\forall T_P(x_i) \in U / P \Rightarrow \exists T_Q(x_i) \in U / Q$,then $T_Q(x_i) \subseteq T_P(x_i)$

Definition 6. Let $S = (U, A, V, f)$ be an incomplete decision table, $A = C \cup D$, $\forall a \in B \subseteq C$, if $POS_B(D) = POS_{B-\{a\}}(D)$, a is unnecessary for B relative to D, Otherwise a is necessary for B. if arbitrary element of B is unnecessary, we call that B is independent with respect to D.

3   The Algorithm of Computing Conflict Region

In the section, we give the definition of conflict at first, and then we design an algorithm of computing conflict region in incomplete decision table [16-20].

As the object of study incomplete decision table, with the study of the notion of conflict region, the definition of attribution reduction based on conflict region in incomplete decision table is provided.

Many scholars design several of algorithms for attribution reduction in complete decision table [21-23]. In practical applications due to the measurement error of the data, restrictions of knowledge acquisition and other various reasons, there will be some default values in the decision table. So the decision table that we often have to deal with is incomplete.

In general, the tolerance class algorithm of object concentrated two comparison, compare them in each attribute property set whether or not to meet the definition, if meet the tolerance class, belong to the same tolerance class; or to the object of every object, according to whether the value judgment of its property belongs to the tolerance existing class.

Definition 7. Let $S = (U, A, V, f)$ be an incomplete decision table, $A = C \cup D$, $B \subseteq C$ .conflict region $Conset(B)$ is defined as: $Conset(B) = \{x_i | x_i \in U, \exists x_m, x_n (x_m, x_n) \in T_B(x_i) \wedge f(x_m, D) \neq f(x_n, D)\}$.

Algorithm $0^{[8]}$.Finding tolerance class $T_P(x)$

a) Input: An incomplete decision table $S = (U, C, D, V, f)$, $U = \{x_1, x_2, \cdots x_n\}$, $c_i \in C$;

Output: $T_{c_i}(x)$

Step1. for $a_i \in C$ statistics maximum value、 minimum value and if there has default value("*") of $f(x_j, c_i)$ $(i=1, 2, \cdots, |U|)$,each mark M_i, M_i, c_i^* ;// $c_i^* = 1$ note exist default value, $c_i^* = 0$ note not exist default value;

Step2:distribution:: if $c_i^* = 1$ note exist default value, establish Empty queue of $A = M_i - m_i + 2$, or establish Empty queue of $A = M_i - m_i + 1$, and each mark the $front_k$ and end_k ($k = 0, 1, 2, \cdots, M_i - m_i$ or $M_i - m_i + 1$) link the front point and the rear point of the k'th queue, $front_*$ and end_* each link the front point and the rear point of the last queue ("*"queue) when $c_i^* = 1$ .we can put the element of the list into the $f(x, a_i)$-m_i queue or put into the last queue("*"queue);

Step3.:collection:the header of list link the first not empty queue , and then mortify the each rear point of the not empty queue , link the next front point of the not empty queue , finally make the all queue one list;

Step4. Suppose the object sequence of the list from the step3 note $x_1', x_2', \cdots, x_n'$;

$t=1$; $E_1 = \{x_1'\}$;

$for(j = 2; j <= n; j++)$

$if\, (f(x_j', a_i) = f(x_{j-1}', a_i))$

$E_t = E_t \cup \{x_j'\}$;

$else\{t=t+1; E_t = E_t \cup \{x_j'\};\}$

Step5. $if\,(c_i^* = 1)$

$for(v = 1; v < t; v++)$

$\{T_v = E_v \cup E_t$;

$T_t = E_t \cup E_v\,(v=1,2,\cdots,t\text{-}1);\}$

$else$

$T_v = E_v$;

The algorithm completes the first time division based on one of the attributes, The algorithm learns from the thought of equivalence partitioning which work in complete decision table.

b) Algorithm of positive region

Step1. $U_{pos} = \varnothing, U_{neg} = \varnothing, T_{C \cup \{c_i\}}$;

Step2.for (j=1; j<r; j++)

$if\,\, (f(x, c_j) = *)\,\, T_{C \cup \{c_j\}}(x) = T_C(x)$;

else

  $\forall y \in T_C(x)$;

$if\,(f(y, c_j) == * \vee f(y, c_j) == f(x, c_j))$

    $T_{C \cup \{c_j\}}(x) = T_{C \cup \{c_j\}}(x) \cup \{y\}$;

Step3.if j>r,   then goto Step4

else $C = C \cup \{c_j\}$, goto Step2;

Step4. Output $T_C(x)$.

By adding a partition condition number, each equivalence class made corresponding change. At last, all conditions are joined the attribution reduction set, and we get the final equivalence class.

As the object of study incomplete decision table, with the study of the notion of conflict region, the definition of attribution reduction based on conflict region in incomplete decision table is provided.

Algorithm 1. Finding the conflict region $Conset(P)$ with respect to P.

Input： An incomplete decision table $S=(U,A,V,f)$, $A=C\cup D$, $\emptyset\neq P\subseteq C$;

Output: $Conset(P)$;

Step 1. Initialize $Conset(P)=\emptyset$;

Step 2. Use the algorithm 1 and 2 from reference [8] to find the tolerance class $T_P(x)$

Step 3. $for(i=1; i\leq |U|; i++)$

$\quad if\,(|T_P(x_i)|>1)$

$if\,(\forall x_m, x_n \in T_P(x_i)\,,\,f(x_m,D)\neq f(x_n,D))$;

$Conset(P)=Conset(P)\cup\{x_i\}$;

Step 4. Output $Conset(P)$;

4   New Attribute Reduction Algorithm

In the section, we will introduce how to compute attribute reduction based on conflict region, and prove that it is equivalent to the algorithm based on positive region.

Theorem 1. Let $S=(U,A,V,f)$ be an incomplete decision table, $A=C\cup D$, $P\subseteq Q\subseteq C$, then $\left|conset(P)\right|\geq\left|conset(Q)\right|$, $\left|conset(P)\right|$ is the number of set $conset(P)$.

Proof. We just need to proof $conset(Q)\subseteq conset(P)$, $\forall x_i\in conset(Q)$, we know that there are two objects decision values are not equal in $T_Q(x_i)$, then $\exists x_i\in T_P(x_i)$ and $T_Q(x_i)\subseteq T_P(x_i)$ based on definition 5, so there are two objects decision values are not equal in $T_P(x_i)$. Due to the arbitrariness of the x_i, we know the result is true.

Theorem 2. Let $S=(U,C\cup D,V,f)$ be an incomplete decision table, $R\subseteq C$, R is the attribute reduction of S, if and only if $\left|ConSet(R)\right|=\left|ConSet(C)\right|$.

Proof. We just need to prove $\left|ConSet(R)\right|=\left|ConSet(C)\right|\Leftrightarrow POS_R(D)=POS_C(D)$. First we prove $POS_R(D)=POS_C(D)\Rightarrow\left|ConSet(R)\right|=\left|ConSet(C)\right|$. $POS_R(D)=POS_C(D)$ We can know $\left|POS_R(D)\right|=\left|POS_C(D)\right|$, and $\left|U-POS_R(D)\right|=\left|U-POS_C(D)\right|$, so the result is $\left|ConSet(R)\right|=\left|ConSet(C)\right|$.

Second we prove $\left|ConSet(R)\right|=\left|ConSet(C)\right|\Rightarrow POS_R(D)=POS_C(D)$, Due to $R\subseteq C$, then $POS_R(D)\subseteq POS_C(D)$ if $POS_R(D)\supseteq POS_C(D)$ is not established, then $POS_R(D)\subset POS_C(D)$. so $\exists x_0\in U$, $x_0\in POS_C(D)$ and $x_0\notin POS_R(D)$. then we have $x_0\in U-POS_R(D)$, but $x_0\notin U-POS_C(D)$, so $\left|U-POS_R(D)\right|\neq\left|U-POS_C(D)\right|$.

However $POS_R(D) \subset POS_C(D)$, thus $\left| POS_R(D) \right| < \left| POS_C(D) \right|$, so $\left| ConSet(R) \right| > \left| ConSet(C) \right|$.
So $POS_R(D) \supseteq POS_C(D)$ is true.

Altogether the result is true.

Theorem 3. The attribute reduction based on conflict region is equivalent to the attribute reduction based on positive region.

Proof. The theorem 3 is easy to be proved true based on the theorem 2.so we don't prove anymore.

Definition 8. Let $S = (U, A, V, f)$ be an incomplete decision table, $A = C \bigcup D$,
$R \subset C$, $a \in C - R$,the significance of attribute a is defined as $Sig(a, R, D) = \left| ConSet(R) \right|$
$- \left| ConSet(R \cup \{a\}) \right|$.

From definition 8 we can see that the important degree of attribute a is bigger and the conflict region reject can be separate faster. So the algorithm can speed up the convergence.

Algorithm 2. attribute reduction algorithm based on conflict region
Input ： An incomplete decision table $S = (U, A, V, f)$, $A = C \cup D$, $U = \{x_1, x_2, \cdots x_n\}$,
$C = \{c_1, c_2, \cdots c_r\}$;
Output ： Attribute reduction R
Step 1.we can compute $conset(C)$ based on Algorithm 1. Initialize $R = \varnothing$,
$conset(\varnothing) = U$;
Step 2. $while(\left| ConSet(R) \right| \neq \left| ConSet(C) \right|)$
$for(i = 1; i <= r; i++)$
compute $conset(R \cup c_i)$ and choose the c_i of the largest of $sig(c_i, R, D)$;
$R = R \cup c_i$;
Update $conset(R)$;
Step 3. $for(i = 1; i <= \left| R \right|; i++)$
$R' = R - c_i$;
$if(\left| conset(R') \right| = \left| conset(R) \right|)$
$R = R'$;
Step 4.output R;

In Algorithm 2, based on the reference [8], we know the time complexity of Step 1 is $O(\left| K \right| \left| C \right| \left| U \right|)$ $K = \max\{\left| T_C(x_i) \right|, x_i \in U\}$. In the worst case, Step 2 need circle $\left| C \right|$ times, so its time complexity is $O(\left| K \right| \left| C \right|^2 \left| U \right|)$. Step 3 is same to Step 2. Therefore, the total time complexity of algorithm 2 is $O(|K||C|^2|U|)$, $K = \max\{\left| T_C(x_i) \right|, x_i \in U\}$.

5 Case Analysis

We use an example to illustrate that the algorithm, and the example (table 1) is as follow:

Table 1 is a incomplete decision table, in order to facilitate the description of the algorithm steps, the data set we used has only 8 objects. However the relationship between these 8 objects covers almost the entire situation. ("*"is marked as the default value). In general, the tolerance class algorithm of object concentrated two comparison, compare them in each attribute property set whether or not to meet the definition, if meet the tolerance class, belong to the same tolerance class; or to the object of every object, according to whether the value judgment of its property belongs to the tolerance existing class. Through calculation examples, learn how to improve the efficiency of classification. As the object of study incomplete decision table, with the study of the notion of conflict region, the definition of attribution reduction based on conflict region in incomplete decision table is provided.

Table 1. Incomplete decision table

U	a1	a2	a3	a4	a5	D
x1	1	1	2	1	*	1
x2	2	*	2	1	1	1
x3	*	*	1	2	2	2
x4	1	*	2	2	1	1
x5	*	*	2	2	1	3
x6	2	1	2	*	1	1
x7	1	1	2	1	2	1
x8	2	*	2	1	*	1

Step 1. we get $conset(C) = \{x_4, x_5, x_6\}$ based on Algorithm 1.

Step 2. Due to $R = \varnothing$, $conset(\varnothing) = U$, then $conset(R) = U$, $|ConSet(R)| \neq |ConSet(C)|$ .So go into the while circle, in the first circle , add attribution a_1 $conset(a_1) = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$, $sig(a_1, R, D) = 0$; add attribution a_2 $conset(a_2) = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$, $sig(a_2, R, D) = 0$; add attribution a_3 $conset(a_3) = \{x_1, x_2, x_4, x_5, x_6, x_7, x_8\}$, $sig(a_3, R, D) = 1$; $conset(a_4) = \{x_3, x_4, x_5, x_6\}$, $sig(a_4, R, D) = 4$; we choose attribute a_4 , $R = \{a_4\}$. because $|ConSet(R)| \neq |ConSet(C)|$, we continue to increase attribute. $conset(a_4, a_1) = \{x_3, x_4, x_5, x_6\}$, $sig(a_1, R, D) = 0$; $conset(a_4, a_2) = \{x_3, x_4, x_5, x_6\}$, $sig(a_2, R, D) = 0$; $conset(a_4, a_3) = \{x_4, x_5, x_6\}$, $sig(a_3, R, D) = 1$; we choose attribute a_3 , $R = \{a_4, a_3\}$ .because of $|ConSet(R)| = |ConSet(C)|$,and then the Algorithm is end.

Step 3 attribution a_4, a_3 are calculated can not be deleted, and $R = \{a_4, a_3\}$ is the result.

After verification the result is same to the algorithm based on the positive region. It is proved that the attribute reduction is equivalent to the attribute reduction based on positive region.

6   Experimental Comparison

We use 6 data sets in the UCI database as the experimental object, and then compare with the algorithm in the reference [10]. The running time of reference [10] is marked as T1, and the running time of this paper is marked as T2. The result is as follows. In order to enhance the reliability of experimental results, the average time of 5 experiments is taken as the final time in this paper.

As the object of study incomplete decision table, with the study of the notion of conflict region, the definition of attribution reduction based on conflict region in incomplete decision table is provided. In the follow table we just give the name of data sets, the final attribute reduction and the running time of the two algorithms. For rigorous this paper has described the detailed experiment environment. The detailed data is as follows.

In this experiment, Computer hardware configuration is as follows: Intel Pentium(R) D 3.20 GHZ, memory: 1G, Development platform: VS2008. The detailed experiment data is as follows:

Table 2. Experimental result

Data set	Attribute reduction	T1(ms)	T2(ms)
Credit	2,3,8,6	14271.30	10370.20
Car	2,1,4,6,5,3	40230.62	12630.55
Hepatitis	1,16,6,3,5	1840.365	735.382
Soybean-large	17,7,16,12,1,22,6,15,8,35,4,9	13570.75	10260.65
Vote	9,4,3,13,1,2,11,10,12,16,15	3190.625	2169.535
win	2	510.125	387.532

The second column is described as the number of the attribution in the database. These data sets include complete and incomplete. To verify the efficiency of the algorithm we use the different kinds of data sets. As can be watched from the experimental data in the table, in all experiments the running time of this paper are faster than it in reference [10]. Based on the different experimental environment, there are some partial data that are not same with reference [10]; however, the result does not affect the efficiency of the algorithm. This algorithm is proved to be correct and efficient.

The running time of the program is almost always not the same, this paper repeat testing several times and method the mean value to ensure the reliability of data. In order to get accurate data, this paper use MS as a unit, and retained two decimal

places. At last, we proved that the experimental results demonstrate that the algorithm is effective.

7   Conclusion

Attribute reduction is one of the important research content of rough set theory. Equivalence partitioning is one of the difficulties. In general, the tolerance class algorithm of object concentrated two comparison, compare them in each attribute property set whether or not to meet the definition, if meet the tolerance class, belong to the same tolerance class; or to the object of every object, according to whether the value judgment of its property belongs to the tolerance existing class. As the object of study incomplete decision table, with the study of the notion of conflict region, the definition of attribution reduction based on conflict region in incomplete decision table is provided.

These paper researches some attribute reduction algorithms in incomplete decision table, and find their time complexity are not better. Then a new attribute reduction algorithm is designed which based on conflict region, its time complexity is $O(|K||C|^2|U|)$, $K = \max\{|T_C(x_i)|, x_i \in U\}$.

then it is proved that the attribute reduction is equivalent to the attribute reduction based on positive region. At last, an example is used to prove the algorithm is correct and efficient.

References

1. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information Science 117, 3–37 (2007)
2. Pawlak, Z., Wong, S.K.M., Ziarko, W.: Rough sets: probabilistic versus deterministic approach. Computational Intelligence 29, 81–95 (1988)
3. Hu, F., Huang, H., Wang, G.Y.: Granular Computing in Incomplete Information Systems. Mini-Micro Systems 26, 1335–1339 (2005)
4. He, W., Liu, C.Y., Zhao, J.: An algorithm of Attributes Reduction in Incomplete Information System. Computer Science 31, 117–119 (2004)
5. Shu, W.H., Xu, Z.Y., Qian, W.B.: Attribution reduction algorithm based on discernibility matrix of incomplete decision table. Computer Engineering and Applications 47, 105–110 (2011)
6. Huang, B., Zhou, X.Z., Zhang, R.R.: Attribution reduction algorithm based on information content in incomplete information Systems. Systems Engineering Theory and Practice 25, 55–60 (2005)
7. Ge, H., Li, L.S., Yang, C.J.: An Efficient Attribute Reduction Algorithm Based on Conflict Region. Chinese Journal of Computers 35, 342–350 (2012)

8. Shu, W.H., Xu, Z.Y., Qian, W.B.: Quick Attribution Reduction Algorithm Based on Incomplete Decision Table. Journal of Chinese Computer Systems 32, 1867–1871 (2011)
9. Wong, S.K.M., Ziarko, W.: Optimal decision rules in decision table. Bulletin of Polish Academy of Sciences 33, 693–696 (1985)
10. Wang, W., Xu, Z.Y., Li, X.Y.: Attribute reduction algorithm based on object matrix in incomplete decision table. Computer Science 39, 201–204 (2012)
11. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences 11, 341–356 (1982)
12. Skowron, A., Rauszer, C.: The Discernibility Functions Matrics and Functions in Information Systems. In: Slowinski, R. (ed.) Intelligent Decision Support - Handbook of Applications and Advances of the Rough Sets Theory, pp. 331–362. Kluwer Academic Publisher, Dordrecht (1992)
13. Hu, X.H., Cercone, N.: Learning in Relational Database: A Rough Set Approach. International Journal of Computional Intelligence 11(2), 323–338 (1995)
14. Nguyen, H.S., Nguyen, S.H.: Some Efficient Algorithms for Rough Set Methods. In: The Sixth International Conference, Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 1996), Granada, Spain, July 1-5, vol. 2, pp. 1451–1456 (1996)
15. Wang, G.Y., Yu, H., Yang, D.C.: Decision Table Reduction Based on Conditional Information Entropy. Chinese Journal of Computers 25(7), 759–766 (2002)
16. Liu, S.H., Cheng, Q.J., Shi, Z.Z.: A New Method for Fast Computing Positve Region. Journal of Computer Research and Development 40(5), 637–642 (2003)
17. Wang, J., Wang, J.: Reduction Algorithms Based on Discernibility Matrix: the Ordered Attributed Method. Journal of Computer Science and Technology 11(6), 489–504 (2001)
18. Zhao, M., Wang, J.: The Data Description Based on Reduct. PhD Thesis, Institute of Automation, Chinese Academy of Sciences, Beijing, China (2004) (in Chinese)
19. Moshkov, M.J., Piliszczuk, M., Zielosko, B.: On Partial Covers, Reducts and Decision Rules with Weights. In: Peters, J.F., Skowron, A., Düntsch, I., Grzymała-Busse, J.W., Orłowska, E., Polkowski, L. (eds.) Transactions on Rough Sets VI. LNCS, vol. 4374, pp. 211–246. Springer, Heidelberg (2007)
20. Qin, Z.R., Wu, Y., Wang, G.Y.: A Partition Algorithm for Huge Data Sets Based on Rough Set. Pattern Recognition and Artificial Intelligence 19(2), 249–256 (2006)
21. Hu, F., Wang, G.Y.: Analysis of the Complexity of Quick Sort for Two Dimension Table. Chinese Journal of Computers 30(6), 963–968 (2007) (in Chinese)
22. Hu, F., Wang, G., Xia, Y.: Attribute Core Computation Based on Divide and Conquer Method. In: Kryszkiewicz, M., Peters, J.F., Rybiński, H., Skowron, A. (eds.) RSEISP 2007. LNCS (LNAI), vol. 4585, pp. 310–319. Springer, Heidelberg (2007)
23. Wang, G.Y.: Rough Set Theory and Knowledge Acquisition. Xi'an Jiaotong University Press, Xi'an (2001) (in Chinese)

Study on Economic Development
Based on Factor and Cluster Analysis

Jingfang Guo, Haiping Li, Heju Yang, and Rongxin Wang

Collage of Science, Hebei University of Science and Technology,
Shijiazhuang Hebei China 050018
lishuxue@126.com

Abstract. In this paper, we use factor and cluster analysis of multivariate statistical analysis to study the economic development of 31 provinces and cities in china in 2008. From so many economic development data, we find three factors which can explain the comprehensive economic development and then classify, compare and comprehensive evaluation. At last put forward rational proposal in view of current social policy.

Keywords: economic development, factor analysis, cluster analysis.

1   Introduction

Since the reform and opening up, China's economic develops rapidly. Many factors lead to the unbalance of the economic of 31 provinces and cities, so how to analysis and extract the main factors which affect the regional social economic development should be studied.

Reference [1-6] study the regional economic development by the factor analysis or cluster analysis. Reference [7-9] study the comprehensive economic strength of each province of China. This paper will also study the comprehensive economic development of each province in China. In this paper, we take 11 indexes of economics in 31 provinces and cities of China as sample，erects an index system to measure the economic power of provinces and cities and compare them. With the help of SPSS, we uses the factor analysis to try to find the underlying, unobservable, random quantities among them, and then use cluster analysis which gives the features of urbanization and the relative superiority of economic development in these different places.

2   Factor Analysis

This research data comes from "China statistical yearbook 2008" [10], according to the economic development of China's 31 provinces and cities, we select 11 indexes to comprehensively evaluate regional economic development at present. Reflecting regional economic efficiency index GDP x1 (yuan). Reflecting macroeconomic aspects indexes are as follows: total financial income x2 (million yuan); total fiscal expenditure x3 (million yuan); total fixed asset investment x4(100 million yuan).

Reflecting resident's income indexes are as follows: the average wage for urban workers x5(yuan). Rural residents income:x6(yuan). Reflecting consumer indexes are as follows: the consumer index x7(previous year=100); the retail price index x8(previous year=100); total retail sales of social consumer goods x9(100 million yuan). Reflecting residents' leisure life indexes are as follows: international tourism foreign exchange earnings x10(billion); the number of private car every million people x11(vehicles).

   This research use SPSS15[11,12]. In order to overcome the influence of the statistical analysis by different dimensions, dimensionless processing of the original data has been done before factor analysis, and the KMO and Bartlett test is shown in table 1.

Table 1. KMO and Bartlett Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.697
Bartlett's Test of Sphericity	Approx. Chi-Square	447.306
	df	55
	Sig.	.000

2.1   Extract Impact Factors by Principal Component Analysis

Factor analysis of the initial results is shown in table 2.

Table 2. Factor analysis initial results

	Initial	extraction
X1	1.000	0.901
X2	1.000	0.982
X3	1.000	0.915
X4	1.000	0.883
X5	1.000	0.859
X6	1.000	0.882
X7	1.000	0.948
X8	1.000	0.941
X9	1.000	0.960
X10	1.000	0.812
X11	1.000	0.752

   The second column is the common degree of the variable according to the initial solution by factor analysis. The third column is the common degree of the variable according to final solution by factor analysis.We extracted the common factor whose minimum eigenvalue is 1, that is the first three. The common degree of variable 1 (per capita GDP) is 0.906, which can be understood as the variance is 90.6% variable 1 (per capita GDP) explained by three public factors, the other variables explanation are similar.

2.2　Total Variance

Explained variance is shown in table 3

Table 3. Total variance explained

Compone	Total	%of Variance	Cumulative %	Total	% Variance	Cumulative %	Tatal	%of Variance	Cumulative %
1	6.339	57.627	57.627	6.339	57.627	57.627	4.145	37.677	37.677
2	2.246	20.420	78.048	2.246	20.420	78.048	3.726	33.876	71.677
3	1.250	11.363	89.411	1.250	11.363	89.411	1.964	17.857	89.411
4	0.370	3.361	92.772						
5	0.314	2.855	95.626						
6	0.242	0.202	97.828						
7	0.107	0.970	98.798						
8	0.059	0.533	99.331						
9	0.041	0.369	99.700						
10	0.026	0.233	99.033						
11	0.007	0.067	100						

The second column is the variance contribution of the factor variables. The third column is the variance contribution rate of the factor variables.The fourth column is the cumulative variance contribution rate of the factor variables. The fifth to the seventh columns are extracted three factors from the initial solution in accordance with certain standards. We can see the 89.411% of the total variance can be explained by the first three common factors. The eighth to tenth columns are variance contribution value, variance contribution rate and cumulative variance contribution which is after factor rotation.

2.3　The Scree Plot of the Eigenvalue

The scree plot of the eigenvalue is shown in figure 1.
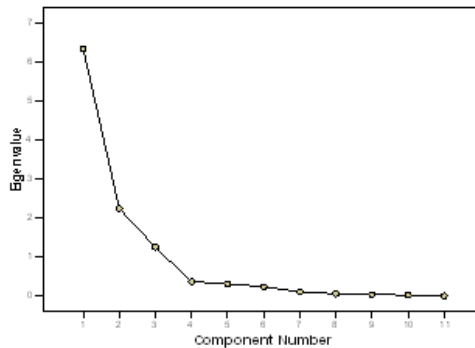
Fig. 1. Scree plot

2.4   The Component Matrix Before Rotation

The component matrix before rotation is shown in table 4.

Table 4. Component matrix before rotation

	Component		
	1	2	3
Zscore(x2)	.934	-.229	.239
Zscore(x6)	.855	.385	.057
Zscore(x10)	.841	.015	.324
Zscore(x1)	.826	.453	.112
Zscore(x9)	.819	-.514	.157
Zscore(x3)	.807	-.500	.118
Zscore(x4)	.678	-.651	-.012
Zscore(x11)	.625	.600	.031
Zscore(x5)	.567	.726	.102
Zscore(x8)	-.623	-.070	.740
Zscore(x7)	-.681	.099	.688

Extraction Method: Principal Component Analys
a. 3 components extracted.

Rotated component matrix is shown in table 5.

Table 5. Rotated component matrix

	Component		
	1	2	3
Zscore(x9)	.953	.145	-.177
Zscore(x3)	.923	.140	-.209
Zscore(x4)	.892	-.083	-.285
Zscore(x2)	.873	.450	-.135
Zscore(x10)	.675	.597	-.014
Zscore(x5)	-.045	.921	-.095
Zscore(x1)	.316	.874	-.191
Zscore(x11)	.056	.845	-.187
Zscore(x6)	.363	.828	-.255
Zscore(x8)	-.165	-.270	.917
Zscore(x7)	-.330	-.189	.896

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 5 iterations.

2.5   Factor Analysis after Rotation

We use the max variance method to rotate the factor loading matrix, the results is shown in table 6. In this table, the first common factor reflect the importance of variable X9,X3,X4,X2,X10, and the second common factor reflect the importance of variable X5,X1,X11,X6, and the third common factor reflect the importance of variable X8 and X7.

Table 6.   Rotated factor loading matrix

	1	2	3
X9	0.953	0.145	-0.177
X3	0.923	0.140	-0.209
X4	0.892	-0.083	-0.285
X2	0.873	0.450	-0.135
X10	0.675	0.597	-0.014
X5	-0.045	0.921	-0.095
X1	0.316	0.874	-0.191
X11	0.056	0.845	-0.187
X6	0.363	0.828	-0.255
X8	-0.165	-0.270	0.917
X7	-0.330	-0.189	0.896

In the first factors:X9,X3,X4,X2,X10 reflect the macroeconomic, named as "total factor"; the second factors:X5,X1,X11,X6 reflect the degree of economic development from the perspective of the individual, named as" average factor". The third factors: X8 and X7, these two indexes have larger load and they mainly reflects the price index, named "the price factor".

2.6   The Component Plot in Rotated Space

The component plot in rotated space is shown in figure 2.
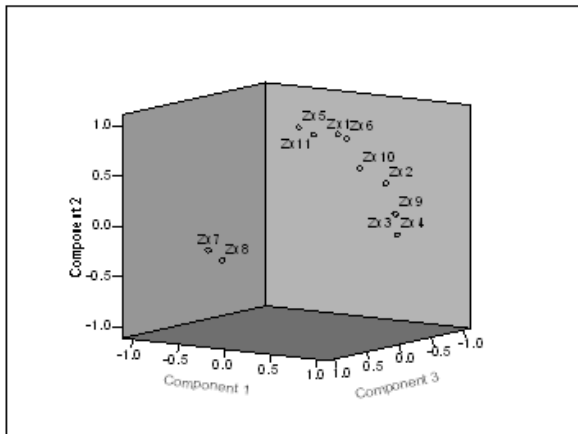
Fig. 2. Component plot in rotated space

2.7   The Provinces in Our Country Comprehensive Evaluation Score

The component score matrix is shown in table 7.

Table 7. Component score matrix

	Component		
	1	2	3
Zscore(x1)	-.013	.253	.041
Zscore(x2)	.227	.055	.119
Zscore(x3)	.262	-.070	.033
Zscore(x4)	.260	-.157	-.058
Zscore(x5)	-.123	.318	.053
Zscore(x6)	-.004	.223	-.003
Zscore(x7)	.064	.090	.552
Zscore(x8)	.131	.047	.585
Zscore(x9)	.277	-.067	.061
Zscore(x10)	.167	.143	.191
Zscore(x11)	-.097	.269	-.006

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
Component Scores.

From the upper matrix, we can make the factor score function as follows:

$$\begin{cases} F_1 = -0.013x_1 + 0.227x_2 + ... - 0.097x_{11} \\ F_2 = 0.253x_1 + 0.055x_2 + ... + 0.269x_{11} \\ F_3 = 0.041x_1 + 0.119x_2 + ... - 0.006x_{11} \end{cases}.$$

According to score function, we calculate the total score zF and make the ranking based on the provinces and cities of the composite score, the results is shown in table 8.

Table 8. Factor score function

place	FAC1_	FAC2_	FAC3	zF	rank
Beijing	-0.46526	3.51304	-0.57292	1.02	3
Tianjin	-1.09952	1.65545	-0.81551	0	11
Hebei	0.43813	-0.42256	0.03257	0.03	9
Shanxi	-0.29036	-0.10766	0.64839	-0.03	13
Neimonggu	-0.4519	-0.08827	-1.00875	-0.43	25
Liaoning	0.59808	-0.30323	-1.01174	-0.06	15
Jilin	-0.45332	-0.43126	-0.77988	-0.51	27
Heilongjiang	-0.28973	-0.4253	-0.60411	-0.4	24
Shanghai	0.50213	2.7009	0.30301	1.3	2
Jiangsu	2.21622	0.10664	-0.34275	0.91	4
Zhejiang	0.99606	0.91466	-0.11109	0.74	5
Anhui	0.03524	-0.68478	-0.20001	-0.28	20

Table 8. (*continued*)

Fujian	-0.14612	0.00067	-0.88405	-0.24	18
Jiangxi	-0.33883	-0.75815	-0.42591	-0.52	28
Shandong	1.7575	-0.55215	-0.80173	0.37	7
Henan	0.94505	-0.72047	0.71738	0.27	8
Hubei	0.18016	-0.64112	-0.1027	-0.19	16
Hunan	0.04161	-0.63464	-0.49979	-0.32	21
Guangdong	3.02184	0.72074	0.96577	1.74	1
Guangxi	-0.1601	-0.43508	1.04913	-0.02	12
Hainan	-1.06474	-0.2833	0.14745	-0.53	29
Chongqing	-0.63571	-0.50076	-1.00069	-0.66	30
Sichuan	0.5275	-0.92072	-1.0196	-0.33	22
Guizhou	-0.60734	-0.60663	0.66363	-0.35	23
Yunnan	-0.39677	-0.60963	-0.46927	-0.49	26
Xizang	-1.86896	0.35219	-1.44038	-0.94	31
Shanxi	-0.21942	-0.50318	0.14584	-0.25	19
Gansu	-0.55251	-0.54897	1.24632	-0.19	16
Qinghai	-0.73335	0.24484	3.15783	0.41	6
Ningxia	-1.01698	0.17244	1.57706	-0.05	14
Xinjiang	-0.46862	-0.20369	1.43652	0.01	10

2.8   Clustering Result of Provinces and Cities

In this paper, we study FAC1, FAC2 and FAC3 by cluster analysis (also called hierarchical cluster). The cluster result is output by a tree diagram and shown in Figure3.
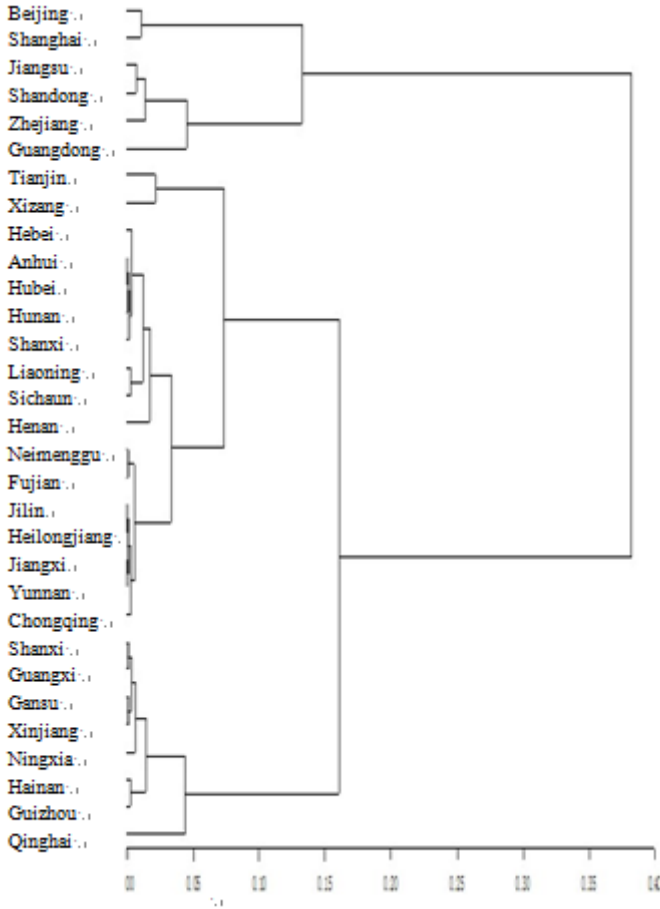
Fig. 3. The clustering results

3   Conclusion

This paper take 11 indexes of economics of 31 provinces and cities in china, and build an index system to measure the economic power of provinces and cities and compare them. By the factor analysis to try to find the underlying, unobservable, random quantities among them, and by cluster analysis gives the features of urbanization and the reasons that they are relative superiority in economic development.

From synthesized, we can see that the provinces have stronger competitive ability are: Guangdong, Shanghai, Beijing, Jiangsu, Zhejiang, Qinghai, Shandong, Henan, Hebei, and Xinjiang. In medium positions of the competitive are: Tianjin, Guangxi, Shanxi, Ningxia, Liaoning, Gansu, Hubei, Fujian, Shanxi, Anhui. Other provinces and cities are in the final. In recent years, the development of Guangdong province has exceeded Shanghai. Beijing still remains its previous position. It is not difficult to see that the southeast of china: Jiangsu, Zhejiang, Shandong and other coastal cities have

been more vigorous development in recent years. In addition, we can not neglect that Qinghai and Xinjiang, have a great development in economic compared with previous. But there are still some western provinces develop slowly relatively. In recent years, government has accelerated the pace of development in the western region, and increased the investment in the west.

From the tree we can see that: Beijing, Shanghai belong to the first class; Guangdong, Zhejiang, Shandong, Jiangsu belong to the second class; Shanxi, Guangxi, Gansu, Xinjiang, Ningxia, Hainan, Guizhou, Qinghai belong to the third class, Other provinces and cities belong to the fourth class.

In conclusion, the economic development is not balance in China, the southeast coastal areas have the strongest competitive, the eastern region following, then the central region, the most weakest is the west. The Midwest is rich in resources, remains to be the development strategic areas, also be China's potential market. Only the mid-west developed, the eastern region's economic development will have more broad market support. Some suggestions to the development of the western region are as follows: The government should pay special attention to its transportation. Some highways and railways should be built. In order to protect the nature, farmers should make farmland revert to forest. Also, measures should be taken to prevent pollution. Western China is rich in natural resources, which can be exploited and make full use of. To meet the need of development, all kinds of experts and skilled workers are to be brought in. As the result of the great project, we believe that central and western regions of China will surely achieve better and faster development in the new century.

References

1. Wu, Q.: The Factor and Cluster analysis in Comparison with Competitive Power of the Cities in Xinjiang. Journal of XinJiang Finance and Economics Institute 1, 67–71 (2005) (in Chinese)
2. Liu, X., et al.: The Regional Economic Development Study by the Factor Analysis in HeNan. Statistics and Decision 5, 119–121 (2010) (in Chinese)
3. Yan, B., Zhang, C.: Analysis evaluation of Hebei provinces economic power by factor analysis. The National Business: Economic Theory Research 10, 17–18 (2007) (in Chinese)
4. Meng, D., Li, B.: Appliction of cluster analysis in regional economy reseach. Hebei Journal of Industrial Science and Technology (2), 116–119 (2012) (in Chinese)
5. Li, P., Wang, H., Chen, Y.: Study on the regional disparities of economic development in Hebei province based on the methods of principal component analysis and cluster analysis. Journal of Science of Teachers' College and University 31(2), 80–84 (2011) (in Chinese)
6. Yuan, Z., Wang, Y., Zhang, J.: Evalution of regional economic competitiveness based on the method of AHP and gray evaluation in Hebei province. Hebei Journal of Industrial Science and Technology (1), 1–4 (2013) (in Chinese)

7. Gan, S.: Comparison and analysis of the comprehensive economic strength of each province of China. Modern Finance & Economics 22(7), 58–62 (2002) (in Chinese)
8. Li, N.: A Comparison of Competitive Powers of Chinese Provinces and Cities Based on Factor Analysis. Journal of Shandong Institute of Commerce and Technology (6), 530–534 (2006) (in Chinese)
9. Qiao, H.: The Multivariate Statistical Analysis on the Economic Development of Thirty One Provinces, Cities and Autonomous Regions in China. Sci-tech Information Development & Economy 21(1), 160–162 (2011) (in Chinese)
10. China Statistical Yearbook 2008. China Statistics Press, Beijing (2009) (in Chinese)
11. Yu, X., Ren, X.S.: Multivariate Statistical Analysis. China Statistics Press, Beijing (1999) (in Chinese)
12. He, X.: Data Statistical Analysis and SPSS Application. People's Post and Telecommunications University Press, Beijing (2003) (in Chinese)

Gray Correlation Model of Enterprise Commercial Credit Risk Assessment

Hao Huang

Guangdong University of Foreign Studies,
No.2, Baiyun Road North, Guangzhou, Guangdong Province, 510420, China
`104342351@qq.com`

Abstract. The modern economy is credit economy. Commercial credit is important for enterprises to gain market share and obtain competitive advantages. Credit risk management mechanism is the key element to make use of commercial credit effectively and expand market sales. In this paper, a commercial credit risk index system is built up with the gray system analysis method. Compared to traditional commercial credit risk evaluation system, it is more comprehensive, systematic and scientific. It solves the problem of incomplete and poor information in credit risk assessment.

Keywords: Gray correlation, Attribute Matrix, Credit Risk Management.

1   Introduction

Commercial credit in the course of commodity trading is the mutual credit on credit sales, prepaid and accelerating the circulation of commodities. It is recognized more and more clearly that strengthening credit risk management is important for enterprise reform and the establishment of a modern enterprise system. Building an advanced credit management system is the key to enhance credit image and credit status, as well as the ability of controlling the credit risk. Therefore, the establishment of the credit risk management mechanism will comprehensively improve the quality of business management and the competitiveness of enterprises at home and abroad. Credit risk management is a professional, technical and comprehensive work. Only by establishing a comprehensive credit risk management scheme can we minimize the credit risk. This article build commercial credit risk evaluation index system on the base of analyzing commercial credit risk, then evaluate the commercial credit risk by the gray correlation model. Hence the commercial credit risk management is a set of risk analysis and risk assessment in one integrated management system.

2   Enterprise Credit Evaluation Index System

With the analysis of the various factors that affect customer credit, the main factors to affect commercial credit risk are decomposed into five aspects consisting of corporate

information credit risk, enterprise basic capability risk, the financial credit risk, the brand credit risk, the trading credit risk and credit management risk. There are 37 indicators in total to examine customer credit.

Corporate Information Credit Risk U_1

Corporate information credit risk refers to credit risk loss due to wrong corporate information such as wrong names and wrong addresses of customers and so on. According to the degree of risk caused by the corporate information, credit risk indicators can be summarized as follows: ① Registered name and address of the enterprise (u_{11}) ②Corporation's legal form and registered capital(u_{12}) ③ Enterprise ownership(u_{13}) ④ Business scope and industry of the enterprise(u_{14}) ⑤Registration date and risk of operations duration (u_{15}) ⑥ Enterprise internal organizational structure and the chief managers(u_{16}).

Enterprise Basic Quality Risk U_2

Enterprise basic quality is the analysis of enterprise basic condition and the comprehensive evaluation of enterprise. Indicators of enterprise basic quality risk are: ① Enterprise system construction (u_{21}), mainly on the Board of Directors (u_{211}),development of corporate strategy (u_{212}),standardized system construction (u_{213}). ②Management efficiency (u_{22}) mainly includes: compliance rate of contracts (u_{221}),achievement of aims (u_{222}) and average time of delivery (u_{223}).The three sub-indicators can better reflect the enterprise condition of achieving objectives in various stages. ③Quality of employees (u_{23}), is mainly to study the quality of corporate executives (u_{231}),technical staff (u_{232})and general employees (u_{233}).Education and work experience are the two aspects to assess these indicators. ④ Asset quality (u_{24})refers to science and technology content of the assets and condition of equipment. Therefore the main indicator we choose is the condition of equipment (u_{241}).

Financial Credit Risk U_3

The customer financial credit is the client's financial capacity and financial structure which reflect the customer's credit status and solvency. In practice, enterprises need to focus on the indicators of the customers financial conditions as the following aspects: ① Customer's profit ability and profit growth (u_{31}), including the following rates: a. Capital Profit Rate (u_{311}) b. Return on Sales (u_{312}) c. Cost-Profit Ratio (u_{313}) d. Return on Assets (u_{314}). ② Customer's capital structure (u_{32}) is the premise of its survival and development, but also one of the reflections of its solvency. Capital structure can be subdivided into the following two sub-indicators: a. Debt Ratio (u_{321}) b. Owner's   Equity Ratio (u_{322}).③Liquidity ratio index (u_{33}) refers to the degree of enterprises' ability of converting properties into cash. There are two indicators: a. Cash Ratio (u_{331}) b. Current Ratio (u_{332}).

Brand Credit Risk U_4

The brand suggests the competitiveness and operating stability of an enterprise. The brand credit includes:①Productive characteristics of the products (u_{41}).It can be decomposed into mode of production (u_{411}), raw materials supply (u_{412}). ②Quality characteristics of the products (u_{42}).Quality is the basis of reputation of the brand. Quality does not only affect the performance of a brand in the market, but also determine its life term. Quality characteristic index of customer products can be categorized into product design (u_{421}), style (u_{422}) and quality (u_{423}). ③ Demand characteristics of the product (u_{43}).It can be decomposed into the market capacity (u_{431}),target customers of the product (u_{432}) and sales trends (u_{433}).④ Competitive characteristics of the products (u_{44}).Market competitiveness is one of the important factors to determine the prospects of the business. It can be further broken down into market share (u_{441}) and industry reputation (u_{442}).

Trading Credit Risk U_5

Trading credit is the credit worthiness of the customer according to the customer's transactions (payment) record. Trading credit risk can be decomposed into three indicators as the fairness of credit transaction contract (u_{51}), the compliance degree of credit transaction (u_{52}) and average days sales outstanding (u_{53}).

3   Methods of Commercial Credit Risk Evaluation

The information incompleteness is precisely the main feature in customer commercial credit risk assessment. In this article, gray system analysis method is applied to evaluate the commercial credit risk. As a combination of cybernetics and mathematics, the gray system theory has developed a set of theories and solutions to deal with information incompleteness.

Attribute Matrix

Set the credit risk evaluation system has n assessment objects. There are m evaluation factors for each assessment object. The attribute value of each assessment object under the respective evaluation factor is $X' = \left(x_{ij}\right)$, where the x'_{ij} is the attribute value of the j th assessment object under the i th evaluation factor.

Normalized Matrix

Variables can be turned into dimensionless by normalization. The attribute index/indicator of an object can be categorized into the following three types:

The bigger the optimal type

$$x_{ij} = \frac{x'_{ij} - \min_i x'_{ij}}{\max_i x'_{ij} - \min_i x'_{ij}} \tag{1}$$

The smaller the optimal type

$$x_{ij} = \frac{\max_i x'_{ij} - x'_{ij}}{\max_i x'_{ij} - \min_i x'_{ij}} \tag{2}$$

Standard numerical type
That is, the closer to a standard value, the better.

$$x_{ij} = 1 - \frac{\left| x'_{ij} - \gamma_i \right|}{\max_i \left| x'_{ij} - \gamma_i \right|} \tag{3}$$

Obviously, $0 \leq x_{ij} \leq 1$. A larger x_{ij} suggests that the evaluation of the j th customer on the i th factor is better; a smaller x'_{ij} suggests that the evaluation of the j th customer on the i th factor is worse.

Determine the Priority Vector G and Secondary Vector B

$$G = (g_1, g_2, \cdots, g_m) = \begin{pmatrix} x_{11} \vee x_{12} \vee \cdots \vee x_{1n}, \vee x_{21} \vee x_{22} \vee \cdots \\ \vee x_{2n}, \cdots, \vee x_{m1} \vee x_{m2} \vee \cdots \vee x_{mn} \end{pmatrix} \tag{4}$$

$$B = (b_1, b_2, \cdots, b_m) = \begin{pmatrix} x_{11} \wedge x_{12} \wedge \cdots \wedge x_{1n}, \wedge x_{21} \wedge x_{22} \wedge \cdots \\ \wedge x_{2n}, \cdots, \wedge x_{m1} \wedge x_{m2} \wedge \cdots \wedge x_{mn} \end{pmatrix} \tag{5}$$

Which $\vee$ and $\wedge$ respectively take big operator and small operator.

Correlation degree analysis

The correlation degree analysis of gray system theory is the quantitative comparison of the systems development trend. This comparison is a comparison of system statistics fitted to a geometric relationship. It is believed that the curves constituted by various statistical data columns are closer, then the variation trends are closer, and correlation degree is greater. Formula (6) and (7) determine the correlation coefficient ξ_i.

The correlation coefficient of vector X_j of the j th customer and priority vector G is:

$$\xi_i(X_j,G) = \frac{\min\limits_{j}\min\limits_{i}\left|x_{ij}-g_i\right| + \rho\max\limits_{j}\min\limits_{i}\left|x_{ij}-g_i\right|}{\left|x_{ij}-g_i\right| + \rho\max\limits_{j}\min\limits_{i}\left|x_{ij}-g_i\right|}$$

$$= \frac{\rho\max\limits_{j}\min\limits_{i}\left|x_{ij}-g_i\right|}{\left|x_{ij}-g_i\right| + \rho\max\limits_{j}\min\limits_{i}\left|x_{ij}-g_i\right|}$$

(6)

The correlation coefficient of vector X_j of the j th customer and secondary vector B is:

$$\xi_i(X_j,B) = \frac{\min\limits_{j}\min\limits_{i}\left|x_{ij}-b_i\right| + \rho\max\limits_{j}\min\limits_{i}\left|x_{ij}-b_i\right|}{\left|x_{ij}-b_i\right| + \rho\max\limits_{j}\min\limits_{i}\left|x_{ij}-b_i\right|}$$

$$= \frac{\rho\max\limits_{j}\min\limits_{i}\left|x_{ij}-b_i\right|}{\left|x_{ij}-b_i\right| + \rho\max\limits_{j}\min\limits_{i}\left|x_{ij}-b_i\right|}$$

(7)

$\rho \in (0, +\infty)$ is called distinguishing coefficient. The smaller is, the greater distinguishing capability it is. Usually, the value of ρ is $[0,1]$. More generally, $\rho = 0.5$ is preferable. The value of ρ can be decided case by case.

Calculate Correlation Degree

The correlation degree of vector X_j of the j th customer and priority vector G is:

$$\gamma(X_j,G) = \sum_{i=1}^{n}\omega_i\xi_i(X_j,G)$$

(8)

The correlation degree of vector X_j of the j th customer and secondary vector B is:

$$\gamma(X_j,B) = \sum_{i=1}^{n}\omega_i\xi_i(X_j,B)$$

(9)

The effect of distinguishing coefficient (ρ) on correlation coefficient (ξ)

Since there is great effect of distinguishing coefficient (ρ) to correlation coefficient (ξ), detailed analysis is shown as below:

$$\xi_i(X_j, G) = \frac{\min\limits_{j} \min\limits_{i} |x_{ij} - g_i| + \rho \max\limits_{j} \min\limits_{i} |x_{ij} - g_i|}{|x_{ij} - g_i| + \rho \max\limits_{j} \min\limits_{i} |x_{ij} - g_i|} \tag{10}$$

When $\min\limits_{j} \min\limits_{i} |x_{ij} - g_i| = |x_{ij} - g_i|$, the maximum of ξ_i is 1. When

$\max\limits_{j} \max\limits_{i} |x_{ij} - g_i| = |x_{ij} - g_i|$, the minimum of ξ_i is $\dfrac{\rho}{1+\rho}$. It can be written as:

$$\frac{\min\limits_{j} \min\limits_{i} |x_{ij} - g_i| + \rho \max\limits_{j} \max\limits_{i} |x_{ij} - g_i|}{\max\limits_{j} \max\limits_{i} |x_{ij} - g_i|(1+\rho)} \geq \frac{\rho}{1+\rho} \tag{11}$$

Therefore, we have:

$$\frac{\rho}{1+\rho} \leq \xi_i \leq 1 \tag{12}$$

If $\rho = 1$, then $\xi_i = 0.5$, and $\xi_i \in [0.5,1]$

If $\rho = 0.1$, then $\xi_i = 0.09$, and $\xi_i \in [0.09,1]$.

As ρ is smaller, distinguishing coefficient is lower; otherwise distinguishing coefficient is higher. Therefore, the value ρ affects the value of ξ, thus affecting the correlation degree. Further analysis of the impact to ξ from ρ is showed as followed.

Set $\xi_i \in \left[\dfrac{\rho}{1+\rho}, 1\right]$, the lower bound of correlation coefficient is $\xi_{min} = \dfrac{\rho_0}{1+\rho_0}$, as shown in figure 3; take the first derivative to ρ_0: $\dfrac{d\xi_i}{d\rho_0} = \dfrac{1}{(1+\rho_0)^2}$, the curve variation is shown in figure 4:
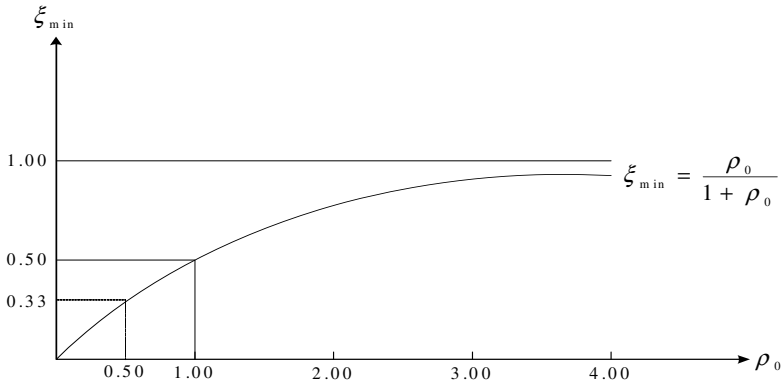
Fig. 3. Relation of ξ and ρ_0
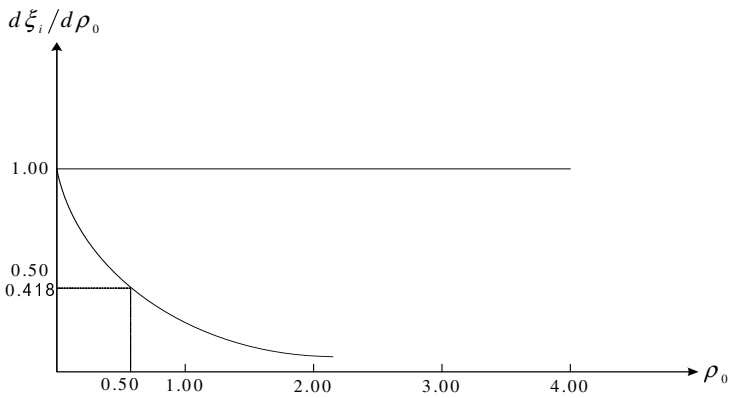
Fig. 4. Relation of ξ and ρ_0

According to figure 3, it suggests:

① ξ_i is a decreasing function of ρ_0, i.e. with the increase of ρ_0, the increase degree is getting smaller;

② When $\rho_0 - +c$, all $\xi_i - 1$, i.e. as the correlation coefficient approaches to a point, correlation degree approaches to a point;

③ The area among $d\xi_i/d\rho_0 = \dfrac{1}{(1 + \rho_0)^2}$, Axis $d\xi_i/d\rho_0$ and Axis ρ_0 can be expressed as Formula (13):

$$S = \int_0^{+\infty} \frac{1}{(1+\rho_0)^2} d\rho_0 = \left[arctg\, \rho_0 \right]_0^{+\infty} = 1 \tag{13}$$

The area among $d\xi_i/d\rho_0 = \dfrac{1}{(1+\rho_0)^2}$, Axis $d\xi_i/d\rho_0$, Axis ρ_0 and $\rho_0 = \tau$ can be expressed as Fomula(14):

$$S_1 = \int_0^\tau \frac{1}{(1+\rho_0)^2} d\rho_0 = arctgx_p \tag{14}$$

$S = 1$ and S1 equals to 0.5, which means $arctgx = 0.5$, then we have $x = 0.5463 \cdot$

When $\rho_0 \div 0.5463$, the decay rate of $d\xi_i/d\rho_0 = \dfrac{1}{(1+\rho_0)^2}$ is the maximum. Then $d\xi_i$ is proportional scaled-down, which suggests that when $\rho_0 \div 0.5463$, resolution is of best quality.

Thus, the correlation coefficient effects to ξ_i can be summarized as follows:

① ρ can adjust the value of ξ and also dominate the interval of ξ ;

② When $\rho \le 0.5463$, it is easier to observe changes in the resolution ratio;

③ When $\rho - +c$, relational analysis cannot be don't, while $\xi_i(k) - 1$ (for all i and j).

Comprehensive Evaluation Model

Assuming that vector X_j of the j th customer subordinates to priority vector G by e_j, and subordinates to secondary vector B by $(1-e_j)$. Expand the rule by least squares method to minimize the objective function:

$$\min\left\{F(u) = \sum_{j=1}^n \left[(1-u_j)\gamma(X_j,G)\right]^2 + \left[u_j\gamma(X_j,B)\right]^2\right\} \tag{15}$$

The optimal solution vector is: $u = (u_1, u_2, \cdots, u_n)$.

Derived from $\dfrac{\partial F(u)}{\partial u_j} = 0$, the gray comprehensive evaluation vector is:

$$u_j = \frac{1}{1 + \left[\dfrac{\gamma(X_j,B)}{\gamma(X_j,G)}\right]} = \frac{1}{1 + \left[\dfrac{\displaystyle\sum_{i=1}^m \omega_i \xi_i(X_j,B)}{\displaystyle\sum_{i=1}^m \omega_i \xi_i(X_j,G)}\right]^2} \qquad j = 1,2,\cdots,n \tag{16}$$

The degree the j th vector X_j subordinate to the priority vector G calculated by the integrated assessment model of gray system reflects the priorities of the program. Various alternatives of customers are prioritized according to the value of u_j, which is considered as gray relational series. The sequence of the gray relational series is the rank of evaluation.

4   Case Study

The profitability indicators of an enterprise are including capital profit margin, return on sales, cost-profit ratio, and return on assets and financial leverage. The major financial risk indicators are as shown in Table 1.

Table 1. 15 Customers' Financial Risk Indicators of an Enterprise

Customers	Capital profit margin (%)	Return on sales (%)	Cost-profit ratio (%)	Return on assets (%)	Financial leverage
1	19.21	9.62	11.55	26.01	2.29
2	15.69	10.71	12.85	17.96	2.05
3	9.26	8.96	10.75	16.14	3.25
4	17.05	14.66	17.29	19.98	1.96
5	21.09	14.29	17.15	27.16	2.12
6	12.19	9.98	11.98	17.62	2.72
7	12.62	16.73	20.01	16.36	1.75
8	15.42	19.55	21.96	21.52	3.01
9	17.56	12.36	14.50	19.59	1.96
10	22.92	17.71	20.27	35.27	4.01
11	20.11	25.66	30.08	23.04	1.97
12	19.91	16.91	19.29	24.59	2.15
13	15.99	14.13	16.92	17.97	1.27
14	19.78	15.59	18.71	26.57	2.67
15	19.75	16.77	20.19	21.25	1.79

The optimal matrixes are : $G = (22.92, 25.66, 30.08, 27.16, 2.50)$,
$B = (6.21, 7.69, 9.21, 8.79, 1.27)$

The correlation coefficient of priority vector G can be determined by gray analysis as shown in Table 2.

Table 2. 15 Customers' Correlation Coefficient of Vector G

Customers	r311	r312	r313	r314	r315
1	0.7110	0.3797	0.3809	0.60712	0.8232
2	0.5580	0.3964	0.3982	0.4538	0.6848
3	0.4006	0.3702	0.3710	0.4293	0.5659
4	0.6086	0.4716	0.4713	0.4845	0.6442
5	0.8330	0.4634	0.4686	0.6379	0.7202
6	0.4597	0.3850	0.3865	0.4490	0.8163
7	0.4699	0.5237	0.5310	0.4321	0.5659
8	0.5490	0.6164	0.5840	0.5109	0.6572
9	06301	0.4247	0.4226	0.4783	0.6442
10	1.0000	0.5525	0.5375	0.9931	0.3931
11	0.7646	1.0000	1.0000	0.5399	0.6485
12	0.7520	0.5287	0.5138	0.5730	0.7364
13	0.5685	0.4599	0.4642	0.4540	0.4429
14	0.7441	0.4936	0.5007	0.6218	0.8519
15	0.7422	0.5248	0.5355	0.5061	0.5794

Table 3. 15 Customers' Correlation Coefficient of Vector B

Customers	r311	r312	r313	r314	r315
1	0.4125	0.8357	0.8297	0.3682	0.6180
2	0.4906	0.7647	0.7580	0.5225	0.6790
3	0.7496	0.8855	0.8810	0.5772	0.4545
4	0.4572	0.5849	0.5852	0.4728	0.7051
5	0.3802	0.7738	0.5895	0.3533	0.6600
6	0.6042	0.8109	0.8045	0.5320	0.5322
7	0.5875	0.5206	0.5135	0.5700	0.7746
8	0.4978	0.4529	0.4721	0.4408	0.4867
9	0.4458	0.6776	0.6831	0.4817	0.7051
10	0.3533	0.4949	0.5076	0.2748	0.3758
11	0.3964	0.3533	0.3533	0.4132	0.7021
12	0.3999	0.5157	0.5308	0.3884	0.6522
13	0.4828	0.6039	0.5966	0.5222	1.0000
14	0.4022	0.5541	0.5454	0.3608	0.5410
15	0.4027	0.5195	0.5094	0.4461	0.7603

The weight of the five risks is set up as $(b_{311}, b_{312}, b_{313}, b_{314}, b_{315}) = (0.25, 0.15, 0.25, 0.15, 0.20)$ by calculation. Gray comprehensive evaluation vector is calculated by the formula (16), the evaluation result is shown in Table 4.

Table 4. 15 Customers' Gray Comprehensive Evaluation Result

Customers	Result	Rank	Customers	Result	Rank	Customers	Result	Rank
1	0.4758	9	6	0.3645	14	11	0.7663	1
2	0.3816	12	7	0.4215	11	12	0.6140	4
3	0.2604	15	8	0.6028	5	13	0.3646	13
4	0.4836	8	9	0.4383	10	14	0.6441	3
5	0.5769	6	10	0.7456	2	15	0.5581	7

Take the result value 0.5 as a decision criterion, enterprise can provide commercial credit sales to the top 7 customers.

5   Conclusion

This article focuses on the analysis of factors that may lead to commercial credit risk. On this basis, a commercial risk evaluation index system is established. Compared to traditional commercial credit risk assessment indicators, the new index system is more comprehensive, scientific and rational. Due to incompleteness of evaluation index data and the uncertainty of information collection, the gray relational analysis method is selected to evaluate the commercial credit risk. This model is more realistic compared with the traditional commercial credit risk assessment model, and high accuracy of the data is not required. It improves the operability and accuracy of the commercial credit risk assessment.

References

1. Wang, H., Wang, G.F.: Capital Adequacy Ratio, Leverage Ratio and Bank Value of Grey Relational Analysis. Co-operative Economy and Science 11, 64–65 (2011)
2. Liu, L., Liu, H.C., Lin, Q.L.: An Improved FMEA Using Fuzzy Evidential Reasoning Approach and Grey Theory. Fuzzy Systems and Mathematics 2, 71–80 (2011)
3. Zhang, M., Zhou, Z.F.: Enterprise Credit Evaluation Model Based on Gray Relation Projection Pursuit. Statistics and Decision 21, 35–37 (2009)
4. Wang, Z.X., et al.: Multi-index Grey Relational Decision-making Based on Cumulative Prospect Theory. Control and Decision 2, 232–236 (2010)
5. Wang, S.W., et al.: Matrix Absolute Grey Relation Degree and its Significance. Journal of Northeast Normal University (Natural Science Edition) 4, 73–78 (2009)
6. Altman, E.: Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. J. Finance, 589–609 (1968)
7. Thorsen, P.: Bankruptcy Prediction: a Comparison with Discriminant Analysis. Neural Networks in the Capital Market. John Wiley&Sons Ltd. (1995)

 8. Wei, W.X.: Comprehensive Method of Corporate Credit Rating Evaluation and its Application. Systems Engineering – Theory and Practice 2, 26–31 (1998)
 9. Liu, S.F., Guo, T.B., Dang, Y.G.: Grey system theory and Its Applications, 2nd edn. Science Press, Beijing (1999)
10. Sun, C.Y., Peng, Q.Y.: Multidimensional Grey Evaluation Model for Suppliers Selection and Evaluation. Journal of Southwest Jiaotong University 39(3), 277–280 (2004)
11. Wu, J.X., Wang, Z.J.: Study of Evaluation of Enterprises Credit Based on Analytic Hierarchy Process. Journal of Huazhong University of Science and Technology, Nature Science Edition 32(3), 109–111 (2004)

Video Key Frame Extraction for Semantic Retrieval

Xian Zhong[1], Yan-Sheng Lu[1], and Lin Li[2]

[1] School of Computer Science and Technology
Huazhong University of Science and Technolgy
cloudy0902@hotmail.com
[2] School of Computer Science and Technology
Wuhan University of Technology
cathylilin@whut.edu.cn

Abstract. With the emergence of a large number of video information, content-based video retrieval becomes a hot research topic. A shot is the basic unit of video analysis, lacking semantic information. How to extract key frames from a shot is the key issues of supporting for semantic video retrieval, which makes the obtained key frames represent the main content of a shot completely and accurately. This paper presents the key frame extraction process from top to down, being divided into three parts, i.e., shot segmentation, shot segmentation and key frame extraction. Each part will use the difference between frames. The most important work is feature extraction which is used during the whole process, so as to provide technical support for semantic-based video retrieval.

Keywords: video, key frame, sub shot, graph center.

1   Introduction

Video structure is usually divided into several levels of the video from big to small, scene, shot, frame and so on. The frame is an independent static image as the smallest unit of video [1-5]. Common Frame rate, so called FPS (Frame Per Second) is in the 24Hz/25Hz/30Hz sequence of frames, which makes person fell coherent [6-10]. The shot is a continuous frame sequence according to time, and it is the basic unit of video analysis, covering a small amount of semantic information.

Key frame is representative of the shot image. Key frame extraction methods can be roughly divided into two types of key frame selection and key frame recombination. Key frame selection is selecting one or multiple frames directly as a key frame from the shot frame sequence [11-15]. Key frame recombination is based on the characteristics of the shot, and to construct a new figure as a key frame. MEPG defines I frame, P frame, B frame. We can directly extract I frame to select the key frame [16-21]. Based on this idea, this paper presents the key frame extraction process.

The extracting procedure can be divided from top to down into three parts: the segmentation of shots, the segmentation of sub shots and the extraction of key frames, as show in Fig 1. The frame difference will be used in each part, and the feature extraction will be applied throughout the entire procedure.
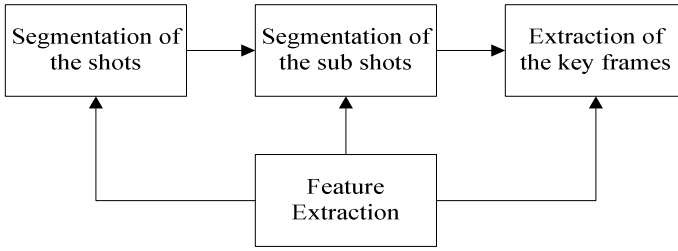
Fig. 1. Procedure of extracting key frames

2   Feature Extraction

MPEG (Moving Graph Experts Group) develops the multimedia content descriptor MPEG-7, which describes the low-layer graph features, including color, texture, shape and motion. The video has the characteristics of space and time. And in one shot, the feature of the video frame sequence gradually changes. This paper uses two color features, i.e., Dominant Color Descriptor (DCD) and Color Layout Descriptor (CLD) [22-25].

2.1   DCD

DCD describes the main color information in a given image area [26-28]. In a given image area, the color information is represented by a small number of representative colors. DCD is mainly used in the HSV color space where H stands for hue (Hue) related to the main light wavelength in visible spectrum. S stands for saturation (Saturation) and V stands for lightness (Value), also called Brightness. Specific process consists of the following four parts.
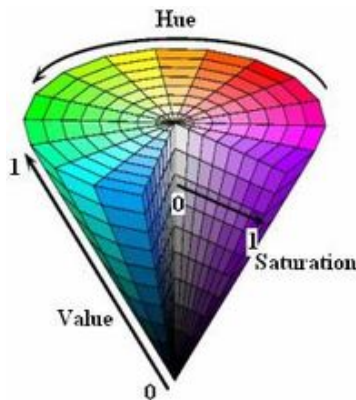
Fig. 2. HSV color space

(1) Coordinate Transformation

As shown in Figure 2, color is mapped from the RGB color space to HSV color space.

(2) Quantify

Color histogram describes the proportion of different colors. Each scale represents a color. We use non-equidistant to quantify the three components of the HSV color space, to distinguish between different colors in Equation 1-3.

$$H = \begin{cases} 0 & h \in [0°, 22°] \cup (330°, 360°) \\ 1 & h \in (22°, 45°] \\ 2 & h \in (45°, 70°] \\ 3 & h \in (70°, 155°] \\ 4 & h \in (155°, 186°] \\ 5 & h \in (186°, 278°] \\ 6 & h \in (278°, 330°] \end{cases} \tag{1}$$

$$S = \begin{cases} 0 & s \in (0.2, 0.65] \\ 1 & s \in (0.65, 1] \end{cases} \tag{2}$$

$$V = \begin{cases} 0 & v \in (0.2, 0.7] \\ 1 & v \in (0.7, 1] \end{cases} \tag{3}$$

As shown in Figure 3, when v<=0.2, that is black. When s<0.2and 0.2<v<0.8, the color is roughly divided into three sections, i.e., light gray, gray and dark gray, respectively. When s<0.2and v>0.8, that is white.
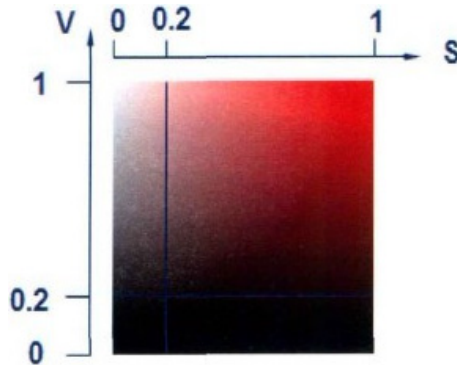
Fig. 3. Hierarchical quantization of color

(3) Main Color Extraction

Quantized colors C_i in accordance with the percentage of the corresponding descending P_i are sorted in descending order. We take the color of cumulative percentage $\sum P_i >= 60\%$ as the main color. If the number of main colors is large than eight, and then we use the one which has the largest percentage among eight main color.

$$F = \{\{C_i, P_i\}\}, (i = 1, 2...N), N \le 8 \tag{4}$$

(4) Similarity Measure

The similarity between two main colors is calculated as follows.

$$F_1 = \{\{C_{1i}, P_{1i}\}\}, (i = 1, 2...N) \tag{5}$$

$$F_2 = \{\{C_{2j}, P_{2j}\}\}, (j = 1, 2...M) \tag{6}$$

$$D^2(F_1, F_2) = \sum_{i=1}^{N} P_{1i}^{2} + \sum_{j=1}^{M} P_{2j}^{2} - \sum_{i=1}^{N} \sum_{j=1}^{M} 2a_{1i2j} P_{1i} P_{2j} \tag{7}$$

$$a_{kl} = \begin{cases} 1 - \dfrac{d_{kl}}{d_{max}} & d_{kl} \le T_d \\ 0 & d_{kl} > T_d \end{cases} \tag{8}$$

d_{kl} is computed by the conical unit space distance formula, and Td is the maximum of the space between the two points of the conical unit space. α is adjustable parameters, and its value is in the rang of 1 to 1.5.

$$d_{kl} = ((v_k - v_l)^2 + (s_k \cos h_k - s_l \cos h_l)^2 + (s_k \sin h_k - s_l \sin h_l)^2)^{1/2} \tag{9}$$

$$T_d = \sqrt{5} \tag{10}$$

$$d_{max} = aT_d \tag{11}$$

2.2   CLD

CLD reflects the color distribution of a given region in the image. The processing is as follows.

(1) Main Color Selection

The video frame is divided into 8 * 8 sub-blocks, and the primary colors of the sub-blocks (N = 1) is extracted, which constitutes the primary color matrix of 8 * 8. As a result the HSV space is mapped into the YCbCr color space.

(2) DCT Transform

The three color components of 8 * 8 main color of YCbCr color space were two-dimensional DCT transformed, getting three 8 * 8 matrix DY[8][8], DCb[8][8] and DCr[8][8], respectively extracting a total of 12 DCT coefficients.

(3)Similarity Measure

The similarity of the two main colors is calculated as follows.

$$D(CLD, CLD') = \sqrt{\sum_i w_{yi}(DY_i - DY'_i)^2}$$
$$+ \sqrt{\sum_i w_{bi}(DCb_i - DCb'_i)^2} + \sqrt{\sum_i w_{ri}(DCr_i - DCr'_i)^2} \tag{12}$$

2.3   Frame Difference

After getting DCD and CLD from the above two steps, their values are normalized. Last, their sum is computed.

3   Key Frames Extraction

3.1   Shot Segmentation

In the literature, there are various works on how to extract key frames [6-10]. Shot segmentation algorithm has been relatively mature. This paper chooses double threshold shot segmentation method which is proved to be effective by researchers.

3.2   Sub-shot Segmentation

This paper presents a feasible idea that divides sub-shot boundary into a small range. We add a sliding window to the video frames sequence, so the time complexity is reduced from n to 1, which is more in line with the real-time characteristics of streaming media.

We have adopted the idea of Fisher's linear discriminant to deal with the sliding window, as shown in the following three equations. C1 and C2 are two classes, m1 and m2 are the mean of the two classes, and s1 and s2 are the inside variance of the two classes. When J reaches the max value, the sub-shot boundary is found.

$$m_k = \frac{1}{N} \sum_{n \in C_k} x_n \tag{13}$$

$$s_k^2 = \sum_{n \in C_k} (x_n - m_k)^2 \qquad (14)$$

$$J = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \qquad (15)$$

Ideally, when all the frames of the sliding window in the same sub-shot, the value of J is approaching 0; When the sliding window enters a new sub-shot, the value of J gradually becomes larger; When the intermediate frame of the sliding window is just in the sub-shot boundary, J obtains a maximum value; Then the value of J gradually becomes smaller, until all the frames of the sliding window again are in the same sub-shot.

3.3   Key Frame Extraction

The frame is seen as the vertices of the graph, and the frame difference is seen as the edge of the connected vertex. That can transform the key frame extraction problem into the calculation of a fully connected graph center. The calculation of the graph center is as follows.

Suppose G is a undirected graph with n vertices, the distance between vertex vi and its farthest vertex is called the radius of vi, as shown in the following equation.

$$\rho(v_i) = \max_{1 \leq j \leq n}\{d(v_i, v_j)\} \qquad (16)$$

The vertex with smallest radius is called graph center, written as v_G, the radius is the radius of the graph, written as ρ_G that is

$$\rho_G = \max_{1 \leq j \leq n}\{d(v_G, v_j)\} = \min_{1 \leq i \leq n}\{\max_{1 \leq j \leq n}\{d(v_i, v_j)\}\} \qquad (17)$$

Graph center is the point of smallest radius, and the corresponding frame is the frame which difference with other frames' maximum frame difference is min. Thus, it can be seen as a representative frame, that is, the key frames.

Segmentation (KFESS).

4   Experimental Results

We used one video as experimental data. Our proposed method includes the shot segmentation, sub-shot segmentation and key frame extraction three steps to verify the effect. Figure 4, Figure 5 and Figure 6 shows the extracted key frames.

Fig. 4. Key Frames Extraction (1)
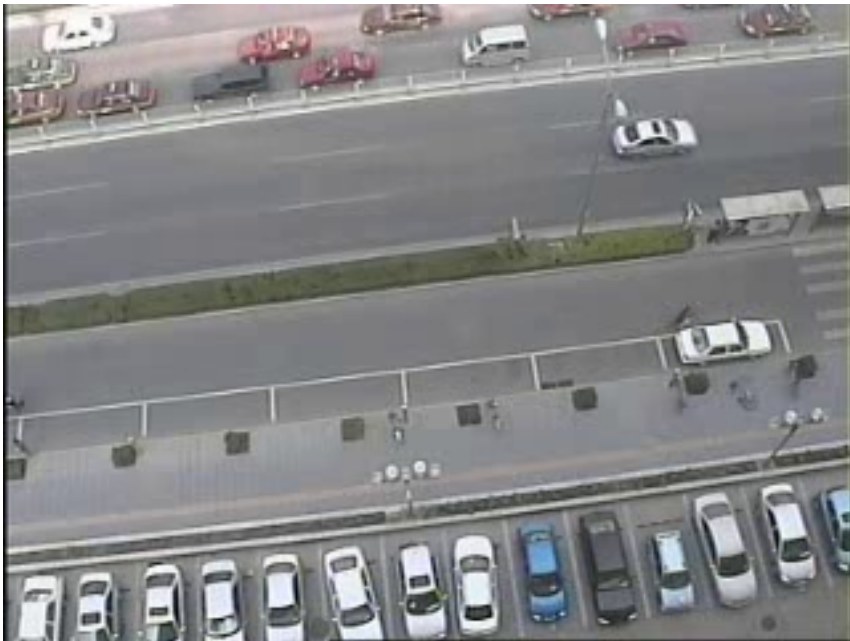
Fig. 5. Key Frames Extraction (2)

Fig. 6. Key Frames Extraction (3)

5   Conclusions

This paper proposes a video key frames extraction approach based on sub-shot segmentation, which is useful for contented based video semantic retrieval. Our approach utilizes the DCD and the CLD color feature to get frame differences by changing, quantization, and similarity computing for the shot segmentation. Then we add the video sequence with a sliding window to find the sub-shot segmentation scheme. After obtaining sub-shots, we find the center of a fully connected graph to extract key frames. The experimental results show that the key frames extraction can better reflect the contents of the video shot.

References

1. ISO/IEC JTC1/SC29/Wg11/N4062: MPEG-7 Requirements document, Singapore (2001)
2. Manjunath, B.S., Ohm, J.-R., Vasudevan, V.V., Yamada, A.: Color and Texture Descriptor. IEEE Transactions on Circuits and Systems for Video Technology 11(6), 703–715 (2001)

3. Chang, H.S., Sull, S., Lee, S.U.: Efficient Video Indexing Scheme for Content-based Retrieval. IEEE Trans. on Circuits and Systems for Video Technology 9(8), 1269–1279 (1999)
4. Luo, S.-L., Ma, S.-J., Liang, J., Pan, L.-M., Feng, Y.: Method of Key Frame Extraction Based on Sub-Shot Clustering. Transactions of Beijing Institute of Technology 31(3), 348–352 (2011)
5. Wang, H.-W., Shi, Z.-P., Shi, Z.-Z., Hu, H.: An Approach Selecting Episode Representation Frames Based on Subshot Clustering. Computer Engineering and Applications (27), 157–159 (2005)
6. Wang, L.: The key-frame based on the video monitor technology research. Beijing Institute of Technology (2008)
7. Hung, E.M., de Queiroz, R.L., Brandi, F., de Oliveira, K.F., Mukherjee, D.: Video Super-Resolution Using Codebooks Derived From Key-Frames. IEEE Trans. Circuits Syst. Video Techn (TCSV) 22(9), 1321–1331 (2012)
8. Lai, J., Yi, Y.: Key frame extraction based on visual attention model. J. Visual Communication and Image Representation (JVCIR) 23(1), 114–125 (2012)
9. Wang, Z., Kumar, M., Luo, J., Li, B.: Extracting key frames from consumer videos using bi-layer group sparsity. ACM Multimedia 1505–1508 (2011)
10. Kim, H.H., Kim, Y.H.: Toward a conceptual framework of key-frame extraction and storyboard display for video summarization. JASIST (JASIS) 61(5), 927–939 (2010)
11. Andon, P.: Accounting-related research in PPPs/PFIs: present contributions and future opportunities. Accounting, Auditing & Accountability Journal 25(5), 876 (2012)
12. Lord, A., Tewdwr-Jones, M.: Is Planning "Under Attack"? Chronicling the Deregulation of Urban and Environmental Planning in England. European Planning Studies 1 (2012)
13. Hodges, R.: Joined-Up Government and the Challenges to Accounting and Accountability Researchers. Financial Accountability & Management 28(1) (2012)
14. Mahony, P., Hextall, I., Richardson, M.: 'Building Schools for the Future': reflections on a new social architecture. Journal of Education Policy 26(3), 341 (2011)
15. Demirag, I., Khadaroo, I.: Accountability and value for money: a theoretical framework for the relationship in public–private partnerships. Journal of Management & Governance 15(2), 271 (2011)
16. Toms, S., Beck, M., Asenova, D.: Accounting regulation and profitability: The case of PFI hospital refinancing. Critical Perspectives on Accounting 22(7), 668 (2011)
17. Appuhami, R., Perera, S., Perera, H.: Coercive Policy Diffusion in a Developing Country: The Case of Public-Private Partnerships in Sri Lanka. Journal of Contemporary Asia 41(3), 431 (2011)
18. Hodkinson, S.: The Private Finance Initiative in English Council Housing Regeneration: A Privatisation too Far? Housing Studies 26(6), 911 (2011)
19. Arnaboldi, M., Lapsley, I.: Asset management in cities: polyphony in action? Accounting, Auditing & Accountability Journal 23(3), 392 (2010)
20. Rangel, T., Galende, J.: Innovation in public–private partnerships (PPPs): the Spanish case of highway concessions. Public Money & Management 30(1), 49 (2010)
21. English, L., Baxter, J.: The Changing Nature of Contracting and Trust in Public-Private Partnerships: The Case of Victorian PPP Prisons. Abacus 46(3) (2010)
22. Hyndman, N., McDdonnell, P.: Governance and Charities: an Exploration of Key Themes and the Development of a Research Agenda. Financial Accountability & Management 25 (1) (2009)

23. Hillier, J., Van Wezemael, J.: 'Empty, Swept and Garnished': the Public Finance Initiative case of Throckley Middle School. Space and Polity 12(2), 157 (2008)
24. Benito, B., Montesinos, V., Bastida, F.: An example of creative accounting in public sector: The private financing of infrastructures in Spain. Critical Perspectives on Accounting 19(7), 963 (2008)
25. Athiyo, R., Smith, N.J., Aritua, B.: Private finance for the delivery of school projects in England. Proceedings of the ICE - Management, Procurement and Law 161(4), 141 (2008)
26. Broadbent, J., Guthrie, J.: Public sector to public services: 20 years of "contextual" accounting research. Accounting, Auditing & Accountability Journal 21(2), 129 (2008)
27. Khadaroo, I.: The actual evaluation of school PFI bids for value for money in the UK public sector. Critical Perspectives on Accounting 19(8), 1321 (2008)
28. Jones, M.J., Mellett, H.J.: Determinants of changes in accounting practices: Accounting and the UK Health Service. Critical Perspectives on Accounting 18(1), 91 (2007)

Granular Sketch Based Uncertain Time Series Streams Clustering

Jingyu Chen[1,2], Ping Chen[2], and Xian'gang Sheng[3]

[1] School of Computer Science and Technology,
Xidian University, 710071, Xi'an, China
[2] Software Engineering Institute, Xidian University,
710071, Xi'an, China
{jychen,chenping}@mail.xidian.edu.cn
[3] College of Information Engineering, Qingdao University,
266071, Qingdao, Shandong, China
stzhibuwu@qdu.edu.cn

Abstract. Uncertainty is inherent in data streams, and presents new challenges to data streams mining. For continuous arriving and huge size of data streams, it requires significantly more space to represent and cluster the uncertain time series data streams. Therefore, it is important to construct compressed representation for storing uncertain time series data. The granular sketches and buckets policy are designed through hash-compressed storage and micro clusters. Then based on the max-min cluster distance measure, an initial cluster centers selection algorithm is proposed to improve the quality of clustering uncertain data streams. Finally, the effectiveness of the proposed algorithm is illustrated through analyzing the experimental results.

Keywords: granulation, sketch, clustering, uncertain time series, data streams.

1 Introduction

With the developing of information technique and wireless sensor networks, a great variety application need to analysis and mining large-scale, continuous and rapid arriving data streams. As a common data form in data streams, time series data streams contain much more useful pattern and information for many applications fields.

Time series is a series of observation data according to a certain time sequence, and aggregates with time and event [9]. Time-series data mining is an important way which mining some useful and potential knowledge from a great deal of time-series. Time series data streams are created by many applications such as RFID, traffic, mobile, economic and finance applications. In these applications, uncertainty often happens because of network failure, noise and sampling error, etc. Probability time series are common uncertain time series data. The probabilistic information should be handled in clustering uncertain time series data streams.

To address the problem of uncertain time series clustering, we suggest a hash based granular sketch approach to reduce storage spaces and computational complexity. The approach is based on granulation mechanism, which granulates uncertain time series data to a set of possible sequences. Then we use hash technique to compress all possible sequences within a sliding window to a granular sketch. Then according hash value of sequences, we map objects to different buckets to put objects which have similar hash value together. Based on buckets, we measure the similarity of objects in same buckets and among different buckets. Then, using the max-min cluster distance measure, the initial cluster centers selection algorithm is proposed to improve the quality of clustering. Based on UKmeans [1], we design an algorithm to cluster uncertain time series data streams. Final, we verify the accuracy and efficiency of the proposed scheme via experiments.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 presents a granular sketch model for uncertain time series data. Section 4 outlines the granular sketch based clustering methods. Simulation methodology and performance evaluation result and analysis are presented in section 5, and we conclude the work in section 6.

2   Related Work

Time series data mining has been attracting much attention in research and practice. Nowadays, the study about time-series mining is mainly about similarity mining. In general, time series distance or similarity measurements can be classified into two broad categories: model-based approaches and similarity-based approaches.

In similarity-based approaches, for the measurement of similarity between two sequences, Euclidean distance based schemes were proposed including EU (Euclidean Distance), DTW (Dynamic Time Warping Distance), ERP (Edit Distance with Real Penalty), LCSS (Longest Common Subsequences), and EDR (Edit Distance in Real Sequence).

In model-based approaches, analytical models are established to fit time series data, and calculate matching degree through the trajectories generation probability of analytical models. Examples of models include regression models, ARMA models, ARIMA models and HMMs [2, 3]. Among these models, ARIMA models are a top-choice linear method. For model-based approaches, it is much easier to exploit the Priori knowledge, and can be applied to trajectories in different lengths.

Uncertain time series often have same characteristics as uncertain data streams, such as uncertain, large-scale, continuous and rapid arriving. To handle the continuous arriving uncertain time series stream, we often use and expand some streams mining algorithms.In recent years there have been a plenty of methods for managing and mining uncertain data streams. Several important stream clustering algorithms have been introduced in recent years. Li et al. propose a method for detecting moving cluster based on the idea of Moving Micro-Cluster [8]. Moving Micro-Clusters are an extension of the idea of micro-clusters used in the BIRCH clustering algorithm [9]. In [10], the authors present an incremental graph-based clustering algorithm RepStream, which uses representative points to update clusters

incrementally and improve the clustering effectiveness. Chau et al. present the UK-means clustering, an algorithm that enhances the K-means algorithm to handle data uncertainty [1]. In [4], authors proposed a GARCH metric to create probabilistic databases for handling uncertain time series data. Ackermann et al. present a new corsets trees based clustering algorithm to improve quality of stream clustering [5]. Tran et al. present the PODS model for processing uncertain data using continuous random variables [6].

Sketch is a popular method for handling huge and fast data streams. Sketch techniques use a sketch vector as a data structure to store the streaming data compactly in a small-memory footprint. The main advantage of using these sketch techniques [11, 12] is that they require a storage which is significantly smaller than the input stream length. Sketch techniques are used in stream data frequent items mining [13, 14], clustering [15] and anomaly detection [16] recently.

Our work is closely related to cluster uncertain time series data. In this paper, we model the uncertain time series data based on granular sketch. We use a granular sketch approach to create hash-compressed representations. We construct buckets to divide objects according the hash value of vectors in granular sketches. We also propose a similarity measurement method to select initial cluster centers optimized.

3   Granular Sketch

One of the most effective ways to deal with imprecise and uncertain data is to employ probabilistic approaches. The possible world model is a common modeling method used for probabilistic and discrete values. The possible world model produces large result sets and great time and space complexities, and is inefficient for processing the continuous arriving uncertain data streams.

Probability time series is a kind of uncertain time series data, which has time, event value and event's probability. Since we may get several probabilistic values at a certain time t, the probability time-series data may have much more size of data. The probability also increases the complexity of modeling and analysis of time series data. The way of modeling the probability of data is a key point of uncertain data stream mining. We use the following definition to represent and store the probability time series data.

Definition 1. (Element) An element $e=\{o_i,<x,p,t>\}$ is a basic probability time series data of a object o_i, where an o_i represents an observation object's id, x represents an observation value , p represents the probability of x value, and t represents observation time.

Definition 2. (Item) An item $I_t(o_i)$ of an object o_i at time t includes all element at time t. $I_t(o_i) = \{o_i, (<x_1,p_1,t>, <x_2,p_2,t>,…)\}$, where o_i is the identity of the observation object.

Examples of Uncertain time series data are showed in Table 1. In Table 1, the object o_1 has two possible items $I_1(o_1) =\{o_1, (<a, 0.6, 1>, <b, 0.4, 1>)\}$ at the time t=1.

Definition 3. (n-length uncertain sequence set) A n-length uncertain sequences set $UT_n(o_i)$ of an object o_i is a set of all possible sequence that include n consecutive

items of uncertain time series UT_n to represent k-length of time for the object o_i, $UT_n(o_i)=\{ I_1(o_i), I_2(o_i),\dots I_n(o_i)\}$.

Table 1. Examples of Uncertain time series data

Time	id=o_1		id=o_2		id=o_3	
	x	p	x	p	x	p
1	a	0.6	c	0.5	f	0.4
1	b	0.4	d	0.5	e	0.6
2	a	0.5	c	0.8	d	0.7
2	b	0.5	d	0.2	f	0.3
3	a	0.3	c	0.2	g	0.2
3	b	0.4	d	0.4	b	0.5
3	e	0.3	e	0.4	d	0.3

The probability increases difficulty for analysis uncertain time series, for it generates much more possible combination of sequences and make sequences model much complex.

Granulation cognition plays an important role for complex data modeling. We can use granulation mechanism to model and represent the complex uncertain time series data. To deal with the diversity of uncertain sequential data, we describe sequences of time points using a set of granules. We consider the different length of subsequences time points as different granules. We define the following concepts for uncertain time series granulating:

Definition 4. (k-Granule) A k-granule $G_k(o_i)$ includes a set of k-length subsequences and probability value of each subsequences. For example, the 2-Granule $G_2(o_1)$ of object o_1 is the set $\{[I_1,I_2], [I_2,I_3]\}=\{[(<a,0.6,1>, <b,0.4,1>), (<a,0.5,2>,<b,0.5,2>)], [(<a,0.5,2>,<b,0.5,2>),(<a,0.3,3>,<b,0.4,3>,<e,0.3,3>)]\}$

Table 2. An example of 2- Granule

G(o_1,2)	
S(G(o_1,2))	P(G(o_1,2))
aa	0.45
ab	0.5
ba	0.35
bb	0.35
ae	0.2
be	0.15

A k-granule $G_k(o_i)$ consists of two parts: $S(G_k(o_i))$ and $P(G_k(o_i))$. $S(G_k(o_i))$ is a sequences set that includes k consecutive elements of a k-length uncertain sequence $UT_n(o_i)$, which is presented as a list of sorted elements. From the 2-Granule $G_2(o_1)$ of object o_1, we also can get a set of sequences $S(G_2(o_1)) = \{[aa,ab,ba,bb],$

[aa,ab,ae,ba,bb,be]}. $P(G_k(o_i))$ is a set of probability value of each subsequence in $S(G_k(o_i))$. The probability value of each sequence is equal to the product of probabilities of each data. For example, for $<a,0.6,1>$ and $<a,0.5,2>$ data, the possibility of a sequence "aa" is $P(aa)=0.6*0.5=0.3$. Correspondingly, probabilities set of each sequence $P(G_2(o_1)) = \{[0.3, 0.3, 0.2, 0.2], [0.15, 0.2, 0.15, 0.15, 0.2, 0.15]\}$. We can shingle the sequences set $S(G_2(o_1))$ to a shingle set {aa,ab,ba,bb,ae,be }. An example of 2-Granule $G_2(o_1)$ is showed in Table 2.

Through this granular approach, we can group subsequences of uncertain time series streams into different length of granules. For the diversity of k-Granule, we also need to compress the storage space. We construct sketches to storage k-Granules compressed, and calculate approximately the similarity between objects' time series by testing similarity between these sketches.

3.1   Granular Sketch

Assume that the number of objects is N in uncertain time series. We adopt sliding windows technique to process the continuous arriving time series data. The size of each sliding window is L. In a time window w_i, we construct a granular sketch for objects to reduce the scale of uncertain time series data. We compress various possible sequences of n-length uncertain time series of N objects to a granular sketch in a time window. Rather than compare all possible sequences, we reduce the similarity calculation on sets by testing similarity on granular sketches.

Definition 5. (Granular Sketch)A granular sketch is a two-dimensional matrix SK[r, c] (r<<N, c=N), c is the number of objects and r is defined according by a hash functions hr. The sequence set of granules are denoted by SJ, such as {aa, ab, …}. Let hr: SJ→{0,…,r} be a hash function that hash a sequence to a row number. The identify set of objects are denoted by oid, such as {o_1, o_2, …}. We store object to a column according it's identify. The initial value of each element in sketch is 0. For each sequence sj in a sequence set SJ of a object o_i, we add possibility value P(sj) of sj in the granular to SK[hr(sj), o_i].

$$SK[hr(sj), o_i]= SK[hr(sj), o_i]+ P(sj) \qquad (1)$$

An example of a granular sketch is showed in Table 3.

Table 3. An example of granular sketch

SK	1 (=o_1)	2 (=o_2)	3 (=o_3)
hr(aa)	0.45	0	0
hr(ab)	0.5	0	0
hr(ba)	0.35	0	0
hr(cc)	0	0.4	0
hr(cd)	0	0.1	0
hr(fd)	0	0	0.28

As the size of the granular sketch even is much less than the size of primitive uncertain streams within a sliding window w_i. As time series data arriving, the granular sketch may be filled full and saturated. So we use a set of m granular sketches $SKS=\{SK_0,SK_1,\ldots SK_m\}$. The set size m is defined by user. We store all time series data within a sliding window to a single granular sketch. As new stream data arrive and sliding window slides from w_{i-1} to next window w_i, we also store uncertain granular data in next granular sketch. We use the m granular sketches cyclically and orderly. When finish the SK_{m-1}, we clear the SK_0 and reuse the granular sketches set from SK_0. We show an example of m granular sketches in figure 1.
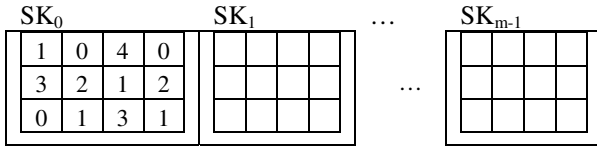
Fig. 1. The m granular sketches

There are m granular sketches, and each of them contains a granular sketch to store approximated uncertain time series data streams.

3.2   Buckets Policy

Based on the set of granular sketch, we use buckets policy to facility the similarity calculation between objects. We create len buckets denoted by $BS=\{B_0,B_1,\ldots,B_{len}\}$.

For improve the similarity calculation, we map those objects that have same hash value to the same bucket. We use consistent hash function [18] h_b: $\{0..r\}*$ to $\{0..len\}$ to map an object to a bucket number according the hash value of its column vector in a granular sketch.

For example, in figure 4, there are four column vectors representing four objects' sketch core. The second and fourth column both have column vector [0, 2, 1], so the corresponding two objects will be hash to a same bucket. For len is far less than the sketch column c and the collision of hash function, so objects with equal or unequal column vectors can map to same of different buckets.

In a bucket B_i, we use a bitmap BT_i to accumulate the number of objects and the identity of sketch cores that objects are mapped from. The bitmap $BT_i[c, m]$ is a two dimensions bit array, and has c (c=N) rows representing hash identifies of objects and m columns representing identifies of sketch core that objects map from.

When new stream arriving and the sliding window is going to switch to next, we use buckets policy to accumulate objects to buckets from current sketch core.

4   Granular Sketch Based Clustering

According the granular sketches and buckets constructed above, based on UK-means we proposed a new algorithm for clustering uncertain time series objects.

4.1   Similarity

Based on Jaccard similarity, the similarity between two objects o_a and o_b in a bucket B_i is Simi(oa , ob) .

$$Sim_i(o_a, o_b) = \frac{BT_i[o_a] \times BT_i[o_b]}{(BT_i[o_a] + BT_i[o_b])^2} \tag{2}$$

where $\times$ means two row vectors inner product and $(\ )^2$ means vector self inner product.

$$SIM(o_a, o_b) = \frac{\sum_{i=0}^{m} Sim_i(o_a, o_b)}{m} \tag{3}$$

The similarity among buckets is the average of similarity insides a bucket.

4.2   Micro Cluster

For hash collision, there are many objects in a bucket. Based on the similarity in a bucket, we can divide objects into different group to form different micro clusters.

The basic idea to construct micro clusters is to select an object which has maximum count value as the first micro cluster center. Then select the farthest object to existing center as a new micro cluster center and form a new micro cluster.

The micro clusters in a bucket B_i are constructed as follows.

Input: a bucket B_i

Output: a set of micro clusters centers

For each objects in the bucket B_i, count the row value of BT_i [hc(o_j)] to a one-dimension array VA_i[hc(o_j)].

Select a object o_{m1} which has the maximum summation value in VA and set o_{m1} as a center of the first micro cluster mc_1.

In a bucket, based on the formula (2), calculate the similarity between o_{m1} and other objects, and select the most different object o_{m2} as the center of second micro cluster mc_2.

Define the similarity $Sim_i(o_{m1}, o_{m2})$ between o_{m1} and o_{m2} as a expect distance Ed_i.

For rest objects in the bucket B_i, if the similarity between an object o_j to a center of existing micro cluster mc_i is larger than α^* Ed_i, set o_j as a member of the micro cluster mc_i; if the similarities between an object o_j to all centers of existing micro clusters are smaller than α^* Ed_i, select as a center of new micro cluster. Where α is a test parameter, usually $1/2 \leq \alpha \leq 1$.

As all objects in the bucket B_i are checked, a set of micro clusters and centers is generated.

The 1) to 3) steps of above algorithm are to select the first and second micro cluster centers. The 4) to 6) steps are to form other micro clusters.

The test parameter α is very important for dividing micro clusters. The larger value is the test parameter α, the more micro clusters will generate. The test parameter α should be choose properly.

4.3   Initial Cluster Centers Selection

UK-means is a known cluster algorithm based on K-means and is used on uncertain streams. The clustering results of UK-means largely depend on the selection of initial cluster centers. If the initial cluster centers are selected unreasonably, the clustering result is likely to be local minima.

To choose the initial cluster centers, we design a selection policy based on a max-min cluster distance algorithm to calculate distances among micro cluster centers of all buckets. The main idea of max-min cluster distance algorithm is to select minimal distance in a bucket and maximum distance among buckets.

The max-min cluster distance based initial cluster centers selection algorithm is shown as following.

Input: all buckets

Output: a set of initial clusters centers

In a bucket B_i, we calculate the similarity between each center of micro clusters as distances of centers.

Select the centers whose average distance to other micro cluster centers is larger than $\alpha * Ed_i$ as a candidate centers set of bucket B_i, and obtain candidate centers sets for each bucket.

Merge each bucket's candidate centers sets onto a candidate centers set CCS. Choose one candidate center cc_j from CCS as the first initial cluster center ic_1.

Based on the formula (2), calculate the similarity between ic_1 and other candidate centers among buckets, and select the candidate center which has the farthest distance to ic_1 as the second initial cluster center ic_2.

Define the similarity $SIM(ic_1, ic_2)$ between ic_1 and ic_2 as a expect initial distance Eid_i.

For rest candidate centers in the CCS, if the similarity between a candidate center cc_j to a existing initial cluster center ic_i is smallar than $\alpha * Eid_i$, select cc_j as a new initial cluster center.

The 1) to 3) steps of above algorithm are to select the first and second initial cluster centers. The 4) to 6) steps are to select other initial cluster centers. We adopt max-min cluster distance algorithm to select initial cluster centers. This algorithm can guarantee each cluster center is much more far from the existing cluster centers. Moreover, the algorithm does not need to be given the number of clusters in advance; it can intelligently determine the number of initial cluster centers.

4.4   USeriesUKm Algorithm

Based on the selection of initial cluster centers and UK-means algorithm, we build our USeriesUKm algorithm for clustering uncertain time series of observation objects.

The USeriesUKm algorithm includes four main phases. First, construct granular sketches to accept and storage continuous arriving uncertain time series of observation objects. Secondly, we use hash function to map objects to different buckets. Thirdly, we construct micro clusters and obtain centers of micro clusters from each bucket. Finally, we select initial cluster centers from micro cluster centers, and cluster the uncertain stream based on UK-means.

5   Experiment

This section shows the results from our experiment to validate the accuracy and efficiency of our proposed clustering algorithm.

5.1   Evaluation Standard

We will show the accuracy of our algorithm based on the evaluation standard of RandIndex[17]. For uncertain time series data set D (includes N objects), let T={T_1, T_2, …, T_k} represent the original clusters, and C={C_1, C_2, …, C_m} be the clusters produced by a clustering algorithm. Let a represent an object that is in a cluster of C and in a cluster of T either. Let b represent an object that is in a cluster of C but NOT in any cluster of T.

$$SRAND = \frac{a + b}{n(n-1)/2} \tag{4}$$

The SRAND represent the degree of matching between T and C. The greater of the SRAND value means the better of the clustering.

5.2   Experimental Results

We compare our USeriesUKm algorithm to UK-means by using the Census 1990 data set. For each record in the data set, we add 5 possible data and probability value, which are used to indicate the probability value of possible point. We assigned an probability following the normal distribution from the range (0,1] to each item in the data set.

Figure 2 shows the comparison of clustering quality between USeriesUKm and UK-means with the increase in the size of objects set. In figure 2, the SRAND value of USeriesUKm is higher than that of UK-means. Because the USeriesUKm increase the calculation of initial cluster centers selection to improve the clustering quality. The clustering quality of USeriesUKm is also very much of variation of the size of testing data set.
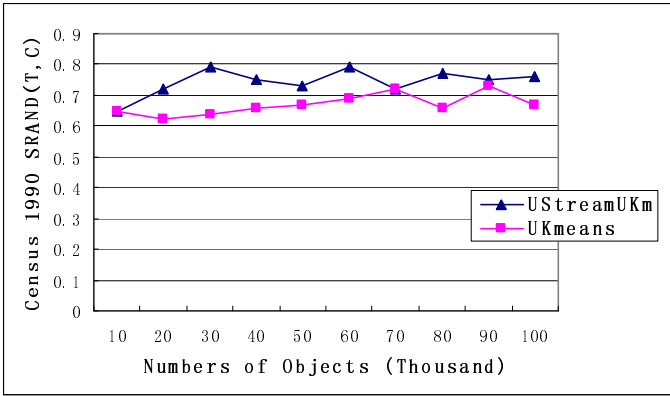
Fig. 2. Clustering quality comparison

Figure 3 shows the comparison of time consuming between USeriesUKm and UK-means with the increase in the size of testing data set. In figure 3, the time consuming of USeriesUKm is higher than that of UK-means. It will cost the USeriesUKm some time to increase the calculation of initial cluster centers selection to improve the clustering quality.
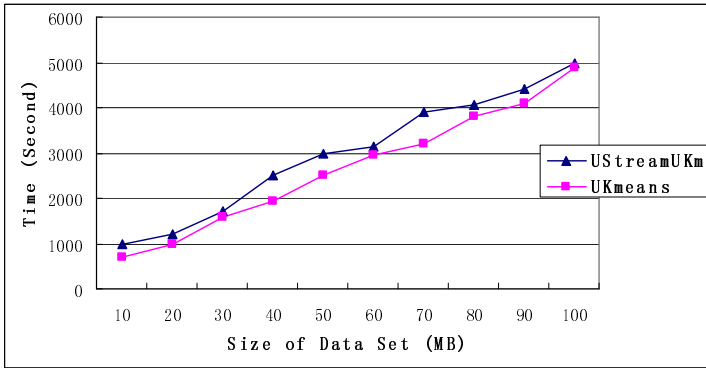
Fig. 3. Time consuming comparison

Figure 4 shows the comparison of memory space consuming between USeriesUKm and UK-means with the increase in the size of objects set. In figure 4, the memory space consuming of USeriesUKm is smaller than that of UK-means. Because the USeriesUKm adopt granular sketch and hash based method to compress storage space.

Experiment results show that the USeriesUKm algorithm can effectively cluster uncertain time series streams. The USeriesUKm benefits from using compressed storage space and initial cluster centers.
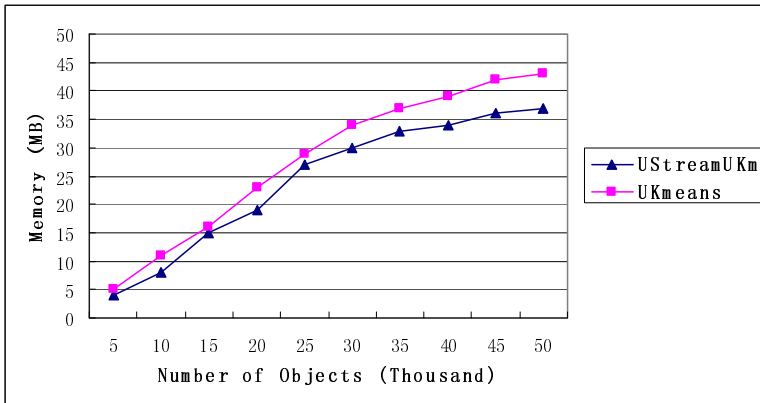
Fig. 4. Memory space comparison

6   Conclusion

This paper has addressed a new efficient granular sketch based clustering model for uncertain time series streams. First, for reducing memory consumption, we construct granular sketches to store continuous arriving uncertain time series data. Secondly, we use a hash-based approach to map objects to different buckets. Then, based on the Max-min cluster distance algorithm, an algorithm for selecting the initial cluster centers is proposed. And then, based on UK-means, our USeriesUKm algorithm for clustering uncertain time series objects is given. Experiment results show that the USeriesUKm algorithm can cluster uncertain data streams with better accuracy and efficiency. In future work, we plan to design distributed sketches and clustering algorithm for distributed uncertain data streams.

References

1. Chau, M., Cheng, R., Kao, B., Ng, J.: Uncertain data mining: An example in clustering location data. In: Ng, W.-K., Kitsuregawa, M., Li, J., Chang, K. (eds.) PAKDD 2006. LNCS (LNAI), vol. 3918, pp. 199–204. Springer, Heidelberg (2006)
2. Gaffney, S., Smyth, P.: Trajectory clustering with mixtures of regression models. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, pp. 63–72 (August 1999)
3. Xiong, Y., Yeung, D.: Mixtures of ARMA models for model-based time series clustering. In: Proceedings of the 2002 IEEE International Conference on Data Mining, Maebashi City, Japan, pp. 717–720 (December 2002)

4. Sathe, S., Jeung, H., Aberer, K.: Creating probabilistic databases from imprecise time-series data. In: Proceedings of the 2011 IEEE International Conference on Data Engineering (ICDE), pp. 327–338 (2011)
5. Ackermann, M.R., Lammersen, C., Martens, M., Raupach, C., Swierkot, K., Sohler, C.: StreamKM++: A Clustering Algorithm for Data Streams. Journal of Experimental Algorithmics (JEA) 17(1) (July 2012)
6. Tran, T.T.L., Peng, P., Li, B.D., Diao, Y., Liu, A.N.: PODS: a new model and processing algorithms for uncertain data streams. In: Proceedings of the 2010 International Conference on Management of Data, Indiana, USA, pp. 159–170 (2010)
7. Shao, F., Yu, Z.: Principle and Algorithm of Data Mining. Water conservancy & water electric press of China, Beijing (2003)
8. Li, Y., Han, J., Yang, J.: Clustering Moving Objects. In: Proc. of the 10th ACM SIGKDD Int'l. Conf. on Knowledge Discovery and Data Mining (2004)
9. Zhang, T., Ramakrishnan, R., Livny, M.: BIRCH: an efficient data clustering method for very large databases. In: Proc. of the 1996 ACM SIGMOD Int'l. Conf. on Management of Data (1996)
10. Luhr, S., Lazarescu, M.: Incremental clustering on dynamic data streams using connectivity based representative points. Data & Knowledge Engineering, 1–27 (2009)
11. Alon, N., Matias, Y., Szegedy, M.: The Space Complexity of Approximating the Frequency Moments. In: ACM Symposium on Theory of Computing, pp. 20–29 (1996)
12. Cormode, G., Muthukrishnan, S.: An Improved Data-Stream Summary: The Count-min Sketch and its Applications. Journal of Algorithms 55(1), 58–75 (2005)
13. Cormode, G., Muthukrishnan, S.: What's hot and what's not: Tracking most frequent items dynamically. In: Proceedings of the 22nd ACM Symposium on Principles of Database Systems, pp. 296–306 (2003)
14. Manerikar, N., Palpanas, T.: Frequent items in streaming data: An experimental evaluation of the state-of-the-art. Technical Report DISI-08-017, University of Trento (March 2008)
15. Aggarwal, C.: A Framework for Clustering Massive-Domain Data Streams. In: IEEE 25th International Conference on Data Engineering (ICDE 2009), pp. 102–113 (2009)
16. Liu, Y., Zhang, L., Guan, Y.: Sketch-based Streaming PCA Algorithm for Network-wide Traffic Anomaly Detection. In: Proc. ICDCS (2010)
17. Somasundaram, R.S., Nedunchezhian, R.: Evaluation of three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values. International Journal of Computer Applications 21(10) (May 2011) 0975–8887
18. Stoica, I., Morris, R., Karger, D., Kaashoek, M.F., Balakrishnan, H.: Chord: A scalable peer-to-peer lookup service for internet applications. In: Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, pp. 149–160. ACM Press (2001)

Analysis of FC-AE-ASM and FC-AE-FCLP Systems

Yiwei Zhou, Feng He, Wentao Chen, and Huagang Xiong

School of Electronic and Information Engineering, Beihang University, Beijing, China
{zhouyiwei,hgxiong}@ee.buaa.edu.cn,
Robinleo@buaa.edu.cn, hg6012135@163.com

Abstract. Fibre Channel (FC) has become the first choice for avionics system interconnection, and the characteristics of FC upper layer protocols have great influence on the whole avionics systems. This paper focuses on the performance test of the implementation systems in FC upper layer. Meanwhile, corresponding methods were proposed to test the delay, throughput and frame loss rate. A test platform has been built to test the FC-AE-ASM and FC-AE-FCLP implementation systems utilizing the proposed methods. The results show that the FC-AE-ASM protocol implementation system has 19.8% less delay and 4.2% more throughput than the FC-AE-FCLP protocol implementation system maximally. It can be concluded that the FC-AE-ASM protocol can use the bandwidth resources more efficiently, and is more suitable for the period data transmission in avionics networks.

Keywords: Fibre Channel, Anonymous Subscribers Messaging, Fibre Channel Lightweight Protocol, Performance analysis.

1 Introduction

Modern avionics network has an increasing need for bandwidth while maintaining the traditional needs of high reliability, fault tolerance and deterministic behavior to support real-time communication. Fibre Channel is a high speed serial transmission protocol, with the properties of high speed, low delay, strong anti-interference [1], and it has become the primary candidate protocol for future avionics network. To meet the demands of mission-critical avionics applications, the American National Standards Institute (ANSI) developed the Fibre Channel Avionics Environment (FC-AE), which has been applied in some advanced fighter planes such as F35/JSF, F18, F16, E2C, etc [2]. The FC-AE recommends several kinds of upper layer protocols, such as Anonymous Subscriber Massaging (ASM), Fibre Channel Lightweight Protocol (FCLP), Remote Direct Memory Access (RDMA), and FC-AE-1553 to meet different avionics environment needs.

At present, the test of FC-AE implement system has drawn a lot of interests from researchers, and different ways have been developed. For example, [3] presented methods to realize the FC-AE-1553 transmission hardware platform, and tested the result to see whether the implementation system can communicate according to the protocol precisely. [4] and [5] simulated the performance of FC-AE-1553 network in

avionics system using OPNET. [1] [6] and [7] analyzed the FC-AE-ASM protocol, but [1] and [6] chose to realize the protocol over hardware systems, and they mainly focused on the conformance of the implementation systems like [3]. In [7], a method of adopting Stochastic Petri nets to determine (DPSN) model to simulate the performances like network load and delay of FC-AE-ASM network was proposed, and it got the result that FC-AE-ASM network can improve real-time, reliability of the data transmission and completely meet the demands of future avionics system. Zhu et al. [8] simulated the end-to-end delay of FC Network with OPNET. But both [7] and [8] stayed at the theoretical level and did not provide instances to certificate their results. At the same time, some previous investigations about the delay measurement had been developed, such as [9]. In [9], they implemented a passive One Way Delay measurement system composed by a set of synchronized measurement points that observed selected packets and reported transit timestamps to a collector station. However, the method is too complex and not suitable for testing the delay of FC upper layer protocols' implementation systems.

As mentioned above, a set of mature methods to test FC upper layer protocols' implementation systems' performance has not been proposed. To built a testing system, we have to cope with those problems: first, the frames used in the test have to strictly comply the protocols; second, the test data flow should not interface the regular communication data flow in the avionics networks; third, the timing of the test system has to be accurate and the clock drift between distributed end systems has to be diminished. In this paper, we choose delay, throughput and frame loss rate as the items of the performance test. As the FC-AE-ASM and FC-AE-FCLP are two of the most widely used FC upper layer protocols today, we test the two protocols' implementation systems as experiments, and we analysis the result of our test.

The reminder of this paper is organized as follows. Chapter 2 provides the introduction of the FC-AE-ASM protocol and the FC-AE-FCLP protocol. In Chapter 3, we present the test system we designed and detailed methods to test the performance of the implementation systems are proposed. The test results and the analysis are provided in Chapter 4. And finally, the paper is concluded in Chapter 5.

2   Analysis of FC-AE-ASM and FC-AE-FCLP Protocol

2.1   FC-AE-ASM Protocol

FC-AE-ASM is intended to support bidirectional communication between two N_Ports in a constrained and carefully defined environment, typical of avionics applications. The intended usage is avionic command, control, instrumentation, simulation, signal processing, and sensor/video data distribution. These applications areas are characterized by a variety of requirements, among them a need for high reliability, fault tolerance, and deterministic behavior to support real-time control/response [10]. The Type code of FC frame that supports FC-AE-ASM protocol is hex'49', and the first four words in the payload of the FC frame are used as the FC-AE-ASM header.

In the process of data transmission of FC-AE-ASM protocol, the sender sends the messages according to some rules and the receiver receive the messages at a

predetermined rate and it does not necessarily know from which sender the messages come. All the activities in a communication process are well designed before the communication happens.

2.2   FC-AE-FCLP Protocol

FC-AE-FCLP follows the FCP and FC-FS and FC-AL-2 standards in its definition of the services necessary to support low-latency, low overhead communication between elements of a mission-critical avionics system [2]. In FC-AE-FCLP protocol formulates that an access channel is a collection of channels which is specified with an access point identifier (APID). An access point and a channel together specify the endpoint for communication. There is a virtual connection between the sender and the receiver. In application, we often use a channel number (chnum) to specify the virtual connection and use an application numbers (apnum) to keep track of APIDs in use. The Type code of FC frame that supports FC-AE-FCLP protocol is hex'08'.

The communication process of FC-AE-FCLP implementation system contains five stages, which are Setup Channel Stage, Acknowledge Channel Setup Stage, Send Data Stage, Close Channel Stage and Return APIDs Stage.

3   Performance Test of FC Upper Level Protocols

We designed a test system that takes three aspects about the FC upper layer protocols into consideration: the delay, the throughput and the frame loss rate. We realize all the test functions by developing the Vxworks real-time operation system and adding some pieces of code to the original FC upper layer realization software. The test architecture is shown in Fig. 1.
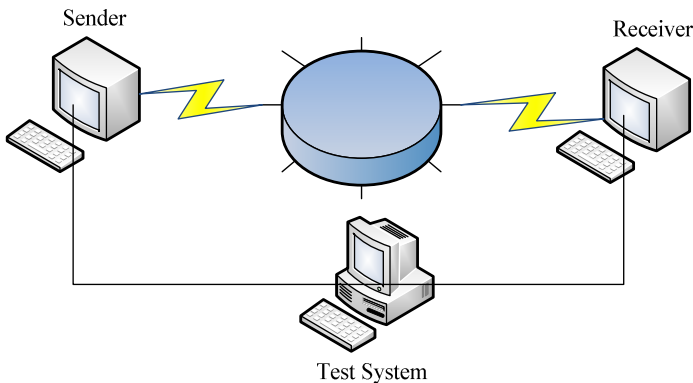
Fig. 1. The test architecture

The detailed methods and algorithms suitable for the test of each aspect of FC upper layer implementation system's performance are demonstrated as follows:

3.1   Delay Test

The delay is the time that a frame spends transfer from one end system to another end system. It reflects the speed of a system to process the frames, which is critical for the real-time transmission tasks on the avionics network.

A Global Time method is proposed to avoid the time drift between different distributed end systems. We used a test PC to provide global time. At the time the sender sends a test message, it sends signal A to the test PC. At the time the receiver receives the message, the receiver sends signal B to the test PC. The test PC then calculates the difference between the time it receives signal B and the time it receives signal A. To increase accuracy, the sender sends the messages the same with the test message with a period of T_{cons}. To guarantee the accuracy of the delay test, we send the test message with a some constant payload length N times, and the period of test messages is $10 \times T_{cons}$.

Fig. 2 is the algorithm flowchart of the delay test.
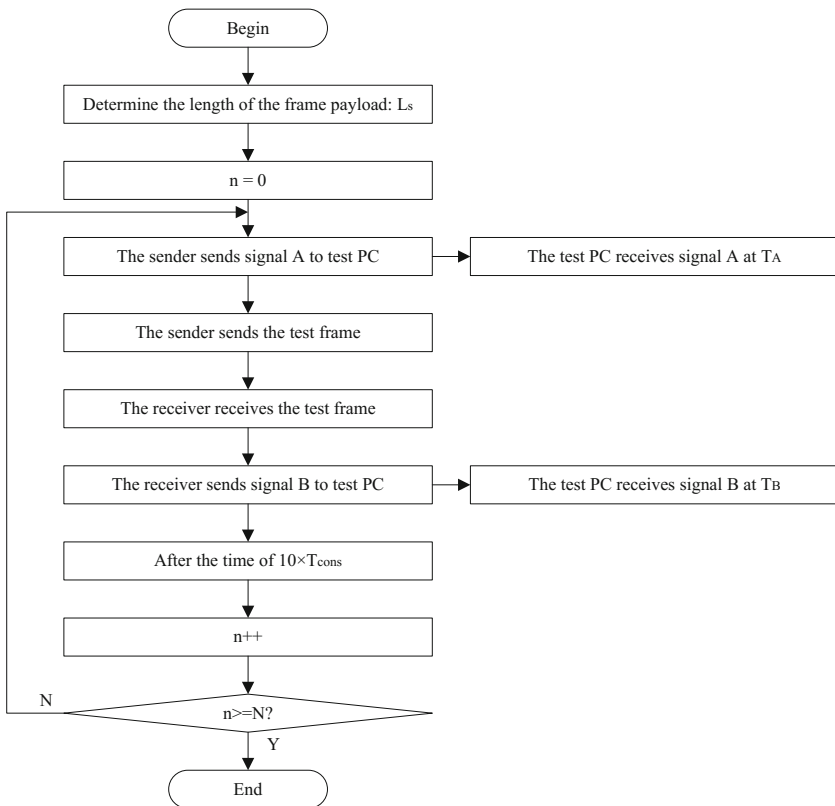
Fig. 2. The algorithm flowchart of delay test

We use T_A to denote the time that the test PC receives signal A and T_B to denote the time that the test PC receives signal B. The actual payload of each frame is L_S. The average delay time can be calculated as follow:

$$T_{aver}(L_s) = \frac{1}{N}\sum_{i=1}^{N}(T_B^i - T_A^i), i = 1, 2, 3, \ldots\ldots, N \tag{1}$$

The main source of error in the delay test is the time used to transmission sending and receiving signal between the end systems and the test PC, but this error is a constant value which is independent from the systems being tested and the payload length of test frames, so it is a fixed error which can be corrected.

3.2   Throughput Test

The throughput is the maximum volume of data traffic that transferred from the end system being tested without frame loss. The throughput reflects the communication capability of an end system, and is usually measured by the bytes or number of packets that transfer through the system per second. Here we choose the bytes per second as the throughput measurement. The development of avionics system asks for more network
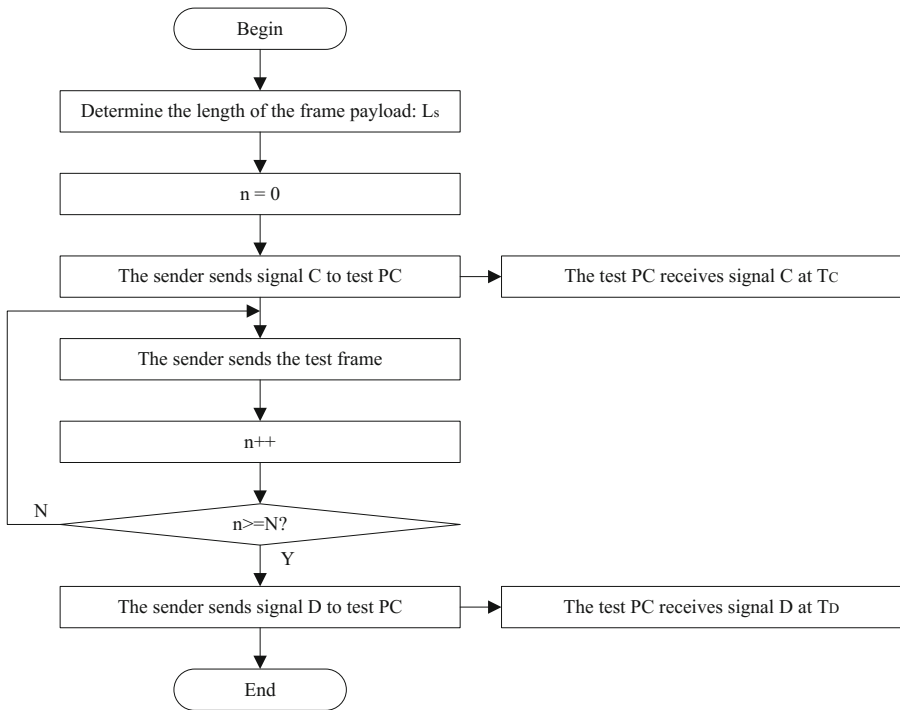
Fig. 3. The algorithm flowchart of throughput test

traffic to support the application service coming both from inside and outside of the aircraft. As a critical node in the avionics network, the network end system is the source or destination of the date transmissions. If the throughput of an end system is too small, the end system will be the bottleneck of the network data transmission and the communication efficiency of the whole network will be greatly diminished.
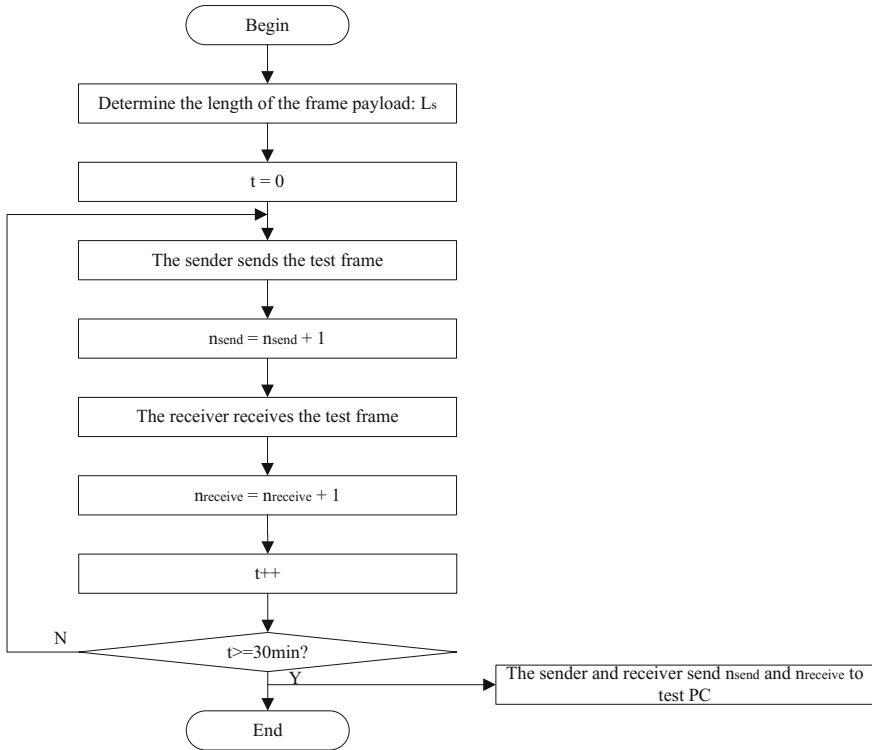
Fig. 4. The algorithm flowchart of frame loss rate test

A particular problem we need to address in the throughput test is that the maximum capability of an end system sending messages is hard to get. To solve these problems, we proposed the Maximum Send Rate method. We choose to measure the time of sending frames with a specify frame payload length L_S N times, and calculate the average time of sending a single frame. After that, we draw a figure which illustrates the number of Bytes the end system can send per second if the frame payload is L_S, and we can find the throughput of the end system from the figure.

Fig. 3 is the algorithm flowchart of the throughput test.

The average number of Bytes the end system can send per second if the frame payload length is L_S can be calculated with the following formula:

$$S = \frac{N \times L_s}{T_D - T_C} \tag{2}$$

3.3   Frame Loss Rate Test

The frame loss rate is the how much percentage of the frames that being lost during the transmission account for the total number of frames that needed to be transmitted. It reflects how much data load an end system can bear. If the frame loss rate is too large under some payload length, it will severely influence the whole performance of the implementation system.

To get the frame loss rate, a method called Continuous Data Transmission method is proposed. We use the test PC to control the end system to send frames of payload length L_S which lasted for *30 min*. A counter at the sender counts the frames being sent which is denoted by n_{send} and a counter at the receiver count the frames being received which is denoted by $n_{receive}$.

Fig. 4 is the algorithm flowchart of the frame loss rate test.

The frame loss rate calculation formula is as follow:

$$p(L_s) = \frac{n_{send} - n_{receive}}{n_{send}} \times 100\% \tag{3}$$

4   Experiments and Discussions

In this section, experimental results of performance test of FC-AE-ASM protocol and FC-AE-FCLP protocol implementation systems utilizing the proposed test methods mentioned above are presented. We developed the test platform utilizing Tornado 2.2, an IDE for real-time operation system Vxworks 5.5.1. The test PC running the software is an ordinary desktop with 1G frequency, 256M memory and Windows XP operation system. The FC upper layer protocols are implemented on FC interface
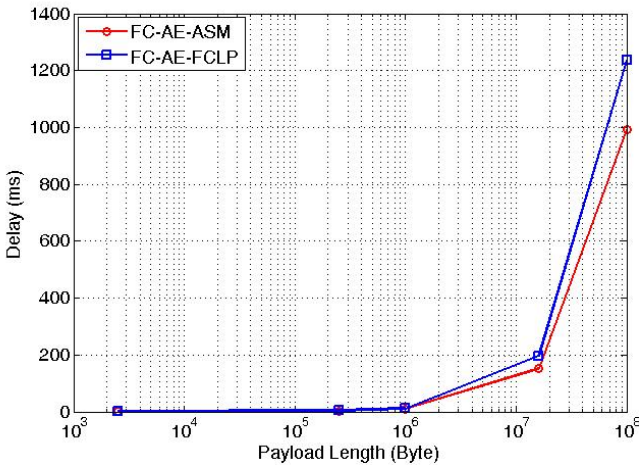
Fig. 5. The delay test result

cards, whose main component is a piece of Virtex-5 series FPGA produced by *Xilinx* Company. The IP cores are used to realize the functions of FC-0, FC-1 and part of FC-2. Another part of FC-2 functions are realized through VHDL logic programming. The Power PC embedded in the FPGA realized the functions of FC-4 and FC upper layer protocols. The link maximum transmission rate is 1Gbps.

Through Fig. 5, it is obvious that both of the FC upper layer protocols' implementation systems are of high efficiency in data transmission, with a delay of about 1s when transferring the message with a payload length of 100MB. When the message payload length is less than 1MB, the delay increases slightly with the increase of the payload length This is because that that time, the link utilization is relatively low, and the main causes of the delay are transmission delay and processing delay. However, when the message payload length is greater than 1MB, the delay increases enormously with the increase of payload length. This is because at this time, the link has been fully used, and the queuing delay increases enormously which lead to the increase of total delay. We also find that the FC-AE-ASM implementation system have better performance in the delay test. FC-AE-ASM protocol greatly simplify the communication process and have high link utilization, on the contrary, the complex process before the actual data transmission and the occurrence of large number of judging operations is the main drawbacks of FC-AE-FCLP protocol compared to FC-AE-ASM protocol. The advantages of FC-AE-ASM protocol can be more obvious in the environment of large amount of data transmission. When the payload length of the frames is 100MB, the delay of FC-AE-ASM implementation system is 19.8% less than the FC-AE-FCLP implementation system.

To decrease the delay of the implementation systems, we can improve the link transmission rate and optimize the underlying hardware and the logic program to improve the processing speed.
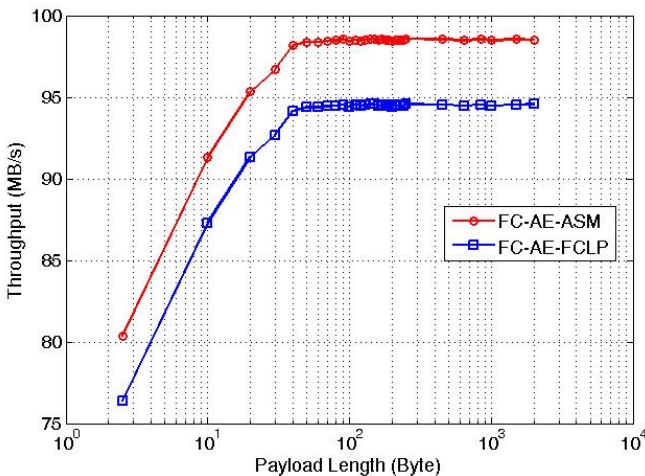
Fig. 6. The throughput test result

Fig. 6 shows the curves from which we can speculate the throughputs of the implementation systems. The theoretical value of the throughput is:

$$\frac{1Gbps}{8bit/byte} \times \frac{8byte}{10byte} = 102.4MB/s$$

However, from the figure, we can find that the throughput of FC-AE-ASM implementation system is 98.5 *MB/s,* and the throughput of FC-AE-FCLP implementation system is 94.5 *MB/s*. When the frame payload is greater than 20KB, the link transmission rate becomes the main factor that influences the throughput. The same reasons listed in the part of delay test result analysis make the FC-AE-FCLP a less efficient protocol in throughput than FC-AE-ASM protocol. The big gaps from the theoretical value and results are resulted from the following factors: the existence of FC-2 frame header, FC-AE-ASM frame header, the EOF, the SOF and the CRC, etc. Moreover, IDLE and other link overhead also causes this phenomenon. To improve the throughputs of the systems, we should improve the link transmission rate and the link utilization during the communication process.

Both of implementation systems have a frame loss rate of 0, which is because the topology from the sender to of receiver is rather simple. And this is also a proof that both of the FC upper layer protocols are of high reliability.

5   Conclusion

This paper studies the performance test methods of FC upper layer protocol implementation systems. The algorithms of delay test, throughput test and frame loss rate test are proposed. And we realized a test platform to test the performance of FC-AE-ASM protocol and FC-AE-FCLP protocol implementation systems. Our result proves the feasibility and accuracy of our algorithms. From the result, it is concluded that the FC-AE-ASM protocol have better performance than FC-AE-FCLP protocol in delay and throughput, which makes the FC-AE-ASM protocol a more suitable protocol for large amount of period data transmission in avionics networks.

References

1. Meng, Z., Tu, X., Xie, J.: Design and Implementation of Fiber Channel Interface Card on Avionics Environment Networks. In: International Conference on Computational Problem-Solving, pp. 399–401. IEEE Press, Lijian (2010)
2. INCTIS. T11/02-041v1. Fibre Channel - Avionics Environment. International Committee for Information Technology Standards, Washington (2002)
3. Leng, Y., Zong, Z., Liu, W.: Design and Implementation of FC-AE-1553 Point to Point Transmission Hardware Platform. In: IEEE Circuits and Systems International Conference on Testing and Diagnosis, pp. 1–4. IEEE Press, Chengdu (2009)
4. Zhu, X., Zhang, F., Wang, Y.: Study and Simulation of Fibre Channel-1553B network in Avionics System Using OPNET. In: 3rd International Symposium on Systems and Control in Aeronautics and Astronautics, pp. 456–461. IEEE Press, Harbin (2010)

5. Zhang, J., Shi, Y., Tian, H., Shi, G.: Applying DSPN (Deterministic and Stochastic Petri Nets) to Performance Analysis of FC-AE-1553 Integrated Avionics Bus System. Journal of Northwestern Polytechnical University 28(3), 388–392 (2010)
6. Zhen, X., Zhai, Z.: Study on Network Layer Test Method for FC-AE-ASM Protocol. Computer Measurement & Control 20(2), 324–327 (2012)
7. Liu, B., Zhang, J., Yang, Q., Li, D.: Modeling and Performance Analysis of FC-AE-ASM Which Base on PETRI Net Theory. In: International Conference on Computational Intelligence and Software Engineering, pp. 1–4. IEEE Press, Wuhan (2012)
8. Zhu, X., Zhang, F., Wang, Y., Bai, Y.: Simulation and Estimation of End-to-End Delay of Fiber Channel Network in Avionic System. Electronics Optics & Control 18(5), 54–58 (2011)
9. Niccolini, S., Molina, M., Raspall, F., Tartarelli, S.: Design and implementation of a one way delay passive measurement system. In: IEEE/IFIP Network Operations and Management Symposium, vol. 1, pp. 469–482 (2004)
10. INCITS. T11/08-013v1. Fibre Channel Avionics Environment - Anonymous Subscriber Messaging (ASM)/Amendment 1. International Committee for Information Technology Standards, Washington (2008)

Attribute Reduction Algorithm Based on the Simplified Discernibility Matrix of Granularity

Zhangyan Xu, Xiaoyu Wang, and Wei Zhang

Guangxi Normal University, College of Computer Science &
Information Engineering, 541004 Guilin, China
`{Xiaoyu Wang 1,527866254}@qq.com`

Abstract. This paper gives the definition of discernibility matrix of granularity in decision table and the corresponding attribute reduction and the definition of discernibility matrix of granularity in the simplified decision table and the corresponding attribute reduction. Therefore, verifying the equivalence of the definition of attribute reduction in the simplified decision table and the definition of attribute reduction based on relative granularity of decision table. On the basis of the above theories, a new algorithm is designed based on knowledge granulation for attribute reduction in simplified decision table, and the corresponding complexity of time is reduced to $O(|C|^2|U'_{pos}||U|)$. Finally, an example is given to illustrate the validity of the new algorithm.

Keywords: Rough set, Knowledge granularity, discernibility matrix of granularity, Attribute reduction.

1 Introduction

Knowledge reduction is one of the core contents in rough set theory [1].However, for all reduction or minimum reduction had been proved to be NP-Hard problem [2]. Therefore, seek the efficient and fast attribute reduction in rough set theory is one of the core areas of study. Liu Qing put forward a rough set theory based on granularity model [3], and constructed the granulation according to equivalence relation of the attribute entropy value. Y.Y.Yao etc studied a series of basic problems and methods for granulation [4,5,6]. Li Xiuhong etc proposed main concept of knowledge granularity and operation method, constructed attribute significance and designed algorithm for attribute reduction from the knowledge perspective [7,8]. The algorithms only dicussed attribute reduction for information systems through concept and the nature of the knowledge granularity, and no further discussion.

It is a good method for discernibility matrix that is designed the attribute reduction algorithm.Thus this paper gives the definition of discernibility matrix based on granularity and the corresponding attribute reduction.The definition of the simplified decision table size difference matrix, this verifies the equivalence of the definition and discernibility matrix of granularity of decision table. On that basis, design a new algorithm based on knowledge granulation for attribute reduction in decision table, its

complexity of time is reduced to $O(|C|^2|U'_{pos}||U|)$. Finally,an example is given to illustrate the validity of the new algorithm.

2   Concepts of Discernibility Matrix Based on Granularity

In this section,we introduce the basic concepts , propose new definitions of discernibility matrix and attribute reduction based on granularity.

Definition 1[7]. Let $S = (U,A,V,f)$ is information systems, marked $S = (U,A)$ where are:

U is a nonempty finite set of objects;

$A = C \cup D$ and $C \cap D = \emptyset$ is a nonempty finite set of attribute, C is the set of condition attributes, D is the set of decision attributes;

$V = \bigcup\limits_{a \in C \cup D} V_a$, V_a denotes   the   value   domain   of a ;

$f : U \times A \to V$,for $\forall a \in C \cup D$,there is a mapping $f(x,a) \in V_a :\; U \to V_a$,where V_a is called the value set of a .

Each subset of attributes $P \subseteq A$ determines a binary indiscernibility relation

$\text{IND}(P)$, as follows

$IND(P) = \{(x,y) \in U \times U \mid \forall a \in A, f(x,a) = f(y,a)\}$

It is easily shown that $\text{IND}(P)$ is an equivalence relation on the set U.

The relation $\text{IND}(P)$, $P \subseteq A$, constitutes a partition of U, which we will denote by $U/\text{IND}(P)$.

Definition 2. Let a decision table $S = (U,C,D,V,f)$ discernibility matrix $M = (m(i,j))$ is defined as:

$$m(i,j) = \begin{cases} \{a \mid a \in C, f(x_i,a) \neq f(x_j,a) \wedge x_i, x_j \in U_{pos}; \\ f(x_i,a) \neq f(x_j,a), f(x_i,D) \neq f(x_j,D) \wedge x_i \in U_{pos} \wedge x_j \in U_{neg}; \\ f(x_i,a) \neq f(x_j,a), f(x_i,D) \neq f(x_j,D) \wedge x_i \in U_{neg} \wedge x_j \in U_{pos}; \\ \emptyset, else \end{cases}$$

Definition 3. Let a decision table $S = (U,C,D,V,f)$, M is the discernibility matrix of granularity,for any subset $B \subseteq C$, if $\forall \emptyset \neq m(i,j) \in M \Rightarrow B \cap m(i,j) \neq \emptyset$;and for any $a \in B$,there at least exists $\emptyset \neq m(i_0,j_0) \in M$ such that $B \cap m(i_0,j_0) = \emptyset$,then B is called an atrribute reducetion of C on D based on discernibility matrix of granularity.

Definition 4. For decision table $S = (U,C,D,V,f)$, denote $U/C = \{[u'_1]_C, [u'_2]_C, \cdots, [u'_m]_C\}$, $U' = \{u'_1, u'_2, \cdots, u'_m\}$,and according to theorem1, assume $POS_C(D) = [u'_{i_1}]_C \cup [u'_{i_2}]_C \cup \cdots \cup [u'_{i_t}]_C$

where $\{u'_{i_1}, u'_{i_2}, \cdots, u'_{i_t}\} \subseteq U'$ and $|[x'_{i_s}]_C / D| = 1 (s = 1, 2, \cdots, t)$,   denote

$U'_{pos} = \{u'_{i_1}, u'_{i_2}, \cdots, u'_{i_t}\}$, $U'_{neg} = U - pos_c(D)$, $U" = U'_{pos} \cup U'_{neg}$,then

$S' = (U", C, D, V, f)$ is called the simplified decision table.

Definition 5. Let a decision table $S = (U, C, D, V, f)$, $S' = (U", C, D, V, f)$ is the simplified decision table ,the simplified discernibility matrix of granularity $M' = (m(i', j'))$ is defined as:

$$m(i', j') = \begin{cases} \{a | a \in C, f(x'_i, a) \neq f(x'_j, a) \wedge x'_i, x'_j \in U'_{pos}; \\ f(x'_i, a) \neq f(x'_j, a), f(x'_i, D) \neq f(x'_j, D) \wedge x'_i \in U'_{pos} \wedge x'_j \in U'_{neg}; \\ f(x'_i, a) \neq f(x'_j, a), f(x'_i, D) \neq f(x'_j, D) \wedge x'_i \in U'_{neg} \wedge x'_j \in U'_{pos}; \\ \varnothing, else \end{cases}$$

Definition 6. Let a decision table $S' = (U", C, D, V, f)$, M' is the simplified discernibility matrix of granularity, for any subset $B \subseteq C$, if $\forall \varnothing \neq m(i', j') \in M \Rightarrow B \cap m(i', j') \neq \varnothing$;and for any $a \in B$,there at least exists $\varnothing \neq m(i'_0, j'_0) \in M$ such that $B \cap m(i'_0, j'_0) = \varnothing$,then B is called an atrribute reducetion of C on D based on discernibility matrix of granularity.

3   Related Theorem Based on Biscernibility Matrix of Granularity

In this section,we provide and proof theorems for discernibility matrix of granularity,and prove this new attribute reduction definition and reduction of relative granularity are equal to each other.

Definition 7. Let a decision table $S = (U, A)$,$U / IND(A) = \{X_1, X_2, \cdots, X_n\}$, then A of   knowledge   granularity   is   called $GK(A)$,and   definition $GK(A) = \sum_{i=1}^{n} |X_i|^2 / |U|^2$。 $\sum_{i=1}^{n} |X_i|^2$ is $\overset{n}{\underset{i=1}{U}}(X_i \times X_i)$ composed of equivalence relation determined cardinal.

Theorem1[8].   Let $S = (U, A)$ be an information system, $P 、 Q \subseteq A$ and $\varnothing \subset X \subseteq U$. If $IND(P) = IND(Q)$ then $GK(P) = GK(Q)$.

Theorem 2. Let a decision table $S = (U, C, D, V, f)$, M is the discernibility matrix of granularity, for anysubset $B \subseteq C$, if $\forall \varnothing \neq m(i, j) \in M \Rightarrow B \cap m(i, j) \neq \varnothing$, then $GK(B) = GK(C)$.

Proof: Suppose $GK(B) \neq GK(C)$,there at least $x_i \in U$ exists, which makes $[x_i]_B \neq [x_i]_C$, then there certainly exists $x_j \in [x_i]_B \wedge x_j \notin [x_i]_C$,and $c_k \in C - B$,which makes $f(x_i, c_k) \neq f(x_j, c_k)$,

when $x_i, x_j \in U_{POS}$,according to the definition 3, there is $c_k \in m(i,j) \Rightarrow m(i,j) \neq \varnothing$ and $x_j \in [x_i]_B$,there is $B \cap m(i,j) = \varnothing$,it conflicts with the precondition, so the hypothesis does not hold. We have $GK(D|B) = GK(D|C)$.

when $x_i \in U_{POS} \wedge x_j \in U_{neg}$ and $f(x_i, D) \neq f(x_j, D)$, according to the definition 3, $c_k \in m(i,j) \Rightarrow m(i,j) \neq \varnothing$ and $x_j \in [x_i]_B$,there is $B \cap m(i,j) = \varnothing$, it conflicts with the condition.

when $x_i \in U_{POS} \wedge x_j \in U_{neg}$ and $f(x_i, D) = f(x_j, D)$,since $x_j \in U_{neg}$, then there certainly exists x_r ,which makes $f(x_i, D) \neq f(x_r, D)$ according to the definition 3, $c_k \in m(i,j) \Rightarrow m(i,j) \neq \varnothing$ and $x_j \in [x_i]_B$,there is $B \cap m(i,j) = \varnothing$, it conflicts with the condition.

when $x_j \in U_{POS} \wedge x_i \in U_{neg}$ and $f(x_i, D) \neq f(x_j, D)$, this case is the same as (2).

when $x_j \in U_{POS} \wedge x_i \in U_{neg}$ and $f(x_i, D) = f(x_j, D)$, this case is the same as (3).

Remark. According to the above discussion,all the cases are not true, so the hypothesis does not hold.Thus we have $GK(B) = GK(C)$.

Theorem 3. Let a decision table $S = (U, C, D, V, f)$, M is the discernibility matrix of granularity, for any subset $B \subseteq C$, if $GK(B) = GK(C)$, then $\forall \varnothing \neq m(i,j) \in M \Rightarrow B \cap m(i,j) \neq \varnothing$.

Proof: Suppose $\forall \varnothing \neq m(i,j) \in M \Rightarrow B \cap m(i,j) = \varnothing$,there   exists $c_k \in C \wedge c_k \notin B$,which makes $c_k \in m(i,j)$, there is $[x_i]_B \neq [x_i]_C$ .Since $B \cap m(i,j) = \varnothing$, according to the definition 4, there is $[x_i]_B = [x_i]_C$ .On the other hand ,since $[x_i]_C \subseteq [x_i]_B \wedge [x_j]_C \subseteq [x_j]_B$,there is $[x_i]_C \cup [x_j]_C \subseteq [x_i]_B \cup [x_j]_B = [x_i]_B$, hence $[x_i]_C \neq [x_i]_B$ and $GK(B) \neq GK(C)$,it conflicts with the condition, so the hypothesis does not hold. Thus we have $\forall \varnothing \neq m(i,j) \in M \Rightarrow B \cap m(i,j) \neq \varnothing$.

Theorem 4. Attribute reduction definitions based on relative granularity and attribute reduction definitions based on discernibility matrix of granularity is equivalence .

Proof: let $RG\operatorname{Re}duce$ and $GDis\operatorname{Re}duce$ be sets of attribute reduction based on relative granularity and the discernibility matrix of granularity, for any $GDis\operatorname{Re}duce$,we have $\forall \varnothing \neq m(i,j) \in M$ and $B \cap m(i,j) \neq \varnothing$ .According to the theorem 2, $GK(D|B) = GK(D|C)$. Since $B \in GDis\operatorname{Re}duce$, for $\forall a \in B$

and $P = B - \{a\}$, there exists $\forall \varnothing \neq m(i_0, j_0) \in M$ and $B \cap m(i_0, j_0) = \varnothing$ at least . According to the negative conversd proposition of theorem 3,for $\forall a \in B$, we have $GK(D \mid B - \{a\}) \neq GK(D \mid C)$,so $B \in RG \operatorname{Re} duce$ .Since B is randomly selected,we have $GDis \operatorname{Re} duce \subseteq RG \operatorname{Re} duce$.

For any $B \in RG \operatorname{Re} duce$,we have $GK(D \mid B) = GK(D \mid C)$. According to the theorem 3, we have $\forall \varnothing \neq m(i_0, j_0) \in M$ and $B \cap m(i_0, j_0) = \varnothing$ .As $B \in RG \operatorname{Re} duce$,then for $\forall a \in B$,there is $GK(D \mid B - \{a\}) \neq GK(D \mid C)$. According to the negative conversd proposition of theorem 2, we can draw the conclusion that there is $B \in GDis \operatorname{Re} duce$,since B is randomly selected ,we have $GDis \operatorname{Re} duce \supseteq RG \operatorname{Re} duce$.

Remark. According to the above two aspects,we can draw the conclusion that $GDis \operatorname{Re} duce$ and $RG \operatorname{Re} duce$ are equialent.That is to say,the proposition is true.

Theorem 5. Let a simplified decision table $S' = (U", C, D, V, f)$, M' is the discernibility matrix of granularity, for any subset $B \subseteq C$, if $\forall \varnothing \neq m(i', j') \in M' \Rightarrow B \cap m(i', j') \neq \varnothing$,then $GK(B) = GK(C)$.

Proof: the same as Theorem3.

Theorem 6. Let a simplified decision table $S' = (U", C, D, V, f)$, M' is the discernibility matrix of granularity, for any subset $B \subseteq C$, if $GK(B) = GK(C)$, then $\forall \varnothing \neq m(i', j') \in M' \Rightarrow B \cap m(i', j') \neq \varnothing$.

Proof: the same as Theorem4.

Theorem 7. Atrribute reduction based on relative granularity is the same as that based on the simplified discernibility matrix of granularity.

Proof: let $RG \operatorname{Re} duce$ and $SGDis \operatorname{Re} duce$ be sets of attribute reduction based on relative granularity and the simplified discernibility matrix of granularity, for any $SGDis \operatorname{Re} duce$,we have $\forall \varnothing \neq m(i', j') \in M' \Rightarrow B \cap m(i', j') \neq \varnothing$ .According to the theorem 5, $GK(D \mid B) = GK(D \mid C)$. Since $B \in SGDis \operatorname{Re} duce$, for $\forall a \in B$ and $P = B - \{a\}$, there exists $\forall \varnothing \neq m(i'_0, j'_0) \in M \Rightarrow B \cap m(i'_0, j'_0) = \varnothing$ at least . According to the negative conversd proposition of theorem 6, for $\forall a \in B$, we have $GK(D \mid B - \{a\}) \neq GK(D \mid C)$,so $B \in RG \operatorname{Re} duce$ .Since B is randomly selected,we have $SGDis \operatorname{Re} duce \subseteq RG \operatorname{Re} duce$.

For any $B \in RG \operatorname{Re} duce$, we have $GK(D \mid B) = GK(D \mid C)$. According to the theorem 6, we have $\forall \varnothing \neq m(i', j') \in M' \Rightarrow B \cap m(i', j') \neq \varnothing$ .As $B \in RG \operatorname{Re} duce$,then for $\forall a \in B$,there is $GK(D \mid B - \{a\}) \neq GK(D \mid C)$. According to the negative conversd proposition of theorem 5,we can draw the conclusion that there at least exists $\forall \varnothing \neq m(i_0, j_0) \in M \Rightarrow B \cap m(i_0, j_0) = \varnothing$, so $B \in SGDis \operatorname{Re} duce$, since B is randomly selected , we have $SGDis \operatorname{Re} duce \supseteq RG \operatorname{Re} duce$.

Remark. According to the above two aspects,we can draw the conclusion that *SGDis* Re *duce* and *RG* Re *duce* are equialent. That is to say,the proposition is true.

4   Attribute Reduction Algorithm Based on the Simplified Discernibility Matrix of Granularity

In this section,we propose a new attribute reduction algorithm based on the simplified discernibility matrix of granularity, and analyze the time complexity of corresponding algorithm

Algorithm 1. Acorrding to the algorithm of literature [10], we can sort $U = \{x_1, x_2, x_3, \cdots, x_n\}$ as $U = \{u_1', u_2', \cdots, u_n'\}$.

Input： a decision table $S = (U, A), U = \{u_1, u_2, \cdots, u_n\}$, $P \subseteq A$ and $P = \{p_1, p_2, \cdots, p_s\}$

Output： U/P

For each $P_i(i = 1, 2, \cdots, s)$ statistic $f(u_j, p_i)(j = 1, 2, \cdots, n)$ of the maximum and minimum values , respectively M_i and m_i ;

In order to static linked list storage object $u_1, u_2, \cdots, u_n$;the head pointer u_1 ;

For(i=1;i<s+1;i++)

3．1 The i trip "distribution": create $M_i - m_i + 1$ empty queue, respectively $front_k$ and $end_k (k = 0, 1, 2, \cdots, M_i - m_i)$ are K head pointer of the queue and the tail pointer. The object $u \in U$ in the list in order to distribute the queue $f(u, p_i) - m_i$ in the list .

3．2 The i trip "collection": header pointer point to the first nonempty queue head pointer, modify every nonempty queue tail pointer, let it points to the next non-empty queue right object,and let $M_i - m_i + 1$ queue to form a linked list.

The third step consists of object sequence list for $u_1', u_2', \cdots, u_n'$;

$t = 1 ; B_t = \{u_1'\}$;

for(j = 2 ; j < n +1 ; j++)

{ for $(i = 1; i <= s; i++)$

$if (f(u_j', p_i) == f(u_{j-1}', p_i))$ $\{B_t = B_t \cup \{u_j'\};\}$

else $t = t + 1$;

$B_t = \{u_j'\}$;

Analysis of Algorithm1.   The complexity of Algorithm1 is $O(|C||U|)$ [10].

Algorithm 2. Algorithm of simplified decision table

Input:Decision table $S = (U, C, D, V, f)$

Output: U'_{pos}、 U'_{neg}

Step1 ： Calculate $U = \{x_1, x_2, x_3, \cdots, x_n\}$ get the list of object sequence $u'_1, u'_2, \cdots, u'_n$ according to Algorithm 1;

Step2 ： t=1; $B_t = \{u'_1\}$;

Step3 ： for (j=2;j<n+1;j++)

if any $c_i \in C(i=1,2,\cdots,k)$,there is $f(u'_j, c_i) = f(u'_{j-1}, c_i)$, then $B_t = B_t \cup \{u'_j\}$;

else {t=t+1; $B_t = \{u'_j\}$;}

step4 ： $U'_{pos} = \varnothing; U'_{neg} = \varnothing$;

for (i=1;i<t+1;i++)

if any $x, y \in B_i$ there is $f(x, D) = f(y, D)$,then we take out the first object of B_i to U'_{pos} ;else we take out all objects of B_i to U'_{neg} ;

Analysis of Algorithm2.The complexity of Step1 is $O(|C\|U|)$ .The complexity of Step2 is $O(|C\|U|)$ .The complexity of Step3 is $O(|C\|U|)$ .So the complexity of Algorithm2 is $O(|C\|U|)$.

Algorithm 3. attribute reduction algorithm based on the simplified discernibility matrix of granularity

Input: Decision table $S = (U, C, D, V, f)$

Output: $reduce(C)$

Step1 ： Calculate $U'_{pos} = \{u'_{i_1}, u'_{i_2}, \cdots, u'_{i_i}\}$ and U'_{neg} according to Algorithm 2;

Step2 ： $reduce(C) = \varnothing$, P=$\varnothing$;

Step3 ： for(i=1;i<s;i++)

for(j=i+1;j<s+1;j++)

if(f(y$_i$',D)≠f(y$_j$',D)){m(i,j)= $\varnothing$;

for(k=1;k<r+1;k++)

if(f(y$_i$',c$_k$)≠f(y$_j$',c$_k$)) m(i,j)= m(i,j)∪{c$_k$};}

Step4 ： for(i=1;i<s+1;i++)

for(j=1;j<t+1;j++){m(i,j)= $\varnothing$;

for(k=1;k<r+1;k++)

if(f(y$_i$',c$_k$)≠f(y$_j$',c$_k$)∧f(y$_i$',D)≠f(y$_j$',D))m(i,j)= m(i,j)∪{c$_k$};}

Step5 ： While(M=(m(i,j))!= $\varnothing$)

1) Judge every element in the matrix, if the element is only one attribute $\{a\}$, then we merge it into $reduce(C)$, and remove all elements that contain the attribute in the matrix;

2) take out any attribute $\{b\}$ in the $C-P-reduce(C)$ and merge it into P , then remove all elements that contain the attribute in the matrix;}

Step6：output $reduce(C)$.

Analysis of Algorithm 3. The complexity of Step1 is $O(|C\|U|)$. The complexity of Step3 and Step4 is $O(|C\|U'_{pos}\|(U'_{pos}+U'_{neg})|)$. The worst time complexity of each pass in the loop of Step5 is $O(|C\|U'_{pos}\|(U'_{pos}+U'_{neg})|)$, the most cycle several times is $|C|-1$, the worst time complexity of Step5 is $O(|C|^2|U'_{pos}\|U|)$. So the complexity of Algorithm3 is $O(|C|^2|U'_{pos}\|U|)$.

5   Example

In this section,we set a example and illustrate the new algorithm.

Table 1. Information table.

U	a	b	c	d
X_1	2	1	2	1
X_2	2	2	2	1
X_3	2	2	2	0
X_4	0	0	0	0
X_5	1	0	1	0
X_6	2	0	1	1
X_7	2	0	1	1
X_8	1	0	1	0

$U/\{a,b,c\}=\{\{X_1\},\{X_2,X_3\},\{X_4\},\{X_5,X_8\},\{X_6,X_7\}\}$,
$U'_{pos}=\{X_1,X_4,X_5,X_6\},U'_{neg}=\{X_2,X_3\}$

Table 2. The simplified information table

U	a	b	c	d
X_1	2	1	2	1
X_2	2	2	2	1
X_3	2	2	2	0
X_4	0	0	0	0
X_5	1	0	1	0
X_6	2	0	1	1
X_7	2	0	1	1

Remark. According algorithm 3, the first step of the first cycle can get core attribute $\{a\}$, remove core attribute $\{a\}$ in the discernibility matrix, if all elements is not all empty in the discernibility matrix, $reduce(C) = \{a\}$;

$C - P - reduce(C) = \{b,c\}$, take any attribute in this set and remove it that contains the attribute in the discernibility matrix. Then judge whether if it is empty for every element of discernibility matrix, if all elements is not all empty, we should continue the cycle. Now, it is all empty and the attribute reduction of decision table is $\{a,b\}$

Table 3. The simplified discernibility matrix of granularity

	X_1	X_4	X_5	X_6
X_1		a,b	a,b,c	b,c
X_2		∅	∅	∅
X_3		a,b	b,c	∅
X_4		∅	a,c	a,c
X_5			∅	a
X_6				∅

6   Conclusion

In complete information systems, it is discussed discernibility matrix of granularity for decision tables and simplified decision tables. Reveal that the granularity of knowledge change reason and nature in the paper, on that basis, put forward attribute reduction algorithm of the simplified decision table. The algorithm reduces the complexity, and it is convenient to practical application.

We will discuss binary discernibility matrix based on knowledge granularity and its attribute reduction, and attribute reduction of information entropy for further study.

References

1. Pawlak, Z.: Rough Sets. International Journal of Computer and Imforation Science 11, 341–356 (1982)
2. Zhangyan, X., Zuopeng, L., Pingru, Y.: A quick attribute reduction algorithm which complexity is max($O(|C||U|),O|C||U / C|$). Int. Journal of Geographical Sciences 29, 391–399 (2006)
3. Duoqian, M., Guoyin, W., Qing, L.: Granular computing: past, present and future. In: Abadi, M. (ed.), Science Press, Beijing (2007)

4. Yao, Y.Y.: Granular computing: Basics and possible solutions. In: Proceedings of the 5th International Conference on Computing and Information, vol. 1, pp. 186–189 (2000)
5. Yao, Y.Y.: Granular computing: Past, present and future. In: 2008 IEEE International Conference on Granular Computing, pp. 80–85 (2008)
6. Yao, Y.Y.: The rise of granular computing. Int. Journal of Chongqing University of Posts and Telecommunications: Natural Science Edition 20, 299–308 (2008)
7. Xiuhong, L., Kaiquan, S.: A attribute reduction method of information system based on knowledge granularity. Int. J. Computer Application 26, 76–77 (2006)
8. Xiuhong, L.: Knowledge Granularity of Rough Set Concept and Operation. Int. J. Computer Engineering and Applications 47, 34–36 (2011)
9. Zhangyan, X., Pingru, Y., Wei, S.: A quick algorithm of computing core based on discernibility matrix. Int. Journal of Geographical Sciences 42, 4–5 (2006)
10. Zhangyan, X.: A quick algorithm of computing core based on positive region. Int. J. Systems Engineering and Electronics 28, 1902–1905 (2006)

Efficient Inference about the Partially Linear Varying Coefficient Model with Random Effect for Longitudinal Data

Wanbin Li[1,2] and Liugen Xue[1]

[1] College of Applied Sciences, Beijing University of Technology, P.R. China, 100124
lgxue@bjut.edu.cn
[2] Department of Mathematics, Yancheng Teachers University, Jiangsu, P.R. China, 224002
lwb@emails.bjut.edu.cn

Abstract. This article considers the estimation for the partially linear varying coefficient model with random effect. This article proposes to use the estimators for the variance component and profile weighted semi-parametric least squares (WSLSE) techniques to estimate the parametric component efficiently and to establish its efficiency and asymptotic properties. The proposed estimator is proved to be more efficient than that naïve estimation with working independence error structure. Some Monte Carlo simulations are conducted to examine the finite sample performance. Moreover, from the empirical study, the newly proposed procedure performs well in moderate-sized samples.

Keywords: longitudinal data, random effect, efficient estimation.

1   Introduction

Longitudinal or clustered data is a common phenomenon in modern data analysis. For example, in social surveys, we know that individual respondents that are collected from the same subject at different times are correlated and that observations from different subjects are often independent. Other examples can be found in econometrics, biology, and so on. Mixed effect models are widely used statistical tools to deal with such kind of data for considering the within-individual correlation.

Parametric mixed effect regression model generally has simple and intuitive interpretations and provides a parsimonious description of the relationship between the response variable and its covariates. However, since that the strong assumption models may induce modeling biases and lead to erroneous conclusions, semi-parametric and nonparametric estimation techniques have attracted much attention among statisticians and econometricians. [1] proposed the nonparametric model with random effects.

$$Y_{it} = \theta(U_{it}) + \alpha_i + v_{it}, i = 1, \cdots n, t = 1, \cdots m,$$

where U_{it} is a $q \times 1$ random variable vector, α_i is $i.i.d.$ $(0, \sigma_\alpha^2)$. v_{it} is $i.i.d.$ $(0, \sigma_v^2)$, α_i and v_{it} are uncorrelated for all $i = 1, \cdots n$ and $\theta(U_{it})$ is an unknown smooth function. However, when the dimension of X is high, [1]'s proposed model suffers from the "curse of dimensionality". Furthermore, to enhance the model's interpretability, we shall consider the generalization of this nonparametric model with random effect.

Another motivation of this paper is the application to the CD4 count data. Many kinds of literatures have focused on the analysis of this dataset with varying coefficient models, see, for example, [2], and [3]. However, none of them has proposed a unified procedure by considering the traditional inference and the within-subject correlation. In fact, we can't reject the null hypothesis significantly when testing for the within-subject effect in the real data application. Therefore, as a natural and necessary extension, and simultaneously taking the within-subject correlation into account, we investigate the following partially linear varying coefficient model (PLVC) with random effect

$$Y_{ij} = X_{ij}^T \beta + Z_{ij}^T g(u_{ij}) + \varepsilon_{ij} \text{ and } \varepsilon_{ij} = \alpha_i + v_{ij}, i = 1, \cdots, n, j = 1, \cdots, m , \text{ (1)}$$

where the nonparametric component $g(\cdot)$ is assumed to be a p-dimension smooth function, Y and X are one dimension and q dimension response variable and covariates vector, respectively. Despite containing [1]'s proposed model as a special case, the generalization has the following two main advantages. Firstly, the proposed model (1) provides a more powerful interpretational capability because of containing the parametric component. Secondly, our proposed model can deal with the multivariate covariant variables without encountering the curse of dimensionality.

In the literature, the partially linear varying coefficient model is an important tool to analyze the cross-sectional data and longitudinal data. Since proposed by [4], the PLVC models have been extensively studied by [5] and [6] for cross-sectional data. And later extended to longitudinal studies by [2], [7], and among others. We consider PLVC models for longitudinal data due to their good balance between flexibility and parsimony. The focus of this paper is to conduct statistical inference with considering the within-individual subject correlation.

There is extensive research focused on modeling longitudinal data. However, literature on the application of varying coefficient models with random effect is limited. Specifically, existing methods are mainly confined by the complexity of the likelihood function and assumption of a working likelihood. Inferential methods based on consistent estimators for the variance component have been developed by a number of researchers, see, for example, [1], [8] and among others. [8] explored the efficient inference for the partially linear mixed effect model, where they proposed an inference procedure to obtain more efficient estimators for the parametric component. [9] developed an estimation for the single-index models with random effect.

To model various kinds of data, we extend the standard nonparametric model to the partially linear varying coefficient model with random effect, and propose to solve the problem of estimating the parametric and variance component simultaneously. There are some advantages about our estimators obtained in this paper, which can be

summarized below. Firstly, the whole procedure is data-driven and can automatically take advantage of the within-subject correlation. It does not need to specify any assumption about the working likelihood. Secondly, the proposed estimation procedure obtains a $\sqrt{n}$ -consistent estimator for the parametric and variance component. Further, the estimators for the variance component, obtained in this paper, allow us to make further statistical inference, such as the construction of the confidence interval for the nonparametric component.

The rest of this paper is organized as follows. In section 2, we present the framework of estimating method and the theoretical results about the estimators are established in section 3. Simulation results are reported in section 4, while an empirical application of the proposed model to the data of the Multi-Center AIDS Cohort study is reported in section 5. The technical proofs are relegated to the Appendix.

2   Estimation Method

2.1   The Preliminary Estimators

Let $b(t) = (b_1(t), \cdots, b_{k_n+\hbar+1}(t))^T$ be a set of B-spline basis functions of order $\hbar+1$ with quasi-uniform internal knots and denote another coefficient vector $\theta_l = (\theta_{l,1}, \cdots, \theta_{l,k_n+\hbar+1})^T$. With the B-spline basis, model (1) can be approximated by

$$Y_{i,j} \approx \sum_{l=1}^{p} \sum_{s=1}^{k_n+\hbar+1} Z_{ij,l}^T b_s^T(u_{ij})\theta_{l,s} + \sum_{s=1}^{q} X_{ij,l}^T \beta_s + \varepsilon_{ij} = B_{ij}^T \Theta + X_{ij}^T \beta + \varepsilon_{ij},$$

where $B_i = I_p \otimes b(U_i) \cdot Z_i, \Theta = (\theta_{l,s}) \in \Re^{p_{k_n}}$, with $p_{k_n} = p(k_n + \hbar + 1)$ and

$b_{ij} = b(u_{ij})$. For details on the construction of B-spline basis functions, the readers are referred to [10].

Therefore, by the linear model theory, we can obtain the preliminary estimator

$$\hat{\beta} = (X^T M_B X)^{-1} X^T M_B Y \quad \text{and} \quad \hat{\Theta} = (B^T B)^{-1} B^T (Y - X\hat{\beta}), \tag{2}$$

where $M_B = I_n - P_B = I_n - B(B^T B)^{-1} B^T$, $X = (X_{11}, \cdots, X_{1m}, \cdots, X_{nm})^T$,

$B = (B(U_{11}), \cdots, B(U_{1m}), \cdots, B(U_{nm}))^T$ and $Y = (Y_{11}, \cdots, Y_{1m}, \cdots, Y_{nm})^T$.

Moreover, the corresponding B-spline estimator of $g(u)$ is

$$\hat{g}(u) = B^T(u)\hat{\Theta}.$$

2.2   Estimators for the Variance Components

In this section, we consider the consistent estimators for the variance component, which reflect the within-subject covariance. Let $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \cdots, \varepsilon_{im})^T$, $(i=1,\ldots,n)$ is an $m \times 1$ random vector. Thus, the variance-covariance matrix of $\boldsymbol{\varepsilon}_i$ is defined as

$$\Omega_i = \sigma_\alpha^2 \mathbf{1}_m \mathbf{1}_m^T + \sigma_\nu^2 \mathbf{I}_m, \tag{3}$$

where $\mathbf{1}_m$ represents an $m \times 1$ vector of ones and $\mathbf{I}_m$ is the $m \times m$ identity matrix and denote that $\Omega = diag(\Omega_i)$. In fact, $E(\varepsilon_{i1}\varepsilon_{i2}) = \sigma_\alpha^2$ and $E(\varepsilon_{11}^2) = \sigma_\alpha^2 + \sigma_\nu^2$. Hence, we can obtain the moment estimators for the variance component of Ω_i as

$$\hat{\sigma}_\alpha^2 = \frac{1}{nm(m-1)} \sum_{i=1}^{n} \sum_{j_1=1}^{m} \sum_{j_1 \neq j_2} \hat{\varepsilon}_{ij_1}\hat{\varepsilon}_{ij_2} \quad \text{and} \quad \hat{\sigma}_\nu^2 = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j_1=1}^{m} \hat{\varepsilon}_{ij_1}^2 - \hat{\sigma}_\alpha^2. \tag{4}$$

where $\hat{\varepsilon}_{ij}$ is defined as $\hat{\varepsilon}_{ij} = Y_{ij} - X_{ij}^T \hat{\boldsymbol{\beta}} - Z_{ij}^T \hat{g}(u_{ij}), i=1,\cdots,n, j=1,\cdots m.$

Similar idea for this estimator has been used in several literatures. See, for example, [8], [9] and among others.

2.3   Weighted Semi-parametric Least Squares Estimation (WSLSE)

When plugging in the estimator for the variance component in (3), we can obtain the estimator $\hat{\Omega}$ for the covariance matrices. Therefore, the weighted estimation function can be defined as follows

$$S^W(\boldsymbol{\beta}, \boldsymbol{\theta}) = (Y - X^T \boldsymbol{\beta} - B^T \boldsymbol{\Theta})^T \hat{\Omega}^{-1} (Y - X^T \boldsymbol{\beta} - B^T \boldsymbol{\Theta}),$$

by which we can obtain the estimator

$$\hat{\boldsymbol{\beta}}^W = (X^T M_B^{\hat{\Omega}^{-1}} X)^{-1} X^T M_B^{\hat{\Omega}^{-1}} Y, \tag{5}$$

where $M_B^{\hat{\Omega}^{-1}} = \hat{\Omega}^{-1} - P_B^{\hat{\Omega}^{-1}} = \hat{\Omega}^{-1} - \hat{\Omega}^{-1} B (B^T \hat{\Omega}^{-1} B) B^T \hat{\Omega}^{-1}.$

Consequently, $\hat{\boldsymbol{\beta}}^W$ are called the WSLSE of $\boldsymbol{\beta}$. With the estimator

$$\hat{\Theta}^W = (B^T \hat{\Omega}^{-1} B) B^T \hat{\Omega}^{-1} (Y - X^T \hat{\boldsymbol{\beta}}^W),$$

we solve the corresponding weighted nonparametric estimator $\hat{g}(u)$ as $B^T(u)\hat{\Theta}^W$.

3   Main Results

Similar to the definition of the varying coefficient class and discussion about the projection on to the varying coefficient space in [13], we also assume that $X_{i,l} = H_l^*(Z_i, U_i) + \Pi_{i,l}$, where $H_l^*(Z_i, U_i)$ is the projection function of $m_l(Z_i, U_i)$ onto the varying coefficient space, where

$$m_l(z, u) = E(X_l \mid Z = z, U = u), \qquad l = 1, \cdots, q.$$

Next, after introducing some notations, we make some conditions to derive the theoretical results. Let β_0 and $g_0(u)$ be the true value of the corresponding parameters β and $\theta(u)$. Denote $f(\cdot)$ to be the density for covariant variable U_{ij}. And some regularity conditions are listed as follows.

(C1) The function $g(\cdot)$ is r th continuously differentiable on $(0,1)$, where $r > 2$.

(C2) The sequence of design points $\{U_{ij}, i = 1, \cdots, n, j = 1, \cdots, m\}$ has a bounded support U and is generated by a "design density" $f(\cdot)$ with

$$0 < \inf_U f(\cdot) \le \sup_U f(\cdot) < 1.$$

(C3) $\Pi_i = (\Pi_{i1}, \cdots, \Pi_{im})^T, i = 1, \cdots, n$ are independent random matrices with mean zero and $E \mid \Pi_i \mid \le c_0 < \infty,$ where $\Pi_{ij} = (\pi_{ij1}, \cdots, \pi_{ijp})^T, j = 1, \cdots, m$ and $\mid \cdot \mid$ denotes Euclidean norm. Further, Π_i is independent of $\varepsilon_i = (\varepsilon_{i1}, \cdots, \varepsilon_{im})^T$.

(C4) For all i and j, the random vectors X_{ij}, Z_{ij} are bounded in probability.

Next theorems reveal the asymptotic properties about the estimators in section 2.

Theorem 1. Suppose that conditions (C1)--(C4) hold, then

$$\sqrt{n}(\hat{\sigma}_\alpha^2 - \sigma_\alpha^2) \to_D N(0, V_\alpha), \quad \text{and} \quad \sqrt{nm}(\hat{\sigma}_v^2 - \sigma_v^2) \to_D N(0, V_v), \quad (6)$$

where $V_\alpha = \mathbf{Var}(\alpha_1^2) + \dfrac{4\sigma_\alpha^2 \sigma_v^2}{m} + \dfrac{2\sigma_v^4}{m(m-1)}$ and $V_v = \mathbf{Var}(v_{11}^2) + \dfrac{2\sigma_v^4}{m-1}$.

Theorem 2. Under the assumptions of Theorem 1, $\sqrt{n}(\hat{\beta}^W - \beta) \to_D N(0, \Sigma_3^{-1})$ as $n \to \infty$, where $\Sigma_3 = (1/n)\Sigma_{i=1}^n (\Pi_i \Omega_i^{-1} \Pi_i^T).$

Remark 1. According to the asymptotic variance of the estimator $\hat{\beta}^W$, Σ_3^{-1} can achieve the semi-parametric information bound (see [11]), which is similar to the discussion in [8]. Therefore, the proposed WSLSE is semi-parametric efficient.

4   Empirical Studies

In this section, we conduct some Monte Carlo simulations to evaluate the finite sample performance of the proposed method. According to the model (1), the data is generated from the following partially linear varying coefficient model

$$y_{ij} = x_{ij1}\beta_1 + x_{ij2}\beta_2 + z_{ij}g(u_{ij}) + \alpha_i + v_{ij}, i = 1,\cdots,n, j = 1\cdots m, \qquad (7)$$

where $x_{ij1} = \zeta_{i1} + \xi_{ij1}, x_{ij2} = \zeta_{i2} + \xi_{ij2}, z_{ij} = \zeta_{i3} + \xi_{ij3}$ and $\zeta_{i1}, \xi_{ij1},\ \zeta_{i2},,$ $\zeta_{ij2},$

$\zeta_{i3},\ \xi_{i3}$ are $i.i.d.$ N(0,0.25), u_{ij} is $i.i.d.$ U(0,1), α_i is $N(0,\sigma_\mu^2)$ and v_{ij} is $N(0,\sigma_v^2)$. Assume that (β_1,β_1) takes (1.5, 2), $g(t) = 2\sin(2\pi t)$. $(\sigma_\mu^2,\sigma_v^2)$ takes (2, 0.5) and (2, 1). Moreover, the number of n and m_i are chosen as follows:

Case 1: $n = 80, m_1 = \cdots = m_{50} = 2$ and $m_{51} = \cdots = m_{80} = 3.$

Case 2: $n = 40, m_1 = \cdots = m_{25} = 4$ and $m_{26} = \cdots = m_{40} = 6.$

The simulation will be performed with 1000 times according to the proposed procedure in this paper. In the simulation, we compare three estimators for the parametric components (β_1,β_2) : the (unweighed) semi-parametric least squares estimator $\hat{\beta}$, the WSLSE $\hat{\beta}^W$ with estimated σ_μ^2 and σ_v^2 and the WSLSE $\tilde{\beta}^W$ with known σ_μ^2 and σ_v^2. Table 1 and 2 summary the simulation results about the estimates of (β_1,β_2) and $(\sigma_\mu^2,\sigma_v^2)$ respectively, including the sample mean (Mean), the sample standard error (SE) for the estimators. From Table 1 and Table 2, we can find:

(1) Generally speaking, all three methods are asymptotic unbiased, because there is little observable bias in the means of the estimators.

(2) The performance of our two weighted estimators are superior to the unweighed one.

(3) Not only the weighted estimated sample means are closer to the true value, but also they all have evident smaller standard deviations than the unweighed one.

(4) For the error estimates, the simulated values of the Mean and SE perform well.

Moreover, we compare the nonparametric performance of with the others. Given the confidence degree 95%, Figure 1 provides us with different estimated confidence interval (CI), which is shown in each figure of figure 1 with different colors. These CI

is obtained by averaged the B-spline estimates for the nonparametric function g over 1000 simulations under the four case. And in each figures, the red, purple and green dashed line indicate the unweighed ,weighted with estimated weight and weighted with known weight estimators, respectively.

From figure 1, we can find that both the unweighed estimator and weighted estimators with estimated weight can estimate the true CI well, but the latter have a higher accuracy, especially in Figure 1 (c and d). In other words, the weighted estimator is superior to the un-weighed one, because of the consideration of the error structure. What's more, when compared the weighted estimators with the estimated weight and true weight, our methods also performs well, especially in the longitudinal data with more observation times in each object, which also show the importance of the correlation between objects in longitudinal data.

5   Real Application

We illustrate the WSLSE proposed in this paper by analyzing a data set from the Multi-Center AIDS Cohort study. The dataset contains the human immunodeficiency

Table 1. Simulation results for the estimators of the parametric components β_1 and β_2

(σ_μ, σ_ν)	methods		β_1		β_2	
			Mean	SE	Mean	SE
$(\sqrt{2}, \sqrt{0.5})$	Case1	$\hat{\beta}$	1.5022	0.0675	2.0003	0.0653
		$\hat{\beta}^w$	1.5017	0.0497	1.9997	0.0491
		$\tilde{\beta}^w$	1.5017	0.0428	1.9999	0.0414
	Case2	$\hat{\beta}$	1.4989	0.0851	2.0022	0.0841
		$\hat{\beta}^w$	1.4997	0.0370	2.0010	0.0366
		$\tilde{\beta}^w$	1.5005	0.0243	2.0000	0.0245
$(\sqrt{2}, 1)$	Case1	$\hat{\beta}$	1.4954	0.0725	2.0006	0.0698
		$\hat{\beta}^w$	1.4964	0.0589	2.0003	0.0566
		$\tilde{\beta}^w$	1.4977	0.0530	1.9997	0.0511
	Case2	$\hat{\beta}$	1.5003	0.0913	1.9987	0.0919
		$\hat{\beta}^w$	1.5005	0.0556	1.9981	0.0582
		$\tilde{\beta}^w$	1.5006	0.0486	1.9984	0.0496

Table 2. Simulation results for the estimators of thevariance components σ_μ^2 and σ_ν^2

(σ_μ, σ_ν)		Mean	SE	Mean	SE
$(\sqrt{2}, \sqrt{0.5})$	Case1	0.7709	0.2725	1.1662	0.2443
	Case2	1.4736	0.5963	1.8463	0.5057
$(\sqrt{2}, 1)$	Case1	0.7723	0.2958	1.5420	0.2749
	Case2	1.4683	0.6298	2.4241	0.5364
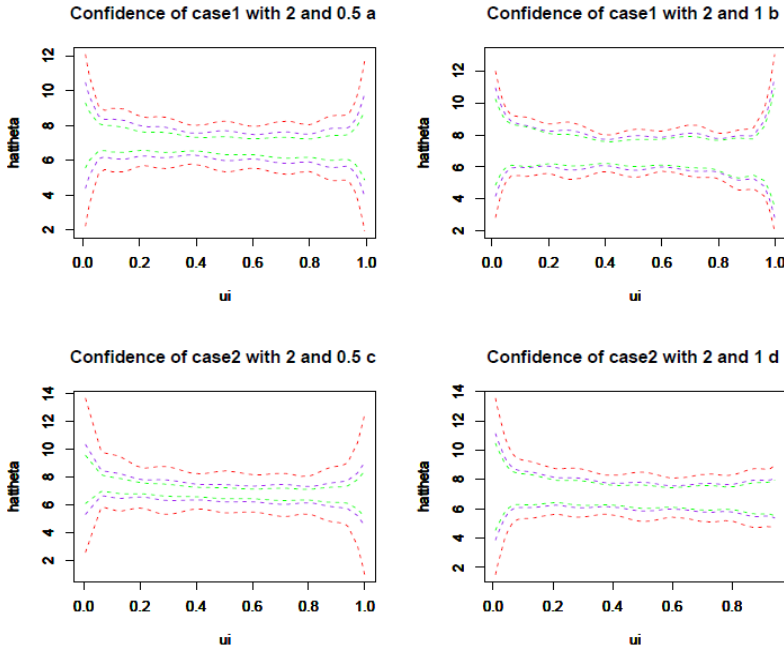
Fig. 1. Confidence bands for the first nonzero function in the four different cases. And the red, purple and green dashed lines indicate the unweighed, weighted with estimated weight and weighted with known weight estimators, respectively.

virus (HIV) status of 283 homosexual males who were infected with HIV during a follow-up period between 1984 and 1991. Each patient had a different number of repeated measurements and different measurements. Details of the experimental design can be found in [12]. Several researchers including [2], [3] and [13] have studied the same data set to analyze the mean CD4 percentage.

Our analysis mainly tries to model the trend of the mean CD4 percentage depletion over time. [2] indicates that neither smoking nor age has a significant impact on the mean CD4 percentage with a p-value of 0.059, while [14] claimed that the effect of PreCD4 is significant time-varying with a p-value of 0.045. [3] found that the mean baseline CD4 percentage decreases very quickly at the beginning of HIV infection, and the rate of decrease somewhat slows down four years after infection. However, there is limited work of them that has considered the within-subject covariance within the random effect

framework. In fact, when we used the Hausman test for the null hypothesis of the random effect, the random effect was approved at 5% level of significance with P-value 9%, which can be calculated by the R function phtest() in the plm package.

In this application, we construct the data as follows. Let Y be the individual's CD4 percentage, X_1 be the centered preCD4 percentage, X_1^2 that represents the quadratic effects of the centered preCD4 percentage, Z_1 be the centered age at HIV infection, and $Z_2 = Z_1^2$ that represents the quadratic effects of the centered age. For the purpose of demonstration and simplicity, the possible effects of other available covariates are omitted. Then, we consider the following model

$$Y = \theta_0(t) + Z_1\theta_1(t) + Z_2\theta_2(t) + X_1\beta_1 + X_2\beta_2 + \varepsilon \tag{8}$$

to fit the data, where the baseline CD4 percentage $\theta_0(t)$, represents the mean CD4 percentage t years after the infection. The data were collected from different objects at different time. That is each subject will be measured several times at serial time. All these features suggest that when analyzing these data, we can't ignore the error structure among each object and between different subjects. So these considerations motivate us to apply the weighted estimators, proposed by this paper, to model (8). Assume that $\varepsilon = (\varepsilon_{11}, \cdots, \varepsilon_{1m}, \cdots \varepsilon_{nm})$ and $\varepsilon_{ij} = u_{ij} + v_{ij}$, $i = 1, \cdots n$, $j = 1, \cdots m$.

By the proposed estimation procedure, the curve of the estimated baseline function is show in Fig 2. Fig 2 shows us that the mean baseline CD4 percentage will drop fast at the beginning of HIV infection, and the drop rate will slows down after the four years' fast decrease. This conclusion is similar to [2] and [3].
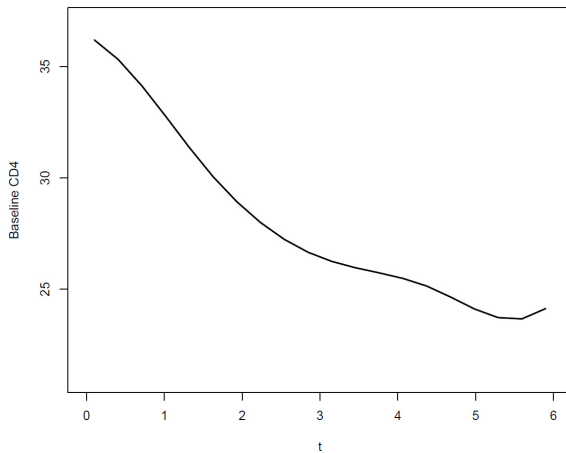
Fig. 2. Application to AIDS data. The estimators for the mean CD4 percentage $\theta_0(t)$.

6 Appendix

In this section, we show the proof about the asymptotic properties in section 3. Our proofs use a strategy to prove the asymptotic result, which is similar to that in [8].

Firstly, let

$$\tilde{\beta}^W = (X^T M_B^{\Omega^{-1}} X)^{-1} X^T M_B^{\Omega^{-1}} Y, \tilde{\Theta}^W = (B^T \Omega^{-1} B)^{-1} B^T \Omega^{-1} (Y - X^T \tilde{\beta}^W),$$

$$\tilde{g}^W(u) = B^T(u) \tilde{\Theta}^W, \text{ and } M_B^{\Omega^{-1}} = \Omega^{-1} - P_B^{\Omega^{-1}} = \Omega^{-1} - \Omega^{-1} B (B^T \Omega^{-1} B)^{-1} B^T \Omega^{-1}.$$

Lemma 1. Under the conditions for Theorem 1,

$$\sqrt{n}(\tilde{\beta}^W - \beta) \to_D N(0, \Sigma_3^{-1})$$

where Σ_3 is defined in Theorem 2.

Proof. Assume that $\Pi_{ijs} = X_{ijs} - H_s^*(u_{ij}, z_{ij}), (i = 1, \cdots, n; j = 1, \cdots, m; s = 1, \cdots, q)$.

Firstly, the sth element of $X^T M_B^{\Omega^{-1}} \varepsilon$ can be decomposed as

$$X_s^{*T} M_B^{\Omega^{-1}} \varepsilon = \Pi_s^{*T} \Omega^{-1} \varepsilon - \Pi_s^{*T} \Omega^{-1} B (B^T \Omega^{-1} B)^{-1} B^T \Omega^{-1} \varepsilon + H_s^{*T} M_B^{\Omega^{-1}} \varepsilon,$$

where $X_s^* = (X_{11,s}, X_{12,s}, \cdots, X_{nm,s})^T, H_s^* = (H_s(U_{11}, Z_{11}), \cdots, H_s(U_{nm}, Z_{nm}))^T.$

It is easy to very that $\Pi_s^{*T} \Omega^{-1} B (B^T \Omega^{-1} B)^{-1} B^T \Omega^{-1} \varepsilon = o_p(n^{1/2}),$

$X_s^{*T} M_B^{\Omega^{-1}} G = o(n^{1/2})$

and $X^T M_B^{\Omega^{-1}} \varepsilon = \Pi \Omega^{-1} \varepsilon + o_p(n^{1/2})$. Hence,

$$\sqrt{n}(\tilde{\beta}^W - \beta) = \sqrt{n}(\Pi^T \Omega^{-1} \Pi)^{-1} \Pi^T \Omega^{-1} \varepsilon + o_p(1).$$

And, Lemma 1 follows from the central limit theorem.

Proof of Theorem 1. According to the estimator (4), $\hat{\sigma}_\alpha^2$ can be written as

$$\hat{\sigma}_\alpha^2 = (1/(nm(m-1))) \sum_{i=1}^n \sum_{j_1=1}^m \sum_{j_2 \neq j_1} \hat{\sigma}_{ij_1} \hat{\sigma}_{ij_2} = I_1 + I_2 + I_3,$$

where $I_1 = (1/(nm(m-1))) \sum_{i=1}^n \sum_{j_1=1}^m \sum_{j_2 \neq j_1} \varepsilon_{ij_1} \varepsilon_{ij_2},$

$I_2 = (2/(nm(m-1))) \Sigma_{i=1}^n \Sigma_{j_1=1}^m \Sigma_{j_1 \neq j_2} \{[X_{ij_1}^T \beta - Z_{ij_1}^T g(u_{ij_1})] -$

$[X_{ij_1}^T \tilde{\beta} - Z_{ij_1}^T \tilde{g}(u_{ij_1})]\} \varepsilon_{ij_2}$

$I_3 = (1/(nm(m-1))) \Sigma_{i=1}^n \Sigma_{j_1=1}^m \Sigma_{j_1 \neq j_2} \{[X_{ij_1}^T \beta - Z_{ij_1}^T g(u_{ij_1})] -$

$[X_{ij_1}^T \tilde{\beta} - Z_{ij_1}^T \tilde{g}(u_{ij_1})]\} \{[X_{ij_2}^T \beta - Z_{ij_2}^T g(u_{ij_2})] - [X_{ij_2}^T \tilde{\beta} - Z_{ij_2}^T \tilde{g}(u_{ij_2})]\}$

Using Taylor formula, we have that $I_1 = o_p(n^{-(1/2)}), I_3 = o_p(n^{-(1/2)}).$

By

$$I_1 - \sigma_\alpha^2 = \frac{1}{n} \sum_{i=1}^n (\alpha_i^2 - \sigma_\alpha^2) + \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m \alpha_i v_{ij} \frac{1}{nm(m-1)} \sum_{i=1}^n \sum_{j_1=1}^m \sum_{j_2 \neq j_1} v_{ij_1} v_{ij_2},$$

we have

$$\sqrt{n}(\hat{\sigma}_{\alpha}^2 - \sigma_{\alpha}^2) = (1/\sqrt{n}) \sum_{i=1}^{n} V_i + o_p(1),$$

where V_i are independent variables with mean zero. Therefore, after denoting

$B_n^2 = \sum_{i=1}^{n} E(V_i^2)$, Theorem 1 follows from the proof of Theorem 4 in [8].

Proof of Theorem 2. After denoting that $G(u_{ij}) = Z_{ij}^T g(U_{ij})$, for all $1 \leq i \leq n$, $1 \leq j \leq m$, the remaining proof can be adopted from the proof of   Lemma 1.

References

1. Su, L., Ullah, A.: More efficient estimation of nonparametric panel data models with random effects. Economics Letters 96, 375–380 (2007)
2. Huang, J.Z., Wu, C.O., Zhou, L.: Varying-Coefficient Models and Basis Function Approximations for the Analysis of Repeated Measurements. Biometrika 89, 111–128 (2002)
3. Zhao, P.X., Xue, L.G.: Variable Selection for Semiparametric Varying Coefficient Partially Linear Errors-in-variable Models. J. Multivariate Anal. 101, 1872–1883 (2010)
4. Hastie, T., Tibshirani, R.: Varying Coefficient Models. J. Roy. Statist. Soc. Ser. B 55, 757–796 (1993)
5. Ahmad, I., Leelahanon, S., Li, Q.: Efficient. Estimation of a Semiparametric Partially Linear Varying Coefficient Model. The Annals of Statistics 33, 258–283 (2005)
6. Fan, J., Huang, T.: Profile Likelihood Inferences on Semiparametric Varying Coefficient Partially Linear Models. Bernoulli 11, 1031–1057 (2005)
7. Chiang, C.T., Rice, J.A., Wu, O.C.: Smoothing Spine Estimation for Varying Coefficient Models with Repeatedly Measured Dependent Variables. J. Am. Statist. Assoc. 96, 454 (2001)
8. You, J., Zhou, X.: Partially Linear Models and Polynomial Spline Approximations for the Analysis of Unbalanced Panel Data. J. Stat. Plan Infer. 139, 679–695 (2009)
9. Pang, Z., Xue, L.G.: Estimation for the Single-Index Models with Random Effects. Comput. Stat. Data An. 56, 1837–1853 (2012)
10. de Boor, C.: A Practical Guide to Spines. Springer, New York (1978)
11. Bickel, P.J., Klaassen, C.A., Ritov, Y., Wellner, J.A.: Efficient and Adaptive Estimation for Semi-parametric Models. Johns Hopkins, Series in the Mathematical Science. Johns Hopkins University Press, Baltimore (1993)
12. Kaslow, R.A., Ostrow, D.G., Detels, R., Phair, J.P., Polk, B.F., Rinaldo, C.R.: The multicenter AIDS cohort study: Rationale, organization and selected characteristics of the participants. American Journal of Epidemiology 126, 310–318 (1987)
13. Fan, J., Huang, T., Li, R.: Analysis of longitudinal data with semi-parametric estimation of covariance function. J. Amer. Statist. Assoc. 102, 632–641 (2007)
14. Qu, A., Li, R.: Quadratic inference functions for varying coefficient models with longitudinal data. Biometrics 62, 379–391 (2006)

Particular Solutions of a Class of Nonlinear Reaction-Diffusion Equations

Hongxue Chu[1,2] and Tongsong Jiang[1]

[1] Chool of Science, Linyi University, Linyi, 276005, P. R. China
[2] School of Mathematical Sciences, Shandong Normal University, Jinan, 250014, P.R. China
jiangtongsong@sina.com, chuhongxue2012@163.com

Abstract. In this paper, we propose the Method of Particular Solutions for Solving a Class of Nonlinear Reaction-Diffusion Equations using collocation points and approximating the solution using multiquadrics (MQ) and the Thin Plate Splines (TPS) Radial Basis Function (RBF). The scheme works in a similar fashion as finite-difference methods. The results of numerical experiments are presented, and are compared with analytical solutions to confirm the good accuracy of the presented scheme.

Keywords: Nonlinear Reaction-Diffusion Equations; The Method of Particular Solutions(MPS); Radial Basis Function (RBF); Multiquadrics (MQ); Thin Plate Splines (TPS).

1   Introduction

This paper is devoted to the numerical computation of a Class of Nonlinear Reaction-Diffusion Equations

$$u_t = \varepsilon \Delta u + \varepsilon^{-1} u^2 \left(1 - u\right). \tag{1}$$

in some continuous domain with suitable initial and Dirichlet boundary conditions

Finite difference methods [1] are known as the first techniques for solving the partial differential equations. Even though these methods are very effective for solving the partial differential equations, conditional stability of explicit finite difference procedures and the need to use large amount of CPU time in implicit finite difference schemes limit the applicability of these methods. Furthermore, these methods provide the solution of the problem on mesh points only and accuracy of the techniques is reduced in nonsmooth and nonregular domains. To avoid the mesh generation, meshless techniques have attracted the attention of researchers in recent years,. In a meshless (meshfree) method [2-4] a set of scattered nodes are used instead of meshing the domain of the problem. Meanwhile, there are several meshless methods using Radial Basis Functions (RBFs). Some meshless schemes are the radial function methods [5-7] the collocation method [8,9], the method of fundamental

solutions [10], and the method of particular solutions [11,12]. The Equations are computed with the Strang Splitting method [13].

The layout of the paper is as follows. In Section 2 we show how we use the radial basis functions to approximate the solution. In Section 3 we apply the method on the nonlinear reaction-diffusion equation. The results of numerical experiments are presented in Section 4. Section 5 is dedicated to a brief conclusion. Finally some references are introduced at the end.

2   Radial Basis Function Approximation

The approximation of a distribution $u(\mathbf{x})$, using radial basis functions, maybe written as a linear combination of N radial functions; usually it takes the following form:

$$u(\mathbf{x}) \approx \sum_{j=1}^{N} \lambda_j \varphi(\mathbf{x}, \mathbf{x}_j) + \psi(\mathbf{x}), \qquad \text{for } \mathbf{x} \in \Omega \subset R^d . \tag{2}$$

Where N is the number of data points, $\mathbf{x} = (x_1, x_2, \cdots, x_d)$, d is the dimension of the problem, the λ's are coefficients to be determined and φ is the radial basis function (2) can be written without the additional polynomial ψ.

In that case φ must be unconditionally positive definite to guarantee the solvability of the resulting system (e.g. Gaussian or inverse multiquadrics). However, ψ is usually required when φ is conditionally positive definite, i.e., when φ has a polynomial growth towards infinity. Examples are thin plate splines and multiquadrics. We will use multiquadrics and thin plate splines for the numerical scheme introduced in Section 3. These are defined as

(i) Multiquadrics(MQ)

$$\varphi(\mathbf{x}, \mathbf{x}_j) = \varphi(r_j) = \sqrt{r_j^2 + c^2} . \tag{3}$$

(ii) Generalized Thin Plate Splines (TPS)

$$\varphi(\mathbf{x}, \mathbf{x}_j) = \varphi(r_j) = r_j^{2m} \log(r_j), \qquad m = 1, 2, 3, \ldots . \tag{4}$$

Where c is the shape parameter of the radial basis function and $r_j = \| \mathbf{x} - \mathbf{x}_j \|$ is the Euclidean norm. Optimal shape values are found experimentally and these values are written for the problem.

If $\mathcal{P}_q^d$ denotes the space of d-variate polynomials of order not exceeding than

q, and letting the polynomials $P_1, \cdots, P_m$ be the basis of $\mathcal{P}_q^d$ in R^d, then the polynomial $\psi(x)$, in (2), is usually written in the following form

$$\psi(\mathbf{x}) = \sum_{i=1}^{m} \zeta_i P_i(\mathbf{x}).$$

(5)

Where $m = (q-1+d)! / (d!(q-1)!)$.

To determine the coefficients $(\lambda_1, \cdots, \lambda_N)$ and $(\zeta_1, \cdots \zeta_m)$, the collocation method is used. However, in addition to the N equations resulting from collocating (2) at the N points, an extra M equations are required. This is ensured by the M conditions for (2),

$$\sum_{j=1}^{N} \lambda_j P_i(\mathbf{x}_j) = 0, \qquad i = 1, \cdots, m.$$

(6)

In a similar representation as (2), for any linear partial differential operator $\mathcal{L}$, $\mathcal{L}u$ can be approximated by

$$\mathcal{L}u(\mathbf{x}) \approx \sum_{j=1}^{N} \lambda_J \mathcal{L}\varphi(\mathbf{x}, \mathbf{x}_j) + \mathcal{L}\psi(\mathbf{x}).$$

(7)

3   Nonlinear Reaction-Diffusion Equation

Let us consider the following Nonlinear Reaction-Diffusion Equations

$$\frac{\partial u}{\partial t} = \varepsilon \Delta u + \varepsilon^{-1} u^2 (1-u) \qquad (x, y) \in \Omega \subset R^2, \ t \in [0, T].$$

(8)

with the initial condition

$$u(x, y, 0) = u_0(x, y), \qquad (x, y) \in \Omega.$$

(9)

and Dirichlet boundary conditions

$$u(x, y, t) = g(x, y), \qquad (x, y) \in \partial\Omega, \quad t \in (0, T].$$

(10)

where Δ is the Laplace differential operator, $u_0(x, y)$ and $g(t)$ are known functions, and the function $u(x, y, t)$ is unknown.

First, let us discretize (8) according to the following θ-weighted scheme

$$\frac{u(x, y, t+\delta t) - u(x, y, t)}{\delta t} = \theta \varepsilon \Delta u(x, y, t+\delta t)$$
$$+(1-\theta)\varepsilon \Delta u(x, y, t+\delta t) + \varepsilon^{-1}(u^2(x, y, t) - u^3(x, y, t)).$$

(11)

Where $0 \leq \theta \leq 1$ and $\delta t = t^{l+1} - t^l$ is the time step and $t^l = l\delta t$, $l \geq 0$ is the time at l step. Rearranging (11), using the notation $u^l(x, y) = u(x, y, t^l)$, we obtain

$$u^{l+1}(x, y) - \delta t \theta \varepsilon \Delta u^{l+1}(x, y) =$$
$$u^{l+1}(x, y) + \delta t(1 - \theta)\varepsilon \Delta u^l(x, y) + \delta t \varepsilon^{-1} u^{2l}(x, y)(1 - u^l(x, y)). \tag{12}$$

Rearranging (12), we obtain

$$\Delta u^{l+1}(x, y) = \frac{u^{l+1} - u^l}{\delta t \theta \varepsilon} - \frac{1 - \theta}{\theta} \Delta u^l - \frac{\varepsilon^{-2}}{\theta}((u^2)^l - (u^3)^l). \tag{13}$$

Assuming that $u^{l+1}(\mathbf{x})$ is the solution of the equation, we can set the right half part of (13) is $F(x, y)$, This means (13) is a standard Poisson equation

$$\Delta u^{l+1}(x, y) = F(x, y). \tag{14}$$

Assuming that there are a total of $(N - 3)$ interpolation points, $F(x, y)$ can be approximated by

$$F(x, y) \approx \sum_{j=1}^{N-3} \lambda_j^{l+1} \phi_j(r_j) + \lambda_{N-2}^{l+1} x + \lambda_{N-1}^{l+1} y + \lambda_N^{l+1}. \tag{15}$$

To determine the interpolation coefficients $(\lambda_1, \lambda_2, \cdots \lambda_{N-2}, \lambda_{N-1}, \lambda_N)$, the collocation method is used by applying (15), at every point (x_i, y_i), $i = 1, 2, \cdots, N - 3$. Thus we have

$$F(x_i, y_i) \approx \sum_{j=1}^{N-3} \lambda_j^{l+1} \phi_j(r_{ij}) + \lambda_{N-2}^{l+1} x_i + \lambda_{N-1}^{l+1} y_i + \lambda_N^{l+1}. \tag{16}$$

where $r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. The additional conditions due to (6) are written as

$$\sum_{j=1}^{N-3} \lambda_j^{l+1} = \sum_{j=1}^{N-3} \lambda_j^{l+1} x_j = \sum_{j=1}^{N-3} \lambda_j^{l+1} y_j = 0. \tag{17}$$

Thus, we approximate $u^{l+1}(x, y)$ has the following form

$$u^{l+1}(x, y) \approx \sum_{j=1}^{n} \lambda_j^{l+1} \Phi_j(r_j). \tag{18}$$

where $\Phi(x, y)$ is obtaine d by an alytically solving

$$\Delta\Phi_j(x, y) = \phi_j(x, y).$$

(19)

Writing (18) together with (17) in a matrix form we have

$$[u]^l = \mathbf{A}[\lambda]^l.$$

(20)

where

$$\mathbf{A} = \begin{bmatrix} \Phi_{11} & \cdots & \Phi_{1N-3} & x_1 & y_1 & 1 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \Phi_{(N-3)1} & \cdots & \Phi_{(N-3)(N-3)} & x_{N-3} & y_{N-3} & 1 \\ x_1 & \cdots & x_{(N-3)} & 0 & 0 & 0 \\ y_1 & \cdots & y_{(N-3)} & 0 & 0 & 0 \\ 1 & \cdots & 1 & 0 & 0 & 0 \end{bmatrix}.$$

(21)

Assuming that there are $k < (N-3)$ internal points and $(N-3-k)$ boundary points, then the $(N \times N)$ matrix $\mathbf{A}$ can be split into $\mathbf{A} = \mathbf{A}_d + \mathbf{A}_b + \mathbf{A}_e$, where

$$\begin{aligned} \mathbf{A}_d &= \begin{bmatrix} a_{ij} & for\,(1 \le i \le k, 1 \le j \le L) & and & 0 & elsewhere \end{bmatrix} \\ \mathbf{A}_b &= \begin{bmatrix} a_{ij} & for\,(k+1 \le i \le N-3, 1 \le j \le N) & and & 0 & elsewhere \end{bmatrix} \\ \mathbf{A}_e &= \begin{bmatrix} a_{ij} & for\,(N-2 \le i \le N, 1 \le j \le N) & and & 0 & elsewhere \end{bmatrix} \end{aligned}$$

(22)

Using the notation $\mathcal{L}\,\mathbf{A}$ to designate the matrix of the same dimension as $\mathbf{A}$ and containing the elements $\tilde{a}_{ij}$, where $\tilde{a}_{ij} = \mathcal{L}a_{ij}$, $1 \le i, j \le N$ then (14) together with (10) can be written, in matrix form, as

$$\mathbf{B}[\lambda]^{l+1} = \mathbf{C}[\lambda]^l + \delta t \varepsilon^{-1} u^l .* u^l * (1 - \mathbf{A}_d[\lambda]^l) + [\mathbf{G}]^{l+1}.$$

(23)

where

$$\mathbf{B} = \mathbf{A}_d - \delta t \theta \varepsilon \Delta \mathbf{A}_d + \mathbf{A}_b + \mathbf{A}_e,$$
$$\mathbf{C} = \mathbf{A}_d + \delta t(1-\theta)\varepsilon \Delta \mathbf{A}_d,$$
$$[\mathbf{G}]^l = [0 \cdots 0\, \mathbf{G}_{k+1}^l \cdots \mathbf{G}_{N-3}^l\, 0\, 0\, 0].$$

Thus, the solution of the complex system has been reduced to solving the real variable system. Since the coefficient matrix is unchanged in time steps, we use the **LU** factorization to the coefficient matrix only once and use this factorization in our algorithm.

Remark. Although (23) is valid for any value of $\theta \in [0,1]$, we will use $\theta = 1/2$ (the famous Crank–Nicolson scheme).

4   Numerical Examples

In order to study the validity and effectiveness of the numerical method mentioned in the previous sections, we give one example. In addition, we use L_∞, L_2 error and the root-mean-aquare error (RMSE)[13] to measure the difference between the numerical and analytical solutions.

The L_∞ error norm defined as

$$L_\infty = \|u^{exact} - u_N\|_\infty = max_j |u_j^{exact} - (u_N)_j|.$$

The L_2 error norm defined as

$$L_2 = \|u^{exact} - u_N\|_2 = \sqrt{h \sum_{j=0}^{N} |u_j^{exact} - (u_N)_j|^2}.$$

The root-mean-aquare error (RMSE) defined as

$$\text{RMSE} = \sqrt{\frac{1}{n_1} \sum_{j=1}^{n_1} (\hat{u}_j - u_j)^2}.$$

Example. We consider (8) in the region $0 \le x, y \le 1,\ 0 < t \le 5$

$$u_t = 0.1(u_{xx} + u_{yy}) + 10u^2(1-u). \tag{24}$$

The analytical solution [14] of the equation is

$$u(x,y,t) = \frac{1}{1+\exp(5(x+y-t))}. \tag{25}$$

The initial and boundary conditions can be found from the analytical solution as

$$u(x,y,0) = \frac{1}{1+\exp(5(x+y))}, \quad 0 \le x, y \le 1. \tag{26}$$

and

$$u(0,y,t) = \frac{1}{1+\exp(5(y-t))}, \qquad u(1,y,t) = \frac{1}{1+\exp(5(1+y-t))},$$

$$u(x,0,t) = \frac{1}{1+\exp(5(x-t))}, \qquad u(x,1,t) = \frac{1}{1+\exp(5(x+1-t))}.$$
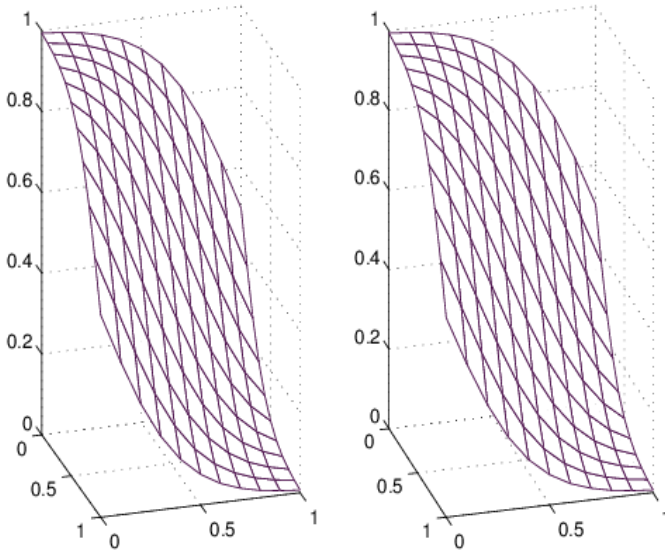
Fig. 1. Analytical solutions and numerical solutions with $T = 1$
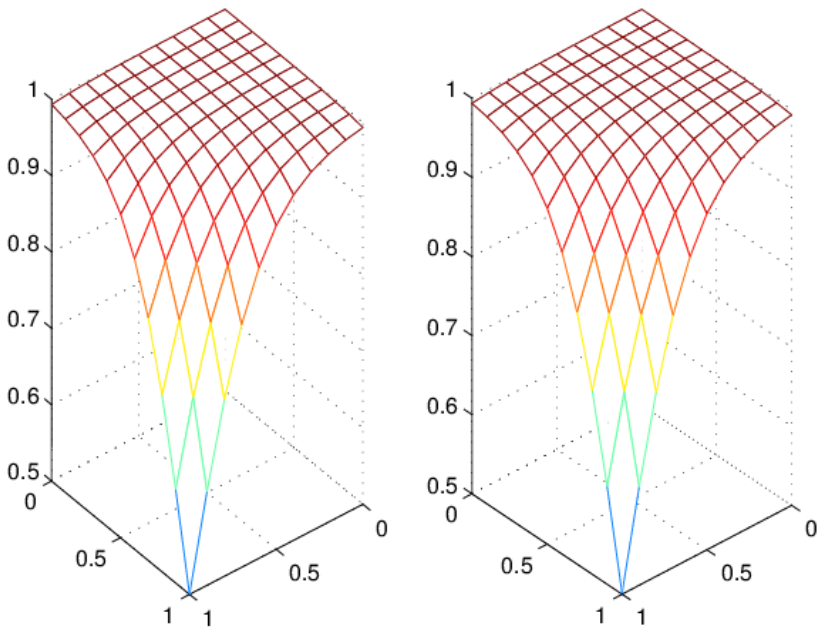
Fig. 2. Analytical solutions and numerical solutions with $T = 2$

For this example, we gain the figures of analytic solutions and numerical solutions using MQ with $c = 0.5$, $\delta t = 0.001$ and $\delta x = \delta y = 0.1$ in Fig.1 with $T = 1$ and in Fig.2 with $T = 2$.

Table 1. L^∞, L^2 errors and RMSE obtained using MQ, TPS for $T = 1, 1.5, 2, 2.5$

T		1	1.5	2	2.5
L^∞ errors	MQ	1.5×10^{-3}	1.5×10^{-3}	1.5×10^{-3}	6.4574×10^{-4}
	TPS	1.5×10^{-3}	1.5×10^{-3}	1.5×10^{-3}	6.4574×10^{-4}
L^2 errors	MQ	1.8×10^{-3}	1.7×10^{-3}	1.5×10^{-3}	3.6692×10^{-4}
	TPS	1.9×10^{-3}	1.8×10^{-3}	1.4×10^{-3}	3.8424×10^{-4}
RMSE	MQ	5.0893×10^{-4}	4.7812×10^{-4}	4.2886×10^{-4}	1.0420×10^{-4}
	TPS	5.2593×10^{-4}	5.0936×10^{-4}	4.0014×10^{-4}	1.0912×10^{-4}
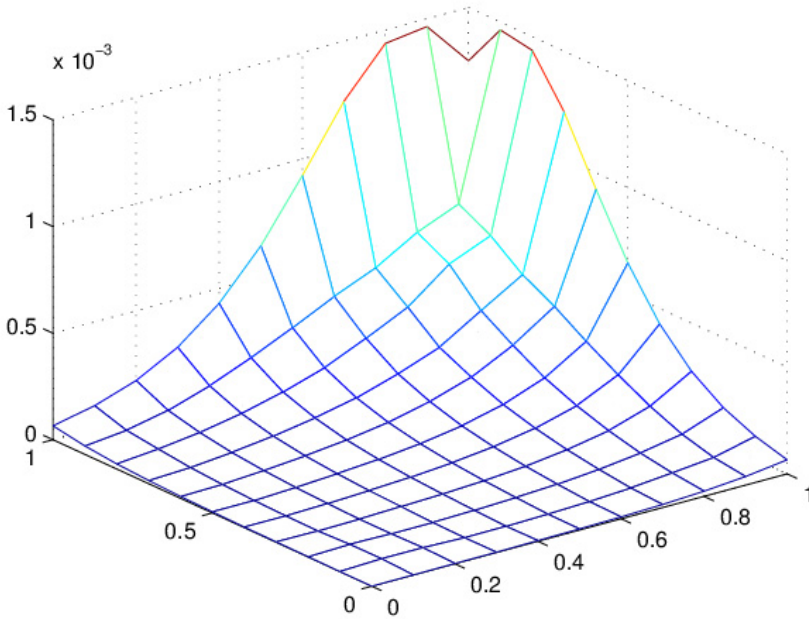
Fig. 3. The errors between analytical solutions and numerical solutions using TPS with $T = 2$

In Table 1, we can gain the L^∞, L^2 errors and RMSE obtained using MQ with $c = 0.5$ and TPS with $m = 2$ for $T = 1, 1.5, 2, 2.5$, $\delta t = 0.001$ and $\delta x = \delta y = 0.1$. in Fig.3, we gain the figures of the errors between analytic solutions and numerical solutions, using TPS with $m = 2$ for $\delta t = 0.001$, $\delta x = \delta y = 0.1$ and $T = 2$.

5   Conclusion

In this paper, we propose the Method of Particular Solutions for Solving a Class of Nonlinear Reaction-Diffusion Equations using collocation points and approximating the solution using multiquadrics (MQ) and the Thin Plate Splines (TPS) Radial Basis Function (RBF). The scheme works in a similar fashion as finite-difference methods. The results of numerical experiments are presented, and are compared with analytica solutions to confirm the good accuracy of the presented scheme. The numerical results given in the previous section demonstrate the good accuracy of this scheme.

References

1. Yang, Y., Wang, Y.M.: A finite difference scheme for solving nonlinear reaction diffusion equations. East China Norma University (Natural Science) 4, 1–8 (2002)
2. Kansa, E.J.: Multiquadrics-a scattered data approximation scheme with applications to computational fluid dynamics-I. Comput. Math. Appl. 19, 127–145 (1990)
3. Quan, S.: A meshless method of lines for the numerical solution of KdV equation using radial basis functions. Engineering Analysis with Boundary Elements 33, 1171–1180 (2009)
4. Li, M., Jiang, T., Hon, Y.C.: A meshless method based on RBFs method for nonhomogeneous backward heat conduction problem. Engineering Analysis with Boundary Elements 34(9), 785–792 (2010)
5. Buhmann, M.D.: Radial Basis Functions. Cambridge University Press, Cambridge (2003)
6. Dehghan, M., Shokri, A.: A numerical method for solution of the two-dimensional sine-Gordon equation using the radial basis functions. Mathematics and Computers in Simulation 79, 700–715 (2008)
7. Jiang, T., Jiang, Z., Kolibal, J.: A numerical method for one dimensional time-dependent Schrodinger equation using radial basis functions. Internatioal Journal of Computational Methods (accepted, 2013)
8. Dehghan, M., Shokri, A.: A numerical method for KdV equation using collocation and radial basis functions. Nonlinear Dyn. 50, 111–120 (2007)
9. Dehghan, M., Shokri, A.: A numerical method for two-dimensional Schodinger equation using collocation and radial basis functions. Computer and Mathematics with Applications 54, 136–146 (2007)
10. Chen, C.S., Fan, C.M., Wen, P.H.: The method of fundamental solutions for solving ellipticpdes with variable coefficients. In: Chen, C.S., Karageorghis, A., Smyrlis, Y.S. (eds.) The Method of Fundamental Solutions-A Meshless Method. Dynamic System Inc. (2008)
11. Wen, P.H., Chen, C.S.: The Method of Particular Solutions for Solving Scalar Wave Equations. Int. J. for Numer. Methods in Biomedical Eng. 26, 1878–1889 (2010)
12. Jiang, T., Li, M., Chen, C.S.: The Method of Particular Solutions for Solving Inverse Problems of a Nonhomogeneous Convection-Diffusion Equation with Variable Coefficients. Numerical Heat Transfer, Part A: Applications 61(5), 338–352 (2012)
13. Buhmann, M.D.: Multivariate cardinal interpolation with radial–basis functions. Constr. Approx. 6, 225–255 (1990)
14. Liu, X.: Study on Time Splitting Algorithms of Local Discontinuous Galerkin method for 2-Dimensional Diffusion-Reaction Problems. J. 11 (2011)

Hopf Bifurcation of Rayleigh Model with Delay

Yan Sun and Wenlian Ma

College of Science, Changchun University of Science and Technology
ChangChun, JiLin, China

Abstract. Hopf bifurcation and stability of Rayleigh model with delay r as parameter have been studied. Analyzing the stability of zero solution, the branch value of Hopf bifurcation has been obtained. Secondly, using theories of normal form and center manifold acquire the calculation formula of the direction of the bifurcation and the stability of the periodic solution. Finally, some numerical simulations are carried out by using MATLAB software, and the results are consistent with analysis results.

Keywords: delay, Rayleigh model, Hopf bifurcation, stability.

1   Introduction

Rayleigh model with delay is as follows:

$$\ddot{x}(t) - \eta(ax^2(t) + \beta x(t) + \gamma)\dot{x}(t) + x(t-r) = 0 \tag{1}$$

In the equation, parameter $\eta > 0$ and a, β, γ are arbitrary constant.

Since the model has significant practical background and extremely rich dynamical behavior. In recent years, it has been paid attention by domestic and overseas scholars. At the present time, researching questions of Rayleigh Model with delay has achieved abundant accomplishment [1,6]. Paper [1] utilizes Hopf bifurcation theorem to discuss the existence of periodic solution of the system (1). It gives the computational formula to determine the nature of bifurcation. The paper [2] presents sufficient condition of a wide range of periodic solutions in model (1).

The paper [5] takes into account the impact of inflationary effects to the price oscillation, and study the stability with multiple delay Rayleigh model and property of local Hopf bifurcation. It gives the results of numerical simulation. While several variable delays have been existed in the system, its dynamic properties become more and more complex [7-10].

The paper researches Rayleigh model as follows:

$$\ddot{x}(t) - \eta(ax^2(t) + \beta x(t) + \gamma)\dot{x}(t-r) + x(t-r) = 0 \tag{2}$$

In this paper, we will make delay r as bifurcation parameter and discuss the stability of zero solutions in the system (2). We can obtain Hopf bifurcation, application center manifold and Hassard "normal form" theorem [11], and determine the direction of Hopf bifurcation and formula of Hopf bifurcation stability [11-12]. An example which is a specific numerical simulation has been given and proves the correctness of results by theoretical analysis.

2   Stability of Zero Solutions and Existence of Hopf Bifurcation

In order to discuss conveniently, we let

$$f(x) = -\eta(ax^2(t) + \beta x(t) + \gamma), \quad g(x) = x(t) \quad \text{and} \quad f(0) = -\eta\gamma = -k,$$

$g(0) = 0, g'(0) = 1$. Obviously, it has an unique equilibrium point (0, 0) in the system(2). Proper equation of linear system at point (0, 0) in the equation (2) is

$$\lambda^2 - ke^{-\lambda r}\lambda + e^{-\lambda r} = 0 \tag{3}$$

Lemma 1

(I) Characteristic equation (3) has a pure imaginary pair eigenvalues $\pm iw_0$ at the point $\tau = \tau_j$, therein,

$$r_n = \frac{\arctan(-kw_0)}{w_0} + 2n\pi \ (n = 0,1,2....), \quad w_0 = \sqrt{\frac{k^2 + \sqrt{k^2 + 4}}{2}}.$$

(II) If Characteristic equation (3) has eigenvalue $\lambda(r) = \alpha(r) + iw(r)$,so

$$\frac{d\alpha(r)}{dr}\bigg|_{r=r_n} > 0.$$

Proof: (I) If $\lambda = iw(w > 0)$ is the root of equation (3), it is substituted into equation (3) and separated real parts and imaginary parts.

$$\begin{cases} -w^2 - wk\sin(wr) + \cos(wr) = 0 \\ wk\cos(wr) + \sin(wr) = 0 \end{cases}$$

So

$$\begin{cases} w^2 k^2 + 1 = w^4 \\ \tan(wr) = -kw \end{cases} \tag{4}$$

Obviously, equation (4) has an unique real positive roof $w_0 = \sqrt{\dfrac{k^2 + \sqrt{k^2 + 4}}{2}}$,

at the same time, $r_n = \dfrac{\arctan(-kw_0)}{w_0} + 2n\pi$,, (n=0,1,2.···).

We know the result (II) from lemma (I). $\lambda(r)$ satisfies the relationship $\alpha(r_n) = 0, w(r_n) = \pm w_0$ and we put $\lambda(r) = \alpha(r) + iw(r)$ into equation (3) and separate parts and imaginary parts.

$$\begin{cases} e^{ar}(\alpha^2 - w^2) + [k\alpha\cos(wr) + kw\sin(wr) + \cos(wr)] = 0 \\ 2\alpha we^{ar} + [kw\cos(wr) - k\alpha\sin(wr) - \sin(wr)] = 0 \end{cases}$$

Equation (3) is differentiated about r on both sides, so

$$\frac{d\lambda(r)}{dr} = \frac{\lambda - \lambda^2 k}{2\lambda e^{\lambda r} + (\lambda kr - k - r)}$$

In the meantime, we can get the equation (5) due to the equation (4) establish.

$$\begin{cases} [k\cos(w_0 r_n) - r_n w_0^2]\dot{\alpha} + [k\sin(w_0 r_n) - 2w_0]\dot{w} = 0 \\ [2w_0 - k\sin(w_0 r_n)]\dot{\alpha} + [k\cos(w_0 r_n) - r_n w_0^2]\dot{w} = w_0^3 \end{cases} \tag{5}$$

So

$$\frac{d\alpha}{dr}\bigg|_{r=r_n} = \frac{-w_0^3[k\sin(w_0 r_n) - 2w_0]}{[k\cos(w_0 r_n) - r_n w_0^2]^2 + [k\sin(w_0 r_n) - 2w_0]^2} > 0$$

Lemma 2

$\omega_n(r) = \omega_n$ is set up, and

(1) The characteristic root of characteristic equation is $\lambda = \pm i\omega_n$, the necessary and sufficient condition is $\gamma = \gamma_n(r) = -\dfrac{a}{\omega_n}\sin(\omega_n r)$;

(2) Each given condition $r \geq \dfrac{2n\pi}{\sqrt{ka}}$ has $\text{Re}\dfrac{\partial\lambda_n}{\partial r}\bigg|_{\gamma=\gamma_n(r)} > 0$;

(3) While $\gamma = \gamma_0(r)$, characteristic equation have pure imaginary roots $\pm i\omega$, and the other roots have negative real parts.

We gain the following conclusions by the lemma and the theorem of Hopf bifurcation [12].

Lemma 3. We can get two conclusions from equation (2).

(I) $r \in (0, r_0)$, the zero solution of it is progressive stability.

(II) $r = r_n (n = 0, 1, \cdots)$ is Hopf bifurcation in the system (2), that is to say system(2) bifurcate small amplitude periodic solution.

3   Analysis of Hopf Bifurcation

Assuming $\dot{x} = y$, we obtain equivalent first order equations in the equation (2).

$$\begin{cases} \dot{x}(t) = y(t) , \\ \dot{y}(t) = \eta\left[ax^2(t) + \beta x(t) + \gamma \right] y(t) - x(t-r) , \end{cases} \tag{6}$$

We should let the right side of equation (6) go to Taylor expansion at a sufficiently small field of the origin, then

$$\begin{cases} \dot{x} = y \\ \dot{y} = ky(t-r) - x(t-r) - \dfrac{g''(0)x^2(t-r)}{2} - \dfrac{g'''(0)x^3(t-r)}{6} - f'(0)xy(t-r) - \dfrac{f''(0)x^2 y(t-r)}{2} - \cdots \end{cases} \tag{7}$$

The nonlinear term of the right end in equation (7) is

$$h = \begin{pmatrix} 0 \\ -\dfrac{g''(0)x^2(t-r)}{2} - \dfrac{g'''(0)x^3(t-r)}{6} - f'(0)xy(t-r) - \dfrac{f''(0)x^2 y(t-r)}{2} - \cdots \end{pmatrix} \tag{8}$$

Recorded as

$$C^k([-1,0],R) = \left\{ \varphi \,\middle|\, \varphi : [-1,0] \to R^2 , \varphi \text{ is } k \text{ times continuously differential} \right\},$$

and $C = C([-r,0],R^2)$. We define the function $\|\varphi\| = \sup\limits_{-r \le \theta \le 0} |\varphi(\theta)|$ while $\phi \in C[-1,0]$, equation (8) can be expressed as

$$\dot{X} = AX_t + RX_t \tag{9}$$

Therein, $X_t = X(t+\theta) \in C^1[-r,0]$, and $X_t = (x_t, y_t)^T$)

Define that

$$A\phi = \begin{cases} \dfrac{d\phi}{d\theta} & -r \le \theta \le 0 \\ L\phi & \theta = 0 \end{cases} \tag{10}$$

$$R\phi(\theta) = \begin{cases} (0,0)^T & -r \le \theta \le 0 \\ h(\phi) & \theta = 0 \end{cases} \tag{11}$$

In which Riesz express theorem, second-order matrix $\eta : [-r,0] \to R^{2\times 2}$ exist component by function of bounded variation, then

$$L\phi = \int_{-r}^{0} [d\eta(s)]\varphi(s) \tag{12}$$

C^* is dual space of C ,there exists $\psi \in C^*$,and makes that A^* is adjoint method of $A = A(0)$.

So

$$A^*\psi(\xi) = \begin{cases} -\dfrac{d\psi}{d\xi} & 0 < \xi \leq r \\ \displaystyle\int_{-r}^{0} [d\eta^T(t)]\psi(-t) & \xi = 0 \end{cases} \tag{13}$$

We define bilinear inner product with $\phi \in C([-r,0], R^2)$, $\psi \in C^*([0,r], R^2)$ and we can get the equation

We define bilinear inner product that

$$<\psi,\phi> = \bar{\psi}^T(0)\phi(0) - \int_{\theta=-r}^{0} \int_{\xi=0}^{\theta} \bar{\psi}^T(\xi-\theta)[d\eta(\theta)]\phi(\xi)d\xi \tag{14}$$

When $r = r_0$ is established, $q(\theta)$ and $q^*(\xi)$ is respectively defined as eigenvector which are corresponded with eigenvalue of iw_0 and $-iw_0$ for operator A and operator A^* , so

$$\begin{cases} Aq(\theta) = iw_0 q(\theta) \\ A^*q^*(\xi) = -iw_0 q^*(\xi) \end{cases} \tag{15}$$

Lemma 4

$q(\theta) = (1,\delta)^T e^{iw_0\theta}, \quad -r \leq \theta \leq \theta$, $q^*(\xi) = N(\beta,1)^T e^{iw_0\xi}, \quad 0 \leq \xi \leq r$, $<q^*,q> = 1$,

In the equation, $\delta = iw_0, \quad \beta = \dfrac{e^{iw_0 r_0}}{iw_0}$

And

$$N = \frac{1}{\beta + \bar{\delta} - (1+k\bar{\delta})r_0 e^{iw_0 r_0}} = \frac{iw_0}{[\cos(w_0 r_0) + w_0^2] + i[\sin(w_0 r_0) - w_0^3 - r_0]}$$

We define the symbol $\Lambda = \{iw_0, -iw_0\}$, then C exists a direct sum decomposition $C = P_\Lambda \oplus Q_\Lambda$. Among the equation, P_Λ is eigenvector of Λ corresponding with operator A , Q_Λ is complementary space of P_Λ .

We set $\phi = (\phi_1, \phi_2)$ to a basis of P_Λ. When $r = r_0$, local center manifold is

$\Omega = \{X_t \in C : X_t = \phi X + u(X, h(\phi))\}$, and $X = (x, y)^T \in U(0), u \in Q_\Lambda$.

$$h(\phi) = \begin{bmatrix} 0 \\ -\dfrac{g''(0)\phi_1^2(-r)}{2} - \dfrac{g'''(0)\phi_1^3(-r)}{6} - f'(0)\phi_1(0)\phi_2(-r) - \dfrac{f''(0)\phi_1^2(0)\phi_2(-r)}{2} - \cdots \end{bmatrix}$$

$X(t)$ is satisfied with autonomous differential equation.

$$\dot{X} = \begin{bmatrix} 0 & -w_0 \\ w_0 & 0 \end{bmatrix} X + \psi(0) h(\phi X) \tag{16}$$

$\psi = (\psi_1, \psi_2)$ is a dual-base of P_Λ, we can learn by lemma 3 that there exist

$$\varphi(\theta) = \begin{bmatrix} \cos(w_0\theta) & \sin(w_0\theta) \\ -w_0 \sin(w_0\theta) & w_0 \cos(w_0\theta) \end{bmatrix} \tag{17}$$

$$\psi(0) = \begin{bmatrix} \operatorname{Re} N \operatorname{Re} \beta - \operatorname{Im} N \operatorname{Im} \beta & \operatorname{Re} N \operatorname{Im} \beta + \operatorname{Im} N \operatorname{Re} \beta \\ \operatorname{Re} N & \operatorname{Im} N \end{bmatrix} \tag{18}$$

So

$$\phi X = \begin{bmatrix} x\cos(w_0\theta) + y\sin(w_0\theta) \\ -w_0 x \sin(w_0\theta) + w_0 y \cos(w_0\theta) \end{bmatrix} \tag{19}$$

$$h(\phi X) = \begin{bmatrix} 0 \\ h_2(x, y) \end{bmatrix} \tag{20}$$

$$h_2(x, y) = I_1 x^2 + I_2 xy + I_3 y^2 + I_4 x^3 + I_5 x^2 y + I_6 xy^2 + I_7 y^3 + O(\rho^4) \tag{21}$$

Among the equation, we know that

$I_1 = -\left[\dfrac{g''(0)}{2}\cos^2(w_0 r_0) + w_0 f'(0)\sin(w_0 r_0) \right]$, $\quad I_2 = \dfrac{g''(0)}{2}\sin(2w_0 r_0) - w_0 f'(0)\sin(w_0 r_0)$,

$I_3 = \dfrac{g''(0)}{2}\sin^2(w_0 r_0), I_4 = -\left[\dfrac{g'''(0)}{6}\cos^3(w_0 r_0) + \dfrac{f''(0)}{2} w_0 \sin(w_0 r_0) \right]$,

$I_5 = \dfrac{g'''(0)}{2}\sin(w_0 r_0)\cos^2(w_0 r_0) - \dfrac{f''(0)}{2} w_0 \sin(w_0 r_0), I_6 = -\dfrac{g'''(0)}{2}\sin^2(w_0 r_0)\cos(w_0 r_0)$,

$I_7 = \dfrac{g'''(0)}{6}\sin^3(w_0 r_0)$.

We put equation (18) and (20) in the equation (16) and obtain the conclusion

$$\begin{cases} \dot{x} = -wy + (\operatorname{Re} N \operatorname{Im} \beta + \operatorname{Im} N \operatorname{Re} \beta) h_2(x, y) \\ \dot{y} = wx + \operatorname{Im} N h_2(x, y) \end{cases} \tag{22}$$

We can compute the results by lemma 3.

$$\operatorname{Re} N \operatorname{Im} \beta + \operatorname{Im} N \operatorname{Re} \beta = \operatorname{Im}(N\beta) = -\frac{w_0^2 \sin(w_0 r_0) + w_0^3 r_0 \cos(w_0 r_0)}{[\cos(w_0 r_0) + w_0^2]^2 + [\sin(w_0 r_0) - w_0^3 r_0]^2}$$

$$\operatorname{Im} N = \frac{w_0[\cos(w_0 r_0) + w_0^2]}{[\cos(w_0 r_0) + w_0^2]^2 + [\sin(w_0 r_0) - w_0^3 r_0]^2}$$

Equation (21) can be changed to the following equation (23) by approximate identical transformation.

$$\begin{cases} \dot{x} = \tilde{a}(x^2 + y^2)x - [w_0 + \tilde{b}(x^2 + y^2)]y + O(\rho^4) \\ \dot{y} = [w_0 + \tilde{b}(x^2 + y^2)]x + \tilde{a}(x^2 + y^2)y + O(\rho^4) \end{cases} \tag{23}$$

We transform equation (23) into polar coordinate.

$$\begin{cases} \dot{\rho} = \tilde{a}\rho^3 + O(\rho^4) \\ \dot{\theta} = w_0 + \tilde{b}\rho^2 + O(\rho^3) \end{cases} \tag{24}$$

The universal unfold of equation (24) is

$$\begin{cases} \dot{\rho} = \mu\rho + \tilde{a}\rho^3 + O(\rho^4) \\ \dot{\theta} = w_0 + \tilde{b}\rho^2 + O(\rho^3) \end{cases} \tag{25}$$

In the equation, μ is a unfolding parameter. $\tilde{a}$ is a coefficient and it can show that

$$\tilde{a} = \frac{1}{16\{[\cos(w_0 r_0) + w_0^2]^2 + [\sin(w_0 r_0) - w_0^3 r_0]^2\} w_0^6}(mg'''(0) + nf''(0))$$

$$+ \frac{1}{16\{[\cos(w_0 r_0) + w_0^2]^2 + [\sin(w_0 r_0) - w_0^3 r_0]^2\} w_0^{10}}(sg''^2(0) + wf'(0)g''(0) + vf'^2(0))$$

In the equation,

$$m = k - kw_0 - r_0 w_0 + k^3 w_0^2 - k^3 w_0^3 - k^2 r_0 w_0^3 + kw_0^4 + k^3 w_0^6$$
$$n = -kw_0^5 - 3k^2 w_0^7 - 3kr_0 w_0^7 - kw_0^9$$
,

$$s = k - kw_0 - r_0 w_0 + k^3 w_0^2 + 2kw_0^4 - k^3 w_0^4 - 2k^2 r_0 w_0^4 - kr_0^2 w_0^4 - kw_0^5 + k^5 w_0^5 - r_0 w_0^5$$

$$+ k^4 r_0 w_0^5 + 2k^3 w_0^6 - k^3 w_0^6 - 2k^4 r_0 w_0^6 - k^3 r_0^2 w_0^6 + kw_0^8 + k^5 w_0^9 + k^4 r_0 w_0^9 + k^3 w_0^{10}$$

$$w = -kw_0^3 + 2k^2 w_0^4 - 4k^2 w_0^5 - k^3 w_0^5 - 4kr_0 w_0^5 - 2kw_0^7 + k^3 w_0^7 + 2k^2 r_0 w_0^7 +$$

$$kr_0^2 w_0^7 + 4k^2 w_0^8 - 2k^4 w_0^8 - 4k^3 r_0 w_0^8 - 2k^2 r_0^2 w_0^8 - 4k^2 w_0^9 - 2k^3 w_0^9 + k^5 w_0^9 -$$

$$4kr_0 w_0^9 + 2k^4 r_0 w_0^9 + k^3 r_0^2 w_0^9 - kw_0^{11} + 2k^2 w_0^{12} - k^3 w_0^{13}$$

$$v = -2k^2 w_0^7 - 4k^3 w_0^9 - 4k^2 r_0 w_0^9 - 4k^2 w_0^{11} + 2k^4 w_0^{11} + 4k^3 r_0 w_0^{11} +$$

$$2k^2 r_0^2 w_0^{11} - 4k^3 w_0^{13} - 4k^2 r_0 w_0^{13} - 2k^2 w_0^{15}$$

Theorem 1

Discrete delays of center manifold Ω have the following properties near the origin.

(I) $\tilde{a}$ determine the direction of bifurcation. When inequality $\tilde{a} > 0$ ($\tilde{a} < 0$) is established, bifurcation periodic solution is lower critical (upper critical).

(II) when $\tilde{a} > 0$, that is to say $r \in [0, r_0)$, system(2) branches a stable periodic solution near the equilibrium point.

(III) when $\tilde{a} < 0$, that is to say $r > r_0$, system (2) bifurcates a unstable periodic solution near the equilibrium point.

(IV)Hopf bifurcation has been occurred in system (2) when $r = r_0$.

4   Numerical Simulation

In order to preferably support the theoretical results, we carry numerical simulation on system (2). The parameter selection is $\eta = 0.1, \ a = 0, \ \beta = 0, \ \gamma = -1$.

It means that system (2) has changed to the following equation.

$$\ddot{x} + 0.1\dot{x}(t - r) + x(t - r) = 0 \tag{26}$$

Hopf bifurcation has been easily calculated to $r_0 = 2.4514$. When we put different numerical number on r, we should observe the change of equation (26).

To get delay $r = 2.3843 < r_0$, origin (0,0) is the stable focal point in system (26), in other words, positive equilibrium point is asymptotically stable in the figure 1; to get delay $r = 2.5379 > r_0$, origin (0,0) has been changed the original stability and it becomes instable point in system (20) in the figure (2). When stability has altered, the upper critical Hopf bifurcation occurs, thus unstable limit cycle has been come into being.
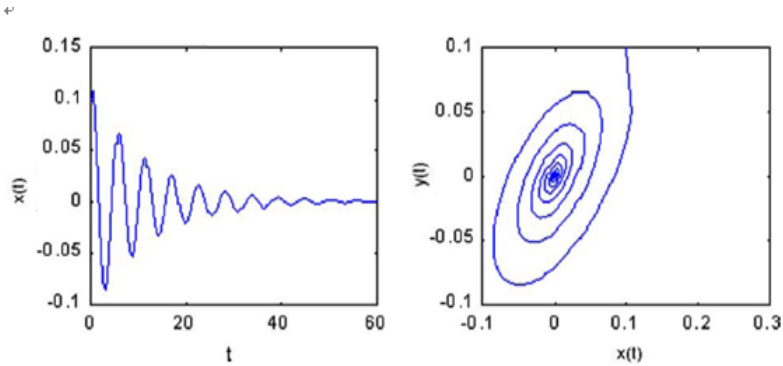
Fig. 1. With $r < r_0$, phase diagram and wave diagram of system (26)
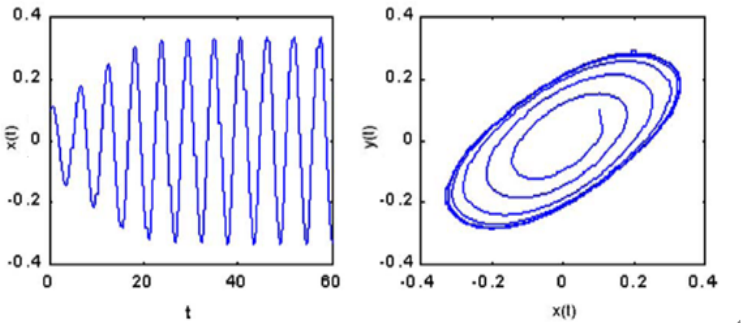
Fig. 2. With $r > r_0$, phase diagram and wave diagram of system (26)

References

1. Lv, T., Liu, Z.: Hopf Bifurcation of Price Reyleigh Equation with Delay. Journal of Jilin University (Science Edition) 47(3), 441–448 (2009)
2. Lv, T., Zhou, L.: Global Periodic Solution of Price Reyleigh Equation with Time Delay. Journal of Changchun University of Science and Technology (Natural Science Edition) 32(3), 515–517 (2009)
3. Lv, T., Zhou, L.: Hopf and resonant codimension two bifurcation in price rayleigh equation with delays. Journal of Northeast Normal University (Natural Science Edition) 44(4), 409–416 (2012)
4. Cao, Z.: The Periodically Perturbed Hopf Bifurcation for Price Reyleigh Equation with Delay. Northeast Normal University, Changchun (2006)
5. Lv, T., Zhou, L.: Hopf and Codimension Two Bifurcation in Price Rayleigh Equation with Two Time Delays. Journal of Jilin University (Science Edition) 50(3), 409–416 (2012)

6. Lv, T.: Neimark-Sacker Hopf Bifurcation for Price Reyleigh Equation with Two Delays. Journal of North University of China (Natural Science Edition) 33(5), 490–494 (2012)
7. Zhang, X.-F., Chen, X., Chen, Y.-Q.: A Qualitative Analysis of Price Model in Differential Equations of Price. Journal of Shenyang Institute of Aeronautical Engineering 21(1), 83–86 (2004)
8. Yu, C.-B., Wei, J.-J., Zhou, X.-F.: Bifurcation Analysis in an Age-Structured Model of a Single Species Living in Two Identical Patches. Applied Mathematical Modelling 34(4), 1068–1077 (2009)
9. Wang, A.Y., Yu, C.-B.: Bifurcation Analysis in an Age-Structured Model of Single Species Living in Two Identical Patches. Differential Equations and Dynamical Systems 16(1/2), 101–120 (2008)
10. Yang, R.-Z.: Stability and Hopf Bifurcation Analysis in Species Model with Two Delays [Master's Degree Thesis]. Harbin Institute of Technology, Harbin (2010)
11. Hassard, B.D., Kazarinoff, N.D., Wan, Y.H.: Theory and Applications of Hopf Bifurcation. Cambridge University Press, Cambridge (1981)
12. Hale, J.K.: Introduction to Functional differential Equations, pp. 227–268. Springer, Berlin (1993)

Route Standardization Based on Polygon Triangulation Cutting Algorithm

Baozhu Zhao, Huifeng Zhang, Yanfeng Jin, and Yanling Su

Shi Jiazhuang Post & Telecommunication Technical College, 050021,
Shijiazhuang, China
zhaobz@cptc.cn, {zhanghf77,yanfengjin}@163.com, suyanl@sohu.com

Abstract. Polygon triangulation cutting algorithm has been widely used because of its excellent characteristics. The method is introduced into route standardization in this paper. Firstly, triangulate the polygon composed by Geometric points with Displacement and time by using Polygon triangulation cutting algorithm. Use the formula to calculate division of the triangle area, and then compute the area of the whole polygon. We can get n different polygons by scanning different period, repeat the algorithm and we can get the area of n polygons. Through the weighted average method to get the best a group of polygon set point, which composed the best path. The results of the experiments prove the validity of the method.

Keywords: Route Standardisation, Polygon cutting algorithm, triangulation cutting.

1   Introduction

Polygon triangulation is the classic problem of computational geometry, originated in an interesting art gallery problem. Currently, there are many different algorithms for polygon triangulation [1-3]. Symmetrical shape and quick calculation are the main two purposes for the triangulation algorithm and also determine two different application concentrations for polygon triangulation [4]. For Symmetrical shape concentration, people study on the nature of triangles and the best criteria and algorithm for shapes very well, especially for Delaunay criterion. These algorithms ensure symmetrical shape as the premise, but also as far as possible to improve the calculation speed [5-8]. For the finite element analysis and many other applications, triangular symmetry is necessary and quality is certainly required for each unit. However, for some applications, there is no strict requirement for triangular symmetry, even no requirement for quality of the unit. For example, when OpenGL displays a graphic, different triangulation strategy has no effect on the graphics. Considering the triangle of symmetry will restrict algorithm thinking, thus affecting the efficiency of the algorithm [9-11]. Therefore pursuing high-efficiency calculations of triangulation algorithm is meaningful.

Polygon triangulation cutting algorithm has been widely used because of its excellent characteristics. The method is introduced into route standardization in this paper. Firstly, triangulate the polygon composed by Geometric points with Displacement and time by using Polygon triangulation cutting algorithm. Use the formula to calculate division of the triangle area, then compute the area of the whole polygon [12]. We can get n different polygons by scanning different period, repeat the algorithm and we can get the area ot n polygons. Through the weighted average method to get the best a group of polygon set point, which composed the best path. The results of the experiments prove the validity of the method.

2   Polygon Triangulation Cutting Algorithm

2.1   Basic Concept

Simple Polygon: graphics constituted by single closed polygonal chain which is not at the intersection. (Contains no holes, side not intersect)

Triangulation: Through a set of greatly diagonal which is not at the intersection with each other, cut a polygon into multiple set of triangles.

Theorem 1: there is (at least) one triangulation in any simple polygon. If the number of vertices is n, its triangulations results contain n-2 triangles.

Theorem 2: (art gallery theorem) any simple polygon containning n vertex, (in the worst case), only needs n / 3 cameras to guarantee the points in polygon is visible in at least a camera at most.

Theorem 3: any simple polygon with n vertex, can always confirm position of n / 3 cameras in polygons in the O (n log n) time, make any point in the polygon in visible of the camera.

Decompose the Polygon into Trapezoids. Let S be a set of non-horizontal, non-intersecting line segments of the polygon. The randomized algorithm is used to create the trapezoidal decomposition of the X-Y plane arising due the segments of set S. This is done by taking a random ordering s1.. sN of the segments in S and adding one segment at a time to incrementally construct the trapezoids. This divides the polygon into trapezoids (which can degenerate into a triangle if any of the horizontal segments of the trapezoid is of zero length). The restriction that the segments be non-horizontal is necessary to limit the number of neighbors of any trapezoid. However, no generality is lost due to this assumption as it can be simulated using lexicographic ordering. That is, if two points have the same y-coordinate then the one with larger x-coordinate is considered higher. The number of trapezoids is linear in the number of segments. Seidel proves that if each permutation of s1.. sN is equally likely then trapezoid formation takes O(n log*n) expected time.

Decompose the Trapezoids into Monotone Polygons. A monotone polygon is a polygon whose boundary consists of two y-monotone chains. These polygons are computed from the trapezoidal decomposition by checking whether the two vertices of the original polygon lie on the same side. This is a linear time operation.

Triangulate the Monotone Polygons. A monotone polygon can be triangulated in linear time by using a simple greedy algorithm which repeatedly cuts off the convex corners of the polygon. Hence, all the monotone polygons can be triangulated in O(n) time.

All the data-structures used in the implementation are statically allocated. The trapezoid formation requires a structure where the neighbors of each trapezoid and its neighboring segments can be determined in constant time. Therefore, for every trapezoid, the indices of its neighbors and the segments are stored in its table-entry T.

The query-structure Q, used to determine the location of a point, is implemented as described by Seidel. The same Q can be later used for fast point-location queries. Both Q and T are updated as a new segment is added into the existing trapezoid formation. This entails splitting in two the trapezoid(s) in which the endpoints of the segment lie, then traversing along the edge of the segment to merge in any neighboring trapezoids which both share the same left and right edges and also share a horizontal edge. All the monotone polygons are stored in a single linked list with pointers to the first vertex in the list stored in a table.

Over a number of algorithms have been proposed to triangulate a polygon

Special cases: A convex polygon is trivial to triangulate in linear time, by adding diagonals from one vertex to all other vertices. The total number of ways to triangulate a convex n-gon by non-intersecting diagonals is the $(n-2)$-th Catalan number, which equals $\dfrac{n\cdot(n+1)\cdots(2n-4)}{(n-2)!}$, a solution found by Leonhard Euler.

A monotone polygon can be triangulated in linear time with either the algorithm of A. Fournier and D.Y. Montuno or the algorithm of Godfried Toussaint.

Ear clipping method: A polygon ear: One way to triangulate a simple polygon is based on the fact that any simple polygon with at least 4 vertices without holes has at least two 'ears', which are triangles with two sides being the edges of the polygon and the third one completely inside it (and with an extra property unimportant for triangulation). The algorithm then consists of finding such an ear, removing it from the polygon (which results in a new polygon that still meets the conditions) and repeating until there is only one triangle left.

This algorithm is easy to implement, but slower than some other algorithms, and it only works on polygons without holes. An implementation that keeps separate lists of convex and concave vertices will run in O(n2) time. This method is known as ear clipping and sometimes ear trimming. An efficient algorithm for cutting off ears was discovered by Hossam ElGindy, Hazel Everett, and Godfried Toussaint.

Using monotone polygons: Breaking a polygon into monotone polygons. A simple polygon may be decomposed into monotone polygons as follows.

For each point, check if the neighboring points are both on the same side of the sweep line, a horizontal or vertical line on which the point being iterated lies. If they are, check the next sweep line on the other side. Break the polygon on the line between the original point and one of the points on this one.

Note that if you are moving downwards, the points where both of the vertices are below the sweep line are 'split points'. They mark a split in the polygon. From there you have to consider both sides separately.

Using this algorithm to triangulate a simple polygon takes O(n log n) time.

Dual graph of a triangulation: A useful graph that is often associated with a triangulation of a polygon P is the so called dual graph. Given a triangulation T_P of P one defines the graph $G(T_P)$ as the graph whose vertex set are the triangles of T_P, two vertices (triangles) being adjacent if and only if they share a diagonal. It is easy to observe that $G(T_P)$ is a tree with maximum degree 3.

2.2   Ear Cutting

Given a polygon P, the main idea of the algorithm is to find ears - a triangle formed by three adjacent vertices and that is contained inside P - and remove them, in such a way that the remaining part is still a simple polygon. We choose some vertex p of P (colored red) and look to the triangle formed by its two adjacent vertices, r and s (colored blue). We need to check if this triangle forms an ear, that is, if this triangle is contained inside P.

First, p must be convex. This can be checked by an orientation test. If the polygon vertices are given in clockwise order, q comes before p and r comes after, then r must be clockwise oriented in relation of the oriented segment defined by p - q so that p is convex. Also, no other point of P can be inside the triangle.

It's possible to show that for any simple polygon, an ear always exists. Each time one is found, we remove a single vertex of P, decreasing its size until it reduces to 3. Since each ear can be found and removed in O(n) and we remove n-3 ears, this algorithm has an $O(n^2)$ complexity.
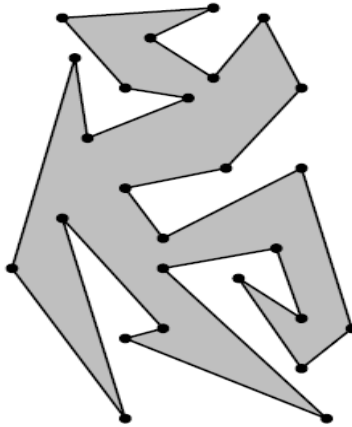
Fig. 1. Two-ears for triangulation

2.3   Polygon Triangulation Algorithms Based on Delaunay Triangulation

Triangulation of the simple polygons can also be achieved by the well-known Delaunay triangulation FLOR92 on a set of points. Namely, the vertices of a polygon can be

considered as individual input points in the plane. When computing the Delaunay triangulation we have to consider that some line segments edges of polygon must exist at the output. That problem is known as a constrained Delaunay triangulation CDT. Triangulation of V in which the circumcircle of each triangle t of T does not contain in its interior any other vertex P of T which is visible from the three vertices of t see Fig. 2c. Another characterization of CDT is given by the empty circle property: a triangle t in a constrained triangulation T is a Delaunay triangle if there does not exist any of her vertex of T inside the circumcircle of t and visible from all three vertices of t see Fig. 2a. See details in FLOR92.

The Delaunay triangulation of simple polygon can be generally computed as follows: the first step computes CDT of edges of simple polygon and the second step removes triangles that are in exterior of simple polygon. The information that input is a simple polygon not just general constraint graph could be useful in step one and therefore algorithms for building a CDT can be subdivided into two groups:

algorithms for computing the CDT when the constraint graph is a simple polygon.
algorithms for computing a CDT for general constraint graph.
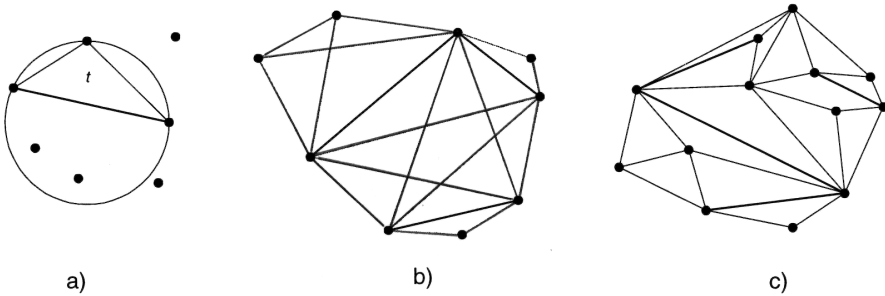
Fig. 2. a) Empty circle property; b) Visibility map; c) Constrained Delaunay triangulation

2.3.1   Constrained Delaunay Triangulation Algorithms for Simple Polygons

Lewis and Robinson described an O (n2)   algorithm based on divide-and-conquer approach with internal points of simple polygon LEWI79. The boundary polygon is recursively subdivided into almost equally sized subpolygons that are separately triangulated together with their internal points. The resulting triangulation is then optimized to produce CDT.

A recursive O (n2) algorithm for CDT based on visibility approach is described by Floriani FLOR85_. The algorithm computes the visibility graph of the vertices of the simple polygon Q in O (n2)   time and the Voronoi diagram P of set of its vertices in O (n log n). The resulting Delaunay triangulation is built by joining each vertex Q of P to those vertices that are both visible from Q and Voronoi neighbors of Q. Another O (n log n) algorithm was described by Lee and Lin in FLOR92. The algorithm is based on Chazelle's polygon cutting theorem. Chazelle has shown that for any simple polygon P with n vertices, two vertices t_1 and t_2 of P can be found in a linear time such that segment t_1t_2 is completely internal to P. Each of the two simple subpolygons

resulting from the cut of P by t_1t_2 has at least $n/3$ vertices. Lee and Lin's algorithm subdivides the given polygon Q into two subpolygons Q_l and Q_r and recursively computes the constrained Delaunay triangulations T_l and T_r. The resulting triangulation T of Q is obtained by merging T_l and T_r. They also proposed a similar algorithm for general constraint graph which runs in $O(n2)$.

2.3.2   Constrained Delaunay Triangulation Algorithms for General Constraint Graphs

Chew describes an $O(n \log n)$ algorithm for the CDT based on the divide-and-conquer approach. The constraint graph $G = (V,L)$ is assumed to be contained in a rectangle, which is subdivided into vertical strips CHEW87. In each strip there is exactly one vertex. The CDT is computed for each strip and adjacent strips are recursively merged together. After last merge we got the final CDT. The major problem here is merging those strips that contain edges, which cross some strip having no endpoint in it.

Algorithm for computing CDT, which includes preprocessing on the constraint segments, is proposed by Bossiant. By preprocessing CDT, the problem is transformed into standard Delaunay problem on a set of points. The idea is to modify the input data by adding points lying on the constraint segments in such a way that resulting Delaunay triangulation is guaranteed to contain such segments. Constraint segment e is a Delaunay edge if the circle having e as diameter does not intersect any other constraint segment. If the circle attached to e intersects some other segment, then e is split into a finite number of subsegments such that none of the circles attached to those segments intersect any constraint. When two constraint segments intersect at an endpoint, one newpoint is inserted into both segments. The circumcircle of the triangle defined by the common endpoint and by the two newpoints does not intersect any other constraint segment. This algorithm takes at most $O(n \log n)$ time and generates at most $O(n)$ additional points. For CDT all the above algorithms require that all points are defined at the beginning of the triangulation process. An algorithm proposed by Floriani and Puppo FLOR92 resolves CDT problem by incrementally updating CDT as new points and constraints are added. The problem of incrementally building of CDT is reduced to the following three subproblems:

- computation of an initial triangulation of the domain,
- insertion of a point,
- insertion of a straight-line segment.

An initial triangulation of the domain can be obtained by different approaches. For example, we can determine a triangle or rectangle made of two triangles, which contain the whole domain. Then, points and straight-lines are incrementally inserted. After each insertion we get new CDT which has more elements than the previous one. After inserting the last point or straight-line, the bounding triangle is removed. Algorithm runs at most in $O(\lg n2)$ where n is number of points and l the number of straightline segments in the final CDT.

2.4   The Triangulation Cutting Algorithm of Convex Polygon P CPTA: (Vertex Series Are $p_1, p_2, \cdots, p_n$)

(1)Calculate the diameter of P , set the two endpoint of the diameter as p_i, p_j ;

(2)Compare their length of $\overline{p_{i-2}p_i}$, $\overline{p_{i-1}p_{i+1}}$, $\overline{p_i p_{i+2}}$,   take a shortest one as the diagonal, delete the corresponding vertex, and output corresponding triangle. Do the same treatment to p_j ;

(3) $P' \leftarrow P - (p_{i-1} \cup p_i \cup p_{j+1}) \cap (p_{j-1} \cup p_j \cup p_{j+1})$;

(4)Repeat step 1 and step 4 to P' until P is cut completely.

2.5   Arbitrary Polygon Triangulations Algorithm APTA

(1)Segmentation polygon P into convex polygons sequence $P_1, P_2, \cdots, P_k$;

(2)Implement algorithm of 1to $p_i (i = \overline{1,k})$.

2.6   Simple Polygon Triangulations Algorithm

(1)If the vertex is cliques and convex point, connect the two adjacent points, delete it, and output of the triangle;

(2)Order the points into increasing sequence according to y dot, segmentate the polygon as left and right chains;

(3)Preset concave (convex) chains are two top vertex, and find the third highest summit p_c ;
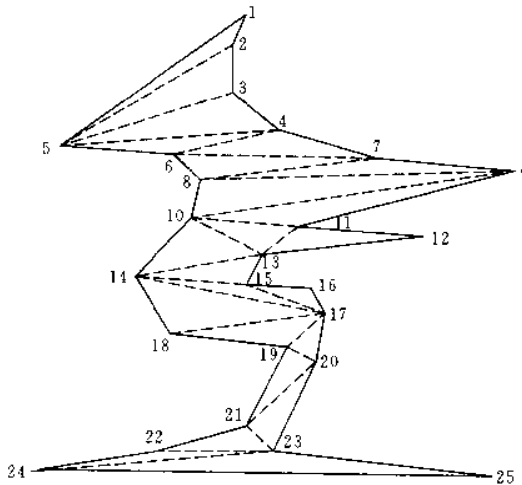
Fig. 3. Simple polygon triangulations

(4)If p_c is in the relative chain of concave (convex) chain, draw diagonal line from p_c to concave (convex) chain vertex p_{c-1}, and cut out the top vertex of concave (convex) chain, if the concave (convex) chain is empty, move down the summits;

(5)If p_c concave (convex) adjacent to the bottom of the chain, and p_{c-1} is a convex point, painting diagonal from p_c to the bottom of concave (convex) chain p_{c-2}, If the concave (convex) chain is empty, the top down, but if p_{c-1} is a concave point, add p_c to the bottom of the concave (convex) chain, then top down.

3   Calculation of the Polygonal Area and Best Path Selection

3.1   Calculation of the Polygonal Area

1.Given three points A(x1,y1)B(x2,y2)and C(x3,y3), the formula for the area is:

$$S(A,B,C) = \begin{vmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{vmatrix} * 0.5$$

(Counterclockwise is positive, clockwise is negative)

the polygons are A1A2A3``````An(both clockwise and counterclockwise are ok), any point P in a plane:

S (A1, A2, A3```````An) = abs(S(P, A1, A2) + S(P, A2, A3)+、、、+S(P, An, A1))

P is any point of the plane, so we can choose(0,0).

3.2   Route Selection

Route selection is actually searching the polygon according to both the time and displacement variables. In a cycle, coordinate is almost completely different in different period, so the polygon also has very big differences. That is the polygon path produced in the process of cycle has irregularities.

The best route is selected by calculating the area of the polygons. Firstly, calculate the area of the polygons, then select polygon which the area is closest with the weighted average area.

(1)According to the operation of the cycle, calculate the probability of each kind of polygon appears in the processx_1, x_2, ……, x_n, treat it as the weight of each polygon area.

(2)Calculate the weighted average of n polygon area $\bar{S} = \dfrac{1}{n}\sum_{1}^{n} S_i(A1, A2,..., An)$,
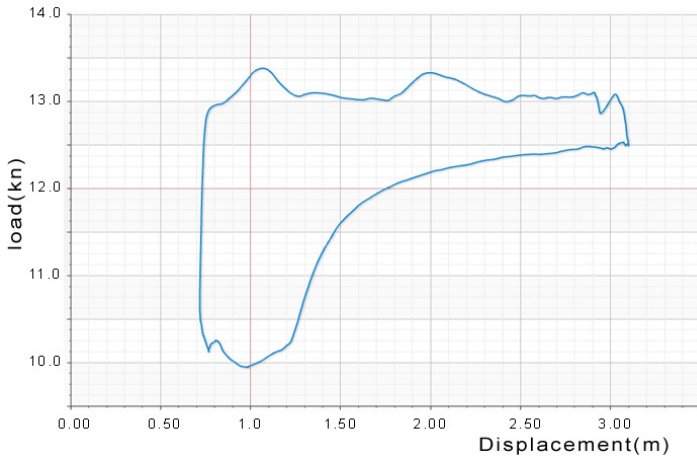
$i = 1,2,...,n.$.

Fig. 4. Single path

(3)calculate the D-value between area of each polygon S_i and the weighted average of n polygon area $\bar{S}$: $\delta_i = | S_i - \bar{S} |$, $i = 1,2,...,n$.

(4)Caculate min{ δ_i }, this is the best path selection.
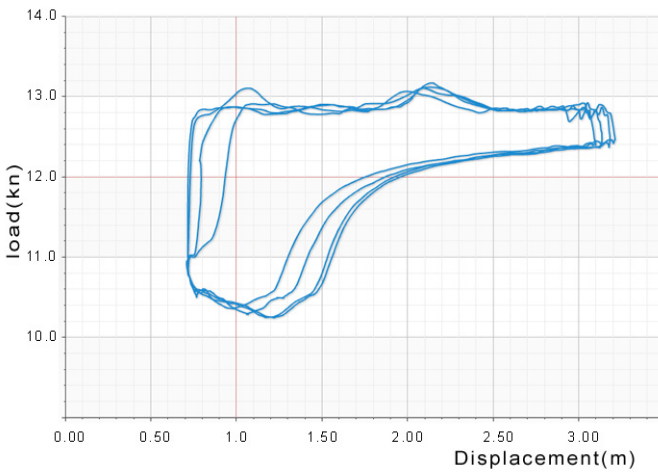
Fig. 5. Four-path
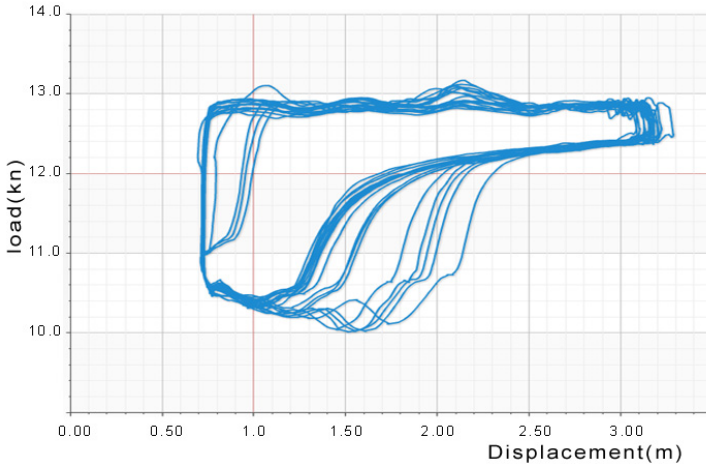
Fig. 6. Multi-path

4   Conclusion

Polygon triangulation cutting algorithm has been widely used because of its excellent characteristics. The method is introduced into route standardization in this paper. Firstly, triangulate the polygon composed by Geometric points with Displacement and time by using Polygon triangulation cutting algorithm. Use the formula to calculate division of the triangle area, and then compute the area of the whole polygon. We can get n different polygons by scanning different period, repeat the algorithm and we can get the area ot n polygons. Through the weighted average method to get the best a group of polygon set point, which composed the best path. The results of the experiments prove the validity of the method.

References

1.   O. Aichholzer, G. Rote, B. Speckmann, and I. Streinu: The Zigzag Path of a Pseudo-Triangulation. Proc. 8th Workshop on Algorithms and Data Structures (WADS 2003), LNCS 2748, 377–388. Springer-Verlag, Berlin, Germany 2003
2.   R. Downey and M. Fellows: Parameterized Complexity. Springer-Verlag, New York, NY, USA 1999
3.   M.R. Garey and D.S. Johnson: Computers and Intractability: A Guide to Theory of NP-Completeness. Freeman, New York, NY, USA 1979
4.   P.D. Gilbert: New Results in Planar Triangulations. Report R-850. University of Illinois, Coordinated Science Lab, 1979
5.   M. Grantson, C. Borgelt and C. Levcopoulos: A Fixed Parameter Algorithm for Minimum Weight Triangulation: Analysis and Experiments. Technical Report LUCS-TR: 2005-234, ISSN 1650-1276 Report 154. Lund University, Sweden 2005

6. M. Hoffmann and Y. Okamoto: The Minimum Triangulation Problem with Few Inner Points. Proc. 1st Int. Workshop on Parameterized and Exact Computation (IWPEC 2004), LNCS 3162, 200–212. Springer-Verlag, Berlin, Germany 2004

7. G.T. Klincsek: Minimal Triangulations of Polygonal Domains. Annals of Discrete Mathematics 9:121–123. ACM Press, New York, NY, USA 1980

8. E. Lodi, F. Luccio, C. Mugnai, and L. Pagli: On Two-Dimensional Data Organization, Part I. Fundaments Informaticae 2:211–226. Polish Mathematical Society, Warsaw, Poland 1979

9. A. Lubiw: The Boolean Basis Problem and How to Cover Some Polygons by Rectangles. SIAM Journal on Discrete Mathematics 3:98–115. Society of Industrial and Applied Mathematics, Philadelphia, PA, USA 1990

10. D. Moitra: Finding a Minimum Cover for Binary Images: An Optimal Parallel Algorithm. Algorithmica 6:624–657. Springer-Verlag, Heidelberg, Germany 1991

11. D. Plaisted and J. Hong: A Heuristic Triangulation Algorithm. Journal of Algorithms 8:405–437 Academic Press, San Diego, CA, USA 1987

12. M. Sharir and E. Welzl: On the Number of Crossing-Free Matchings (Cycles and Partitions), Proc. 17th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2006), to appear

Determinants of the RSFMLR and RSLMFL Circulant Matrices Involving Four Famous Numbers in Codes

Ting-Ting Xu[1,2] and Zhao-Lin Jiang[1,*]

[1] Department of Mathematics, Linyi University, Linyi 276005, P.R. China
[2] School of Mathematical Sciences, Shandong Normal University, Jinan 250014, P.R. China
xutingting655@163.com, jzh1208@sina.com

Abstract. Let A be a row skew first-minus-last right circulant matrix and C be a row are $(F_0, F_1, \ldots, F_{n-1})$, $(L_0, L_1, \ldots, L_{n-1})$, $(P_0, P_1, \ldots, P_{n-1})$ and $(Q_0, Q_1, \ldots, Q_{n-1})$ respectively, where F_n is the Fibonacci numbers, L_n is the Lucas numbers, P_n is the Pell numbers and Q_n is the Pell-Lucas numbers. In this paper, by using the inverse factorization of polynomial of degree n, the explicit determinants of matrices A and C are given only by the four famous numbers.

Keywords: RSFMLR Circulant matrix, RSLMFL Circulant matrix, Determinant, Fibonacci numbers, Lucas numbers, Pell numbers, Pell-Lucas numbers, Codes.

1   Introduction

The Fibonacci, Lucas, Pell and Pell-Lucas sequences [1] are defined by the following recurrence relations, respectively:

$$F_{n+1} = F_n + F_{n-1} \text{ where } F_0 = 0 \; F_1 = 1$$
$$L_{n+1} = L_n + L_{n-1} \text{ where } L_0 = 2 \; L_1 = 1,$$
$$P_{n+1} = 2P_n + P_{n-1} \text{ where } P_0 = 0 \;, P_1 = 1 \;,$$
$$Q_{n+1} = 2Q_n + Q_{n-1} \text{ where } Q_0 = 2 \;, Q_1 = 2.$$

The $\{F_n\}$ is given by the Binet formula $F_n = \dfrac{\alpha^n - \beta^n}{\alpha - \beta}$ and the $\{L_n\}$ is given by the Binet formula $L_n = \alpha^n + \beta^n$, where α and β are the roots of the characteristic equation $x^2 - x - 1 = 0$. The $\{P_n\}$ is given by the Binet formula

---

$$P_n = \frac{\alpha_1^n - \beta_1^n}{\alpha_1 - \beta_1}$$ and the $\{Q_n\}$ is given by the Binet formula $Q_n = \alpha_1^n + \beta_1^n$,

where α_1 and β_1 are the roots of the characteristic equation $x^2 - 2x - 1 = 0$.

Recently, there are many interests in properties and generalization of some special matrices with famous numbers. Jaiswal evaluated some determinants of circulant whose elements are the generalized Fibonacci numbers [2]. Lin gave the determinant of the Fibonacci-Lucas quasi-cyclic matrices [3]. Lind presented the determinants of circulant and skew-circulant involving Fibonacci numbers in [4]. Shen [5] discussed the determinant of circulant matrix involving Fibonacci and Lucas numbers. Akbulak [6] gave the norms of Toeplitz involving Fibonacci and Lucas numbers. The authors [7, 8] discussed some properties of Fibonacci and Lucas numbers. Stanimirović gave generalized Fibonacci and Lucas matrix in [9]. Zhang [10] investigated the Lucas matrix and some combinatorial identities.

Definition 1. A row skew first-minus-last right (RSFMLR) circulant matrix with the first row $(a_0, a_1, \ldots, a_{n-1})$, denoted by $RSFMLRcircfr(a_0, a_1, \ldots, a_{n-1})$, is meant a square matrix of the form

$$A = \begin{pmatrix} a_0 & a_1 & \cdots & a_{n-2} & a_{n-1} \\ -a_{n-1} & a_0 - a_{n-1} & a_1 & \cdots & a_{n-2} \\ \vdots & -a_{n-1} - a_{n-2} & \ddots & \ddots & \vdots \\ -a_2 & \vdots & \cdots & \ddots & a_1 \\ -a_1 & -a_2 - a_1 & \cdots & -a_{n-1} - a_{n-2} & a_0 - a_{n-1} \end{pmatrix}_{n \times n}.$$

It can be seen that the matrix with an arbitrary first row and the following rule for obtaining any other row from the previous one: Get the $i+1^{st}$ row through the first element of the i th row minus the last element of the i th row, and -1 times the last element of the i th row , and then shifting the elements of the i th row (cyclically) one position to the right.

Obviously, the RSFMLR circulant matrix is a $x^n + x + 1$ circulant matrix [11-13], and that is neither the extension of circulant matrix [5, 14] nor its special case and they are two different kinds of special matrices. Moreover, it is a RSFP r LR circulant matrix [15] and is also a ULS r -circulant matrix [16] with $r = -1$.

We define $\Theta_{(-1,-1)}$ as the basic RSFMLR circulant matrix, that is

$$\Theta_{(-1,-1)} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & 0 & 1 \\ -1 & -1 & 0 & \cdots & 0 \end{pmatrix}_{n \times n} = RSFMLRcircfr(0,1,0,\ldots,0).$$

It is easily verified that $g(x) = x^n + x + 1$ has no repeated roots in its splitting field and $g(x) = x^n + x + 1$ is both the minimal polynomial and the characteristic polynomial of the matrix $\Theta_{(-1,-1)}$. In addition, $\Theta_{(-1,-1)}$ is nonderogatory and satisfies $\Theta_{(-1,-1)}^j = RSFMLRcircfr(0,\ldots,0,1,0,\ldots,0)$ and

$$\Theta_{(-1,-1)}^n = -I_n - \Theta_{(-1,-1)}.$$

In view of the structure of the powers of the basic RSFMLR circulant matrix $\Theta_{(-1,-1)}$, it is clear that

$$A = RSFMLRcircfr(a_0, a_1, \ldots, a_{n-1}) = \sum_{i=0}^{n-1} a_i \Theta_{(-1,-1)}^i \tag{1}$$

Thus, A is a RSFMLR circulant matrix if and only if $A = f(\Theta_{(-1,-1)})$ for some polynomial $f(x)$. The polynomial $f(x) = \sum_{i=0}^{n-1} a_i x^i$ will be called the representer of the RSFMLR circulant matrix A.

In addition to the algebraic properties that can be easily derived, we can know that RSFMLR circulant matrices have very nice structure. The product of two RSFMLR circulant matrices is a RSFMLR circulant matrix.

Definition 2. A row skew last-minus-first (RSLMFL) circulant matrix with the first row $(a_0, a_1, \ldots, a_{n-1})$, denoted by $RSLMFLcircfr(a_0, a_1, \ldots, a_{n-1})$, is meant a square matrix of the form

$$A = \begin{pmatrix} a_0 & a_1 & \cdots & a_{n-1} & a_{n-1} \\ a_1 & a_2 & \cdot^{\cdot} & a_{n-1} - a_0 & -a_0 \\ \vdots & \cdot^{\cdot} & \cdot^{\cdot} & -a_0 - a_1 & \vdots \\ a_{n-2} & \cdot^{\cdot} & \cdot^{\cdot} & \vdots & -a_{n-3} \\ a_{n-1} - a_0 & -a_0 - a_1 & \cdots & -a_{n-3} - a_{n-2} & -a_{n-2} \end{pmatrix}_{n\times n}.$$

It can be seen that the matrix with an arbitrary first row and the following rule for obtaining any other row from the previous one: Get the $i+1^{st}$ row through the last element of the ith row minus the first element of the ith row, and -1 times the first element of the ith row, and then shifting the elements of the ith row (cyclically) one position to the left.

By Proposition 5.1 in [13], we deduce the following lemma:

Lemma 1. Let $A = RSFMLRcircfr(a_0, a_1, \ldots, a_{n-1})$, then the eigenvalues of A are

$$f(\varepsilon_k) = \sum_{i=0}^{n-1} (a_i \varepsilon^i),$$

in addition,

$$\det A = \prod_{k=1}^{n} \sum_{i=0}^{n-1} (a_i \varepsilon^i),$$

where $\varepsilon_k (\mathrm{k} = 1,2,\ldots,n)$ are the roots of the

$$x^n + x + 1 = 0 \qquad (2)$$

Lemma 2. $\displaystyle\prod_{k=1}^{n} (c + \varepsilon_k b + \varepsilon_k^2 a)$

$$= c^n + c[(as)^{n-1} + (at)^{n-1}] + [(as)^n + (at)^n] + a^{n-1}(c - b + a),$$

where

$$s = \frac{-b + \sqrt{b^2 - 4ac}}{2a}, \quad t = \frac{-b - \sqrt{b^2 - 4ac}}{2a},$$

and $\varepsilon_k (\mathrm{k} = 1,2,\ldots,n)$ satisfy the equation $(2), a,b,c \in R, a \neq 0$.

Proof.

$$\prod_{k=1}^{n} (c + \varepsilon_k b + \varepsilon_k^2 a) = a^n \prod_{k=1}^{n} (\varepsilon_k^2 + \frac{b}{a}\varepsilon_k + \frac{c}{a}) = a^n \prod_{k=1}^{n} (s - \varepsilon_k)(t - \varepsilon_k),$$

while

$$s + t = -\frac{b}{a}, \quad st = \frac{c}{a}, \quad s = \frac{-b + \sqrt{b^2 - 4ac}}{2a}, \quad t = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

Since $\varepsilon_k (\mathrm{k} = 1,2,\ldots,n)$ satisfy the equation (2), we must have

$$x^n + x + 1 = \prod_{k=1}^{n} (x - \varepsilon_k).$$

So $\displaystyle\prod_{k=1}^{n} (c + \varepsilon_k b + \varepsilon_k^2 a) = a^n (s^n + s + 1)(t^n + t + 1)$

$$= a^n [(st)^n + st(s^{n-1} + t^{n-1}) + (s^n + t^n) + (s + t + st + 1)]$$

$$= c^n + c[(as)^{n-1} + (at)^{n-1}] + [(as)^n + (at)^n] + a^{n-1}(c - b + a)$$

2   Determinants of the RSFMLR and RSLMFL Circulant Matrix with the Four Famous Numbers

In this section, we discuss the explicit determinants of RSFMLR and RSLMFL circulant matrices are given only by the four famous numbers.

Theorem 1. If $F = RSFMLRcircfr(F_0, F_1, \ldots, F_{n-1})$ is a RSFMLR circulant matrix, then

$$\det F = \frac{F_n^n - F_{n-1}^{n-1} + F_{n-1}^{n-1}F_n(x_1^{n-1} + y_1^{n-1}) + F_{n-1}^n(x_1^n + y_1^n)}{1 + L_{n-1} + L_n + (-1)^{n-1}},$$

where

$$x_1 = \frac{-F_{n-1} - F_n - 1 + \sqrt{(F_n - F_{n-1})^2 + 2(F_n + F_{n-1}) + 1}}{2F_{n-1}},$$

$$y_1 = \frac{-F_{n-1} - F_n - 1 - \sqrt{(F_n - F_{n-1})^2 + 2(F_n + F_{n-1}) + 1}}{2F_{n-1}}.$$

Proof. The matrix F can be writen as

$$F = \begin{pmatrix} F_0 & F_1 & \cdots & F_{n-2} & F_{n-1} \\ -F_{n-1} & F_0 - F_{n-1} & F_1 & \cdots & F_{n-2} \\ \vdots & -F_{n-1} - F_{n-2} & \ddots & \ddots & \vdots \\ -F_2 & \vdots & \cdots & \ddots & F_1 \\ -F_1 & -F_2 - F_1 & \cdots & -F_{n-1} - F_{n-2} & F_0 - F_{n-1} \end{pmatrix}_{n \times n},$$

Using the Lemma 1, the determinant of F is

$$\det F = \prod_{k=1}^n (F_0 + F_1\varepsilon_k + \cdots + F_{n-1}\varepsilon_k^{n-1})$$

$$= \prod_{k=1}^n \left(\frac{\alpha - \beta}{\alpha - \beta}\varepsilon_k + \frac{\alpha^2 - \beta^2}{\alpha - \beta}\varepsilon_k^2 + \cdots + \frac{\alpha^{n-1} - \beta^{n-1}}{\alpha - \beta}\varepsilon_k^{n-1}\right)$$

$$= \prod_{k=1}^n \left(\frac{F_n + \varepsilon_k(F_{n-1} + F_n + 1) + \varepsilon_k^2 F_{n-1}}{1 - \varepsilon_k - \varepsilon_k^2}\right).$$

According to Lemma 2, we can get

$$\det F = \frac{F_n^n - F_{n-1}^{n-1} + F_{n-1}^{n-1}F_n(x_1^{n-1} + y_1^{n-1}) + F_{n-1}^n(x_1^n + y_1^n)}{1 + L_{n-1} + L_n + (-1)^{n-1}},$$

where

$$x_1 = \frac{-F_{n-1} - F_n - 1 + \sqrt{(F_n - F_{n-1})^2 + 2(F_n + F_{n-1}) + 1}}{2F_{n-1}},$$

$$y_1 = \frac{-F_{n-1} - F_n - 1 - \sqrt{(F_n - F_{n-1})^2 + 2(F_n + F_{n-1}) + 1}}{2F_{n-1}}.$$

Using the method in Theorem 1 similarly, we also have

Theorem 2. If $F' = RSFMLRcircfr(F_{n-1}, F_{n-2}, \cdots, F_0)$ is a RSFMLR circulant matrix, then

$$\det F' = \frac{(-1-F_{n-1})^n - (F_n+1)^{n-1}(-F_n+F_{n-1})}{(-1)^n - L_{n-1} + L_n + 1}.$$

Theorem 3. If $H = RSFMLRcircfr(F_0, F_1, \cdots, F_{n-1})$ is a RSLMFL circulant matrix, then we have

$$\det H = \frac{(-1-F_{n-1})^n - (F_n+1)^{n-1}(-F_n+F_{n-1})}{(-1)^n - L_{n-1} + L_n + 1}(-1)^{\frac{n(n-1)}{2}}.$$

Proof. The matrix H can be writen as

$$H = \begin{pmatrix} F_0 & F_1 & \cdots & F_{n-2} & F_{n-1} \\ F_1 & F_2 & \cdot\!\cdot\!\cdot & F_{n-1}-F_0 & -F_0 \\ \vdots & \cdot\!\cdot\!\cdot & \cdot\!\cdot\!\cdot & -F_0-F_1 & \vdots \\ F_{n-2} & \cdot\!\cdot\!\cdot & \cdot\!\cdot\!\cdot & \vdots & -F_{n-3} \\ F_{n-1}-F_0 & -F_0-F_1 & \cdots & -F_{n-3}-F_{n-2} & -F_{n-2} \end{pmatrix}$$

$$= \begin{pmatrix} F_{n-1} & F_{n-2} & \cdots & F_1 & F_0 \\ -F_0 & F_{n-1}-F_0 & F_{n-2} & \cdots & F_1 \\ \vdots & -F_0-F_1 & \ddots & \ddots & \vdots \\ -F_{n-3} & \vdots & \ddots & \ddots & F_{n-2} \\ -F_{n-2} & -F_{n-3}-F_{n-2} & \cdots & -F_0-F_1 & F_{n-1}-F_0 \end{pmatrix} \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & \cdots & 0 & 1 & 0 \\ \vdots & \cdot\!\cdot\!\cdot & \cdot\!\cdot\!\cdot & \cdot\!\cdot\!\cdot & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

$$= F'\Gamma,$$

then we can get $\det H = \det F' \det \Gamma$,

where $F' = RSFMLRcircfr(F_{n-1}, F_{n-2}, \cdots, F_0)$, and its determinant could be obtained through Theorem 2,

$$\det F' = \frac{(-1-F_{n-1})^n - (F_n+1)^{n-1}(-F_n+F_{n-1})}{(-1)^n - L_{n-1} + L_n + 1},$$

and since $\det \Gamma = (-1)^{\frac{n(n-1)}{2}}$,

so $\det H = \det F' \det \Gamma = \dfrac{(-1-F_{n-1})^n - (F_n+1)^{n-1}(-F_n+F_{n-1})}{(-1)^n - L_{n-1} + L_n + 1}(-1)^{\frac{n(n-1)}{2}}.$

Theorem 4. If $L = RSFMLRcircfr(L_0, L_1, \ldots, L_{n-1})$ is a RSFMLR circulant matrix, then we have

$$\det L = \frac{(2+L_n)^n + L_{n-1}^{n-1}[(2+L_n)(x_2^{n-1} + y_2^{n-1}) + L_{n-1}(x_2^n + y_2^n) + 3]}{1 + L_{n-1} + L_n + (-1)^{n-1}},$$

where

$$x_2 = \frac{1 - L_{n-1} - L_n + \sqrt{(L_n - L_{n-1})^2 - 10L_{n-1} - 2L_n + 1}}{2L_{n-1}},$$

$$y_2 = \frac{1 - L_{n-1} - L_n - \sqrt{(L_n - L_{n-1})^2 - 10L_{n-1} - 2L_n + 1}}{2L_{n-1}}.$$

Proof. The method is similar to the Theorem 1.

Using the method in Theorem 1 similarly, we also have

Theorem 5. If $L' = RSFMLRcircfr(L_{n-1}, \ldots, L_1, L_0)$ is a RSFMLR circulant matrix, then we get

$$\det L' = \frac{(-1)^n[(-1+L_{n-1})^n + 2^{n-1}(-1+L_{n-1})(x_3^{n-1} + y_3^{n-1})]}{(-1)^n - L_{n-1} + L_n + 1}$$

$$+ \frac{(-1)^n[2^n(x_3^n + y_3^n) + 2^{n-1}(L_{n-1} - L_n)]}{(-1)^n - L_{n-1} + L_n + 1},$$

where

$$x_3 = \frac{1 + L_n + \sqrt{(1+L_n)^2 + 8(1-L_{n-1})}}{-4},$$

$$y_3 = \frac{1 + L_n - \sqrt{(1+L_n)^2 + 8(1-L_{n-1})}}{-4}.$$

Using the method in Theorem 3 similarly, we also have

Theorem 6. If $H' = RSLMFLcircfr(L_0, L_1, \ldots, L_{n-1})$ is a RSLMFL circulant matrix, then

$$\det H' = \frac{(-1)^n[(-1+L_{n-1})^n + 2^{n-1}(-1+L_{n-1})(x_3^{n-1} + y_3^{n-1})]}{(-1)^n - L_{n-1} + L_n + 1}(-1)^{\frac{n(n-1)}{2}}$$

$$+ \frac{(-1)^n[2^n(x_3^n + y_3^n) + 2^{n-1}(L_{n-1} - L_n)]}{(-1)^n - L_{n-1} + L_n + 1}(-1)^{\frac{n(n-1)}{2}},$$

where

$$x_3 = \frac{1 + L_n + \sqrt{(1+L_n)^2 + 8(1-L_{n-1})}}{-4},$$

$$y_3 = \frac{1 + L_n - \sqrt{(1+L_n)^2 + 8(1-L_{n-1})}}{-4}.$$

Theorem 7. If $P = RSFMLRcircfr(P_0, P_1, \ldots, P_{n-1})$ is a RSFMLR circulant matrix, then

$$\det P = \frac{P_n^n + P_{n-1}^{n-1}[P_n(x^{n-1} + y^{n-1}) + P_{n-1}(x^n + y^n) - 1]}{1 + Q_{n-1} + Q_n + 2(-1)^{n-1}},$$

where $x = \dfrac{-P_{n-1} - P_n - 1 + \sqrt{(P_n - P_{n-1})^2 + 2(P_n + P_{n-1}) + 1}}{2P_{n-1}}$,

$$y = \frac{-P_{n-1} - P_n - 1 - \sqrt{(P_n - P_{n-1})^2 + 2(P_n + P_{n-1}) + 1}}{2P_{n-1}}.$$

Proof. The matrix P can be writen as

$$P = \begin{pmatrix} P_0 & P_1 & \cdots & P_{n-2} & P_{n-1} \\ -P_{n-1} & P_0 - P_{n-1} & P_1 & \cdots & P_{n-2} \\ \vdots & -P_{n-1} - P_{n-2} & \ddots & \ddots & \vdots \\ -P_2 & \vdots & \cdots & \ddots & P_1 \\ -P_1 & -P_2 - P_1 & \cdots & -P_{n-1} - P_{n-2} & P_0 - P_{n-1} \end{pmatrix}_{n \times n},$$

Using the Lemma 1, the determinant of P is

$$\det P = \prod_{k=1}^{n} (P_0 + P_1 \varepsilon_k + \cdots + P_{n-1} \varepsilon_k^{n-1})$$

$$= \prod_{k=1}^{n} (\frac{\alpha - \beta}{\alpha - \beta} \varepsilon_k + \frac{\alpha^2 - \beta^2}{\alpha - \beta} \varepsilon_k^2 + \cdots + \frac{\alpha^{n-1} - \beta^{n-1}}{\alpha - \beta} \varepsilon_k^{n-1})$$

$$= \prod_{k=1}^{n} (\frac{P_n + \varepsilon_k(P_{n-1} + P_n + 1) + \varepsilon_k^2 P_{n-1}}{1 - 2\varepsilon_k - \varepsilon_k^2}).$$

According to Lemma 2, we can get

$$\det P = \frac{P_n^n + P_{n-1}^{n-1}[P_n(x^{n-1} + y^{n-1}) + P_{n-1}(x^n + y^n) - 1]}{1 + Q_{n-1} + Q_n + 2(-1)^{n-1}},$$

where

$$x = \frac{-P_{n-1} - P_n - 1 + \sqrt{(P_n - P_{n-1})^2 + 2(P_n + P_{n-1}) + 1}}{2P_{n-1}},$$

$$y = \frac{-P_{n-1} - P_n - 1 - \sqrt{(P_n - P_{n-1})^2 + 2(P_n + P_{n-1}) + 1}}{2P_{n-1}}.$$

Using the method in Theorem 7 similarly, we also have

Theorem 8. If $P' = RSFMLRcircfr(P_{n-1}, P_{n-2}, \cdots, P_0)$ is a RSFMLR circulant matrix, then

$$\det P' = \frac{(-1-P_{n-1})^n - (P_n+1)^{n-1}(-P_n+P_{n-1})}{(-1)^n - Q_{n-1} + Q_n + 2}.$$

Theorem 9. If $C = RSFMLRcircfr(P_0, P_1, \cdots, P_{n-1})$ is a RSLMFL circulant matrix, then we have

$$\det C = \frac{(-1-P_{n-1})^n - (P_n+1)^{n-1}(-P_n+P_{n-1})}{(-1)^n - Q_{n-1} + Q_n + 2}(-1)^{\frac{n(n-1)}{2}}.$$

Proof. The matrix C can be writen as

$$C = \begin{pmatrix} P_0 & P_1 & \cdots & P_{n-2} & P_{n-1} \\ P_1 & P_2 & \ddots & P_{n-1}-P_0 & -P_0 \\ \vdots & \ddots & \ddots & -P_0-P_1 & \vdots \\ P_{n-2} & \ddots & \ddots & \vdots & -P_{n-3} \\ P_{n-1}-P_0 & -P_0-P_1 & \cdots & -P_{n-3}-P_{n-2} & -P_{n-2} \end{pmatrix}$$

$$= \begin{pmatrix} P_{n-1} & P_{n-2} & \cdots & P_1 & P_0 \\ -P_0 & P_{n-1}-P_0 & P_{n-2} & \cdots & P_1 \\ \vdots & -P_0-P_1 & \ddots & \ddots & \vdots \\ -P_{n-3} & \vdots & \ddots & \ddots & P_{n-2} \\ -P_{n-2} & -P_{n-3}-P_{n-2} & \cdots & -P_0-P_1 & P_{n-1}-P_0 \end{pmatrix} \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & \cdots & 0 & 1 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

$$= P'\Gamma,$$

then we can get

$$\det C = \det P' \det \Gamma,$$

where $P' = RSFMLRcircfr(P_{n-1}, P_{n-2}, \cdots, P_0)$, and its determinant could be obtained through Theorem 8,

$$\det P' = \frac{(-1-P_{n-1})^n - (P_n+1)^{n-1}(-P_n+P_{n-1})}{(-1)^n - Q_{n-1} + Q_n + 2},$$

and since $\det \Gamma = (-1)^{\frac{n(n-1)}{2}}$,

so $\det C = \det P' \det \Gamma = \dfrac{(-1-P_{n-1})^n - (P_n+1)^{n-1}(-P_n+P_{n-1})}{(-1)^n - Q_{n-1} + Q_n + 2}(-1)^{\frac{n(n-1)}{2}}.$

Theorem 10. If $Q = RSFMLRcircfr(Q_0, Q_1, \ldots, Q_{n-1})$ is a RSFMLR circulant matrix, then we have

$$\det Q = \frac{(2+Q_n)^n + Q_{n-1}^{n-1}[(2+Q_n)(a^{n-1}+b^{n-1}) - Q_{n-1}(a^n + b^n) + 4]}{1 + Q_{n-1} + Q_n + 2(-1)^{n-1}},$$

where

$$a = \frac{2 - Q_{n-1} - Q_n + \sqrt{(Q_n - Q_{n-1})^2 - 12Q_{n-1} - 4Q_n + 4}}{2Q_{n-1}},$$

$$b = \frac{2 - Q_{n-1} - Q_n - \sqrt{(Q_n - Q_{n-1})^2 - 12Q_{n-1} - 4Q_n + 4}}{2Q_{n-1}}.$$

Proof. The method is similar to the Theorem 7.

Using the method in Theorem 7 similarly, we also have

Theorem 11. If $Q' = RSFMLRcircfr(Q_{n-1}, \ldots, Q_1, Q_0)$ is a RSFMLR circulant matrix, then we get

$$\det Q' = \frac{(2 - Q_{n-1})^n - 2^{n-1}(-1)^n K}{(-1)^n - Q_{n-1} + Q_n + 2},$$

where

$$K = (2 - Q_{n-1})(g^{n-1} + h^{n-1}) - 2(g^n + h^n) + (Q_n - Q_{n-1}),$$

$$g = \frac{Q_n + \sqrt{Q_n^2 - 8Q_{n-1} + 16}}{-4}, \quad h = \frac{Q_n - \sqrt{Q_n^2 - 8Q_{n-1} + 16}}{-4}.$$

Using the method in Theorem 9 similarly, we also have

Theorem 12. If $G = RSLMFLcircfr(Q_0, Q_1, \ldots, Q_{n-1})$ is a RSLMFL circulant matrix, then

$$\det G = \frac{(2 - Q_{n-1})^n - 2^{n-1}(-1)^n K}{(-1)^n - Q_{n-1} + Q_n + 2}(-1)^{\frac{n(n-1)}{2}},$$

where

$$K = (2 - Q_{n-1})(g^{n-1} + h^{n-1}) - 2(g^n + h^n) + (Q_n - Q_{n-1}),$$

$$g = \frac{Q_n + \sqrt{Q_n^2 - 8Q_{n-1} + 16}}{-4}, \quad h = \frac{Q_n - \sqrt{Q_n^2 - 8Q_{n-1} + 16}}{-4}.$$

References

1. Melham, R.: Sums Involving Fibonacci and Pell Numbers. Port. Math. 56, 209–317 (1999)
2. Jaiswal, D.: On Determinants Involving Generalized Fibonacci Numbers. Fibonacci Quart. 7, 319–330 (1969)
3. Lin, D.: Fibonacci-Lucas Quasi-cyclic Matrices. Fibonacci Quart. 40, 280–286 (2002)
4. Lind, D.: A Fibonacci Circulant. Fibonacci Quart. 8, 449–455 (1970)
5. Shen, S.Q., Cen, J.M., Hao, Y.: On the Determinants and Inverses of Circulant Matrices with Fibonacci and Lucas Numbers. Appl. Math. Comput. 217, 9790–9797 (2011)
6. Akbulak, M., Bozkurt, D.: On the Norms of Toeplitz Matrices Involving Fibonacci and Lucas Numbers. Hacet. J. Math. Stat. 37, 89–95 (2008)
7. Lee, G.Y., Kim, J.S., Lee, S.G.: Factorizations and Eigenvalues of Fibonacci and Symmetric Fibonacci Matrices. Fibonacci Quart. 40, 203–211 (2002)
8. Miladinović, M., Stanimirović, P.: Singular Case of Generalized Fibonacci and Lucas Matrices. J. Korean Math. Soc. 48, 33–48 (2011)
9. Stanimirović, P., Nikolov, J., Stanimiroviá, I.: A Generalization of Fibonacci and Lucas Matrices. Discrete Appl. Math. 156, 2606–2619 (2008)
10. Zhang, Z., Zhang, Y.: The Lucas Matrix and Some Combinatorial Identities. Indian J. Pure Appl. Math. 38, 457–465 (2007)
11. Jiang, X.Y., Gao, Y., Jiang, Z.L.: Explicit Determinants and Inverses of Skew Circulant and Skew Left Circulant Matrices Involving the k-Fibonacci Numbers in Communications-I. Far East Journal of Mathematical Sciences 76(1), 123–137 (2013)
12. Jiang, X.Y., Gao, Y., Jiang, Z.L.: Explicit Determinants and Inverses of Skew Circulant and Skew Left Circulant Matrices Involving the k-Lucas Numbers in Communications-II. Accepted by Far East Journal of Mathematical Sciences
13. David, C.: Regular Representations of Semisimple Algebras, Separable Field Extensions, Group Characters. Generalized Circulants, and Generalized Cyclic Codes. Linear Algebra and its Appl. 218, 147–183 (1995)
14. Jiang, Z.L., Zhou, Z.X.: Circulant Matrices. Chengdu Technology University Publishing Company, Chengdu (1999)
15. Jiang, X.Y., Jiang, Z.L.: Efficient Algorithm for Finding Inverse and Group Inverse of the RSFPrLR Circulant Matrix in Codes. JP Journal of Algebra, Number Theory and Applications 29(1), 51–70 (2013)
16. Tian, Z.P.: Fast Algorithms for Solving the Inverse Problem of $Ax = b$ in the Class of the ULS r-circulant (retrocirculant) Matrices. Int. J. Algebra 5, 9403–9411 (2011)

Explicit Determinants of the *k*-Fibonacci and *k*-Lucas RSFPLR Circulant Matrix in Codes

Xiao-Yu Jiang and Kicheon Hong[*]

Dept. of Information and Telecommunications Engineering, The University of Suwon,
Wau-ri, Bongdam-eup, Hwaseong-si, Gyeonggi-do, 445-743, Korea
jxy19890422@gmail.com, Kchong@suwon.ac.kr

Abstract. Let A be a row skew first-plus-last right circulant matrix in codes and C be a row skew last-plus-first left circulant matrix whose first row are $(F_{k,1}, F_{k,2}, \ldots, F_{k,n})$ and $(L_{k,1}, L_{k,2}, \ldots, L_{k,n})$ respectively, where $\{F_{k,n}\}$ is the k-Fibonacci number and $\{L_{k,n}\}$ is the k-Lucas number. In this paper, by using the inverse factorization of polynomial of degree n, the explicit determinants of matrices A and C are expressed by utilizing the k-Fibonacci and k-Lucas numbers.

Keywords: Determinant, k-Fibonacci number, k-Lucas number, RSFPLR circulant matrix, RSLPFL circulant matrix, Codes.

1 Introduction

The k-Fibonacci and k-Lucas sequences [1] are defined by the following recurrence relations, respectively:

$$F_{k,n+1} = kF_{k,n} + F_{k,n-1} \ \text{ where } F_{k,0} = 0, \ \ F_{k,1} = 1,$$
$$L_{k,n+1} = kL_{k,n} + L_{k,n-1} \ \text{ where } L_{k,0} = 2, \ \ L_{k,1} = k.$$

The sequence $\{F_{k,n}\}$ is given by the Binet formula $F_{k,n} = \dfrac{\alpha^n - \beta^n}{\alpha - \beta}$ and the $\{L_{k,n}\}$ is given by the Binet formula $L_{k,n} = \alpha^n + \beta^n$, where α and β are the roots of the characteristic equation $x^2 - kx - 1 = 0$.

There are many interests in properties and generalization of some special matrices with famous numbers. Jaiswal evaluated some determinants of circulant whose elements are the generalized Fibonacci numbers [2]. Lin gave the determinant of the Fibonacci-Lucas quasi-cyclic matrices [3]. Lind presented the determinants of circulant and skew-circulant involving Fibonacci numbers in [4]. Shen [5] discussed

---

[*] Corresponding author.

the determinant of circulant matrix involving Fibonacci and Lucas numbers. Akbulak [6] gave the norms of Toeplitz involving Fibonacci and Lucas numbers. The authors [7, 8] discussed some properties of Fibonacci and Lucas numbers. Stanimirović gave generalized Fibonacci and Lucas matrix in [9]. Zhang [10] investigated the Lucas matrix and some combinatorial identities.

In this paper, we define matrices of forms: $A = \text{RSFPLRcircfr}(a_0, a_1, \ldots, a_{n-1})$ and $C = \text{RSLPFLcircfr}(a_0, a_1, \ldots, a_{n-1})$ be $n \times n$ matrices whose first row are $(F_{k,1}, F_{k,2}, \ldots, F_{k,n})$ and $(L_{k,1}, L_{k,2}, \ldots, L_{k,n})$ respectively, by using the inverse factorization of polynomial of degree n, the explicit determinants of these matrices are given only by the k-Fibonacci and k-Lucas numbers.

Definition 1. A row skew first-plus-last right (RSFPLR) circulant matrix with the first row $(a_0, a_1, \ldots, a_{n-1})$, denoted by $\text{RSFPLRcircfr}(a_0, a_1, \ldots, a_{n-1})$, is meant a square matrix of the form

$$A = \begin{pmatrix} a_0 & a_1 & \cdots & a_{n-2} & a_{n-1} \\ -a_{n-1} & a_0 + a_{n-1} & a_1 & \cdots & a_{n-2} \\ \vdots & -a_{n-1} + a_{n-2} & \ddots & \ddots & \vdots \\ -a_2 & \vdots & \cdots & \ddots & a_1 \\ -a_1 & -a_2 + a_1 & \cdots & -a_{n-1} + a_{n-2} & a_0 + a_{n-1} \end{pmatrix}_{n \times n}.$$

It can be seen that the matrix with an arbitrary first row and the following rule for obtaining any other row from the previous one: Get the $i+1$st row by adding the last element of the ith row to the first element of the ith row, and -1 times the last element of the ith row, and then shifting the elements of the ith row (cyclically) one position to the right.

Note that the RSFPLR circulant matrix is a $x^n - x + 1$ circulant matrix [12], and that is neither the extension of circulant matrix [13] nor its special case and they are two different kinds of special matrices. Moreover, it is a FLS r-circulant matrix [11] with $r = -1$.

If the first row of a RSFPLR circulant matrix be $(F_{k,1}, F_{k,2}, \ldots, F_{k,n})$, then the matrix is called k-Fibonacci RSFPLR circulant matrix. And if the first row of a RSFPLR circulant matrix be $(L_{k,1}, L_{k,2}, \ldots, L_{k,n})$, then the matrix is called k-Lucas RSFPLR circulant matrix. Specially, if the first row of a RSFPLR circulant matrix be $(F_1, F_2, \ldots, F_n)$, then the matrix is called Fibonacci RSFPLR circulant matrix. And if the first row of a RSFPLR circulant matrix be $(L_1, L_2, \ldots, L_n)$, then the matrix is called Lucas RSFPLR circulant matrix.

We define $\Theta_{(-1,1)}$ as the basic RSFPLR circulant matrix, that is

$$\Theta_{(-1,1)} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & 0 & 1 \\ -1 & 1 & 0 & \cdots & 0 \end{pmatrix}_{n \times n}$$

(1)

It is easily verified that $g(x) = x^n - x + 1$ has no repeated roots in its splitting field and $g(x) = x^n - x + 1$ is both the minimal polynomial and the characteristic polynomial of the matrix $\Theta_{(-1,1)}$. In addition, $\Theta_{(-1,1)}$ is nonderogatory and satisfies

$$\Theta_{(-1,1)}^j = \text{RSFPLRcircfr}(0,\ldots,0,1,0,\ldots,0) \quad \text{and} \quad \Theta_{(-1,1)}^n = -I_n + \Theta_{(-1,1)}.$$

Then a matrix A can be written in the from:

$$A = f(\Theta_{(-1,1)}) = \sum_{i=0}^{n-1} a_i \Theta_{(-1,1)}^i \tag{2}$$

if and only if A is a RSFPLR circulant matrix, where the polynomial $f(x) = \sum_{i=0}^{n-1} a_i x^i$ is called the represester of the RSFPLR circulant matrix A.

Definition 2. A row skew last-plus-first left (RSLPFL) circulant matrix with the first row $(a_0, a_1, \ldots, a_{n-1})$, denoted by $\text{RSLPFLcircfr}(a_0, a_1, \ldots, a_{n-1})$, is meant a square matrix of the form

$$A = \begin{pmatrix} a_0 & a_1 & \cdots & a_{n-1} & a_{n-1} \\ a_1 & a_2 & \ddots & a_{n-1}+a_0 & -a_0 \\ \vdots & \ddots & \ddots & -a_0+a_1 & \vdots \\ a_{n-2} & \ddots & \ddots & \vdots & -a_{n-3} \\ a_{n-1}+a_0 & -a_0+a_1 & \cdots & -a_{n-3}+a_{n-2} & -a_{n-2} \end{pmatrix}_{n \times n}.$$

It can be seen that the matrix with an arbitrary first row and the following rule for obtaining any other row from the previous one: Get the $i+1$ st row by adding the first element of the i th row to the last element of the i th row, and -1 times the first element of the i th row, and then shifting the elements of the i th row (cyclically) one position to the left.

If the first row of a RSLPFL circulant matrix be $(F_{k,1}, F_{k,2}, \ldots, F_{k,n})$, $(L_{k,1}, L_{k,2}, \ldots, L_{k,n})$, then the matrix is called k-Fibonacci, k-Lucas RSLPFL circulant matrix, respectively. Specially, If the first row of a RSLPFL circulant matrix be $(F_1, F_2, \ldots, F_n)$, $(L_1, L_2, \ldots, L_n)$, then the matrix is called Fibonacci, Lucas RSLPFL circulant matrix, respectively.

By Theorem 1 in [11], we deduce the following lemma.

Lemma 1. Let $A = \text{RSFPLRcircfr}(a_0, a_1, \ldots, a_{n-1})$, then we have

$$\lambda_i = f(\varepsilon_i) = \sum_{j=0}^{n-1} a_j \varepsilon_i^j \text{ and } \det A = \prod_{i=1}^{n} \sum_{j=0}^{n-1} (a_j \varepsilon_i^j),$$

where $\varepsilon_i (i = 1, 2, \ldots, n)$ are the roots of the

$$x^n - x + 1 = 0 \qquad (3)$$

Lemma 2. $\prod_{i=1}^{n}(y - \varepsilon_i z + \varepsilon_i^2 x) = y^n - y[(ax)^{n-1} + (bx)^{n-1}] + [(ax)^n + (bx)^n] + x^{n-1}(y - z + x),$

where

$$a = \frac{z + \sqrt{z^2 - 4xy}}{2x}, \qquad\qquad b = \frac{z - \sqrt{z^2 - 4xy}}{2x},$$

and $\varepsilon_i (i = 1, 2, \ldots, n)$ satisfy the equation (3), $x, y, z \in R, a \neq 0$.

Proof. $\prod_{i=1}^{n}(y - \varepsilon_i z + \varepsilon_i^2 x) = x^n \prod_{i=1}^{n}(\varepsilon_i^2 - \frac{z}{x}\varepsilon_i + \frac{y}{x})$

$= x^n \prod_{i=1}^{n}(a - \varepsilon_i)(b - \varepsilon_i),$

while

$$a + b = \frac{z}{x}, \quad ab = \frac{y}{x}, \quad a = \frac{z + \sqrt{z^2 - 4xy}}{2x}, \quad b = \frac{z - \sqrt{z^2 - 4xy}}{2x}.$$

Since $\varepsilon_i (i = 1, \ldots, n)$ satisfy the equation (3), we must have

$$x^n - x + 1 = \prod_{i=1}^{n}(x - \varepsilon_i).$$

So $\prod_{i=1}^{n}(y - \varepsilon_i z + \varepsilon_i^2 x) = x^n (a^n - a + 1)(b^n - b + 1)$

$= x^n [(ab)^n + ab - ab(a^{n-1} + b^{n-1}) + (a^n + b^n) - (a + b) + 1]$

$= x^n [(\frac{y}{x})^n + \frac{y}{x} - \frac{z}{x} - \frac{y}{x}(a^{n-1} + b^{n-1}) + (a^n + b^n) + 1]$

$= y^n - y[(ax)^{n-1} + (bx)^{n-1}] + [(ax)^n + (bx)^n] + x^{n-1}(y - z + x)$

2   Determinants of the k-Fibonacci RSFPLR and RSLPFL Circulant Matrix

Theorem 1. Let $A = \text{RSFPLRcircfr}(F_{k,1}, F_{k,2}, \ldots, F_{k,n})$ be a k-Fibonacci RSFPLR circulant matrix, then

$$\det A = \frac{(1+F_{k,n+1})^n - (1+F_{k,n+1})(g_5^{n-1}+h_5^{n-1}) + (g_5^n+h_5^n) + (-F_{k,n})^{n-1}}{1+(-1)^n k + L_{k,n} - L_{k,n-1}},$$

where

$$g_5 = \frac{F_{k,n+1} - F_{k,n} + \sqrt{(F_{k,n+1}+F_{k,n-1})^2 + 4F_{k,n}}}{2},$$

$$h_5 = \frac{F_{k,n+1} - F_{k,n} - \sqrt{(F_{k,n+1}+F_{k,n-1})^2 + 4F_{k,n}}}{2}.$$

Proof. The matrix A can be writen as

$$A = \begin{pmatrix} F_{k,1} & F_{k,2} & \cdots & F_{k,n-1} & F_{k,n} \\ -F_{k,n} & F_{k,1}+F_{k,n} & F_{k,2} & \cdots & F_{k,n-1} \\ \vdots & -F_{k,n}+F_{k,n-1} & \ddots & \ddots & \vdots \\ -F_{k,3} & \vdots & \cdots & \ddots & F_{k,2} \\ -F_{k,2} & -F_{k,3}+F_{k,2} & \cdots & -F_{k,n}+F_{k,n-1} & F_{k,1}+F_{k,n} \end{pmatrix}_{n \times n}$$

From the Lemma 1, the determinant of A is

$$\det A = \prod_{i=1}^{n}(F_{k,1} + F_{k,2}\varepsilon_i + \cdots + F_{k,n}\varepsilon_i^{n-1})$$

$$= \prod_{i=1}^{n}(\frac{\alpha - \beta}{\alpha - \beta} + \frac{\alpha^2 - \beta^2}{\alpha - \beta}\varepsilon_i + \cdots + \frac{\alpha^n - \beta^n}{\alpha - \beta}\varepsilon_i^{n-1})$$

$$= \prod_{i=1}^{n}(\frac{(F_{k,1}+F_{k,n+1}) - \varepsilon_i(F_{k,n+1} - F_{k,n}) + \varepsilon_i^2(-F_{k,n})}{1 - k\varepsilon_i - \varepsilon_i^2}).$$

According to Lemma 2, we can get

$$\prod_{i=1}^{n}[(F_{k,1}+F_{k,n+1}) - \varepsilon_i(F_{k,n+1} - F_{k,n}) + \varepsilon_i^2(-F_{k,n})]$$

$$= (1+F_{k,n+1})^n - (1+F_{k,n+1})(g_5^{n-1}+h_5^{n-1}) + (g_5^n+h_5^n) + (-F_{k,n})^{n-1},$$

and

$$\prod_{i=1}^{n}(1 - k\varepsilon_i - \varepsilon_i^2) = 1 + (-1)^n k + L_{k,n} - L_{k,n-1}.$$

Hence, we can get

$$\det A = \frac{(1+F_{k,n+1})^n - (1+F_{k,n+1})(g_5^{n-1}+h_5^{n-1}) + (g_5^n+h_5^n) + (-F_{k,n})^{n-1}}{1+(-1)^n k + L_{k,n} - L_{k,n-1}},$$

where

$$g_5 = \frac{F_{k,n+1} - F_{k,n} + \sqrt{(F_{k,n+1} + F_{k,n-1})^2 + 4F_{k,n}}}{2},$$

$$h_5 = \frac{F_{k,n+1} - F_{k,n} - \sqrt{(F_{k,n+1} + F_{k,n-1})^2 + 4F_{k,n}}}{2}.$$

Corollary 1. Let $A' = \text{RSFPLRcircfr}(F_1, F_2, \ldots, F_n)$ be a Fibonacci RSFPLR circulant matrix, then

$$\det A' = \frac{(1 + F_{n+1})^n - (1 + F_{n+1})(g_1^{n-1} + h_1^{n-1}) + (g_1^n + h_1^n) + (-F_n)^{n-1}}{1 + (-1)^n + L_{n-2}},$$

where

$$g_1 = \frac{F_{n-1} + \sqrt{(F_{n+1} + F_{n-1})^2 + 4F_n}}{2}, \quad h_1 = \frac{F_{n-1} - \sqrt{(F_{n+1} + F_{n-1})^2 + 4F_n}}{2}.$$

Using the method in Theorem 1 similarly, we also have

Theorem 2. Let $B = \text{RSFPLRcircfr}(F_{k,n}, F_{k,n-1}, \ldots, F_{k,1})$. Then

$$\det B = \frac{(F_{k,n})^n - F_{k,n}(g_6^{n-1} + h_6^{n-1}) + (g_6^n + h_6^n) + (-1)^{n-1}(F_{k,n+1} + F_{k,n})}{1 + (-1)^{n-1}k + (-1)^n(L_{k,n} + L_{k,n-1})},$$

where

$$g_6 = \frac{-F_{k,n+1} - 1 + \sqrt{(F_{k,n+1} + 1)^2 + 4F_{k,n}}}{2},$$

$$h_6 = \frac{-F_{k,n+1} - 1 - \sqrt{(F_{k,n+1} + 1)^2 + 4F_{k,n}}}{2}.$$

Proof. The matrix B can be writen as

$$B = \begin{pmatrix} F_{k,n} & F_{k,n-1} & \cdots & F_{k,2} & F_{k,1} \\ -F_{k,1} & F_{k,n} + F_{k,1} & F_{k,n-1} & \cdots & F_{k,2} \\ \vdots & -F_{k,n} + F_{k,n-1} & \ddots & \ddots & \vdots \\ -F_{k,n-2} & \vdots & \cdots & \ddots & F_{k,n-1} \\ -F_{k,n-1} & -F_{k,n-2} + F_{k,n-1} & \cdots & -F_{k,1} + F_{k,2} & F_{k,n} + F_{k,1} \end{pmatrix}_{n \times n}$$

From the Lemma 1, the determinant of B is

$$\det B = \prod_{i=1}^{n}(F_{k,n} + F_{k,n-1}\varepsilon_i + \cdots + F_{k,1}\varepsilon_i^{n-1})$$

$$= \prod_{i=1}^{n}(\frac{\alpha^n - \beta^n}{\alpha - \beta} + \frac{\alpha^{n-1} - \beta^{n-1}}{\alpha - \beta}\varepsilon_i + \cdots + \frac{\alpha - \beta}{\alpha - \beta}\varepsilon_i^{n-1})$$

$$= \prod_{i=1}^{n} \frac{\varepsilon_i^2 - \varepsilon_i(F_{k,n+1}+1) - F_{k,n}}{\varepsilon_i^2 - k\varepsilon_i - 1}.$$

According to Lemma 2, we can get

$$\prod_{i=1}^{n}[\varepsilon_i^2 - \varepsilon_i(F_{k,n+1}+1) - F_{k,n}]$$

$$= (-F_{k,n})^n + F_{k,n}(g_6^{n-1} + h_6^{n-1}) + (g_6^n + h_6^n) - F_{k,n} - F_{k,n+1},$$

and

$$\prod_{i=1}^{n}(\varepsilon_i^2 - k\varepsilon_i - 1) = (-1)^n + L_{k,n} + L_{k,n-1} - k.$$

Hence, we can get

$$\det B = \frac{(-F_{k,n})^n + F_{k,n}(g_6^{n-1} + h_6^{n-1}) + (g_6^n + h_6^n) - F_{k,n} - F_{k,n+1}}{(-1)^n + L_{k,n} + L_{k,n-1} - k},$$

where

$$g_6 = \frac{F_{k,n+1}+1+\sqrt{(F_{k,n+1}+1)^2+4F_{k,n}}}{2}, \quad h_6 = \frac{F_{k,n+1}+1-\sqrt{(F_{k,n+1}+1)^2+4F_{k,n}}}{2}.$$

Corollary 2. Let $B' = \mathrm{RSFPLRcircfr}(F_n, F_{n-1}, \ldots, F_1)$ be a Fibonacci RSFPLR circulant matrix, then

$$\det B' = \frac{(-F_n)^n + F_n(g_2^{n-1} + h_2^{n-1}) + (g_2^n + h_2^n) - F_{n+2}}{(-1)^n + L_{n+1} - 1},$$

where

$$g_2 = \frac{F_{n+1}+1+\sqrt{(F_{n+1}+1)^2+4F_n}}{2}, \quad h_2 = \frac{F_{n+1}+1-\sqrt{(F_{n+1}+1)^2+4F_n}}{2}.$$

Theorem 3. Let $C = \mathrm{RSLPFLcircfr}(F_{k,1}, F_{k,2}, \ldots, F_{k,n})$ be a k-Fibonacci RSLPFL circulant matrix, then

$$\det C = \frac{(-F_{k,n})^n + F_{k,n}(g_6^{n-1} + h_6^{n-1}) + (g_6^n + h_6^n) - F_{k,n} - F_{k,n+1}}{(-1)^n + L_{k,n} + L_{k,n-1} - k}(-1)^{\frac{n(n-1)}{2}},$$

where g_6 and h_6 are given in Theorem 2.

Proof. The matrix C can be writen as

$$C = \begin{pmatrix} F_{k,1} & F_{k,2} & \cdots & F_{k,n-1} & F_{k,n} \\ F_{k,2} & F_{k,3} & \ddots & F_{k,n}+F_{k,1} & -F_{k,1} \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ F_{k,n-1} & F_{k,n}+F_{k,1} & \ddots & \vdots & -F_{k,n-2} \\ F_{k,n}+F_{k,1} & -F_{k,1}+F_{k,2} & \cdots & -F_{k,n-2}+F_{k,n-1} & -F_{k,n-1} \end{pmatrix}$$

$$= \begin{pmatrix} F_{k,n} & F_{k,n-1} & \cdots & F_{k,2} & F_{k,1} \\ -F_{k,1} & F_{k,n}+F_{k,1} & \ddots & F_{k,3} & F_{k,2} \\ \vdots & \ddots & \ddots & F_{k,3} & \vdots \\ -F_{k,n-2} & -F_{k,n-3}+F_{k,n-2} & \ddots & \vdots & F_{k,n-1} \\ -F_{k,n-1} & -F_{k,n-2}+F_{k,n-1} & \cdots & -F_{k,1}+F_{k,2} & F_{k,n}+F_{k,1} \end{pmatrix} \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & \cdots & 0 & 1 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

$$= B\Gamma,$$

then we get $\det C = \det B \det \Gamma$, where $B = \mathrm{RSFPLRcircfr}(F_{k,n}, F_{k,n-1} \ldots, F_{k,1})$

and its determinant could be obtained through Theorem 2, and since $\det \Gamma = (-1)^{\frac{n(n-1)}{2}}$,

so

$$\det C = \det B \det \Gamma$$
$$= \frac{(-F_{k,n})^n + F_{k,n}(g_6^{n-1}+h_6^{n-1})+(g_6^n+h_6^n)-F_{k,n}-F_{k,n+1}}{(-1)^n+L_{k,n}+L_{k,n-1}-k}(-1)^{\frac{n(n-1)}{2}}.$$

Corollary 3. Let $C' = \mathrm{RSLPFLcircfr}(F_1, F_2, \ldots, F_n)$ be a Fibonacci RSLPFL circulant matrix, then

$$\det C' = \frac{(-F_n)^n + F_n(g_2^{n-1}+h_2^{n-1})+(g_2^n+h_2^n)-F_{n+2}}{(-1)^n+L_{n+1}-1}(-1)^{\frac{n(n-1)}{2}},$$

where g_2 and h_2 are given in Corollary 2.

3   Determinants of the k-Lucas RSFPLR and RSLPFL Circulant Matrix

Theorem 4. Let $D = \mathrm{RSFPLRcircfr}(L_{k,1}, L_{k,2}, \ldots, L_{k,n})$ be a k-Lucas RSFPLR circulant matrix, then

$$\det D = \frac{(k+L_{k,n+1})^n - (k+L_{k,n+1})(g_7^{n-1}+h_7^{n-1}) + (g_7^n+h_7^n) + (k+2)(-L_{k,n})^{n-1}}{1+(-1)^n k + L_{k,n} - L_{k,n-1}},$$

where $g_7 = \dfrac{L_{k,n+1} - L_{k,n} - 2 + \sqrt{(L_{k,n+1}-L_{k,n}-2)^2 + 4L_{k,n}(L_{k,n+1}+k)}}{2}$,

$h_7 = \dfrac{L_{k,n+1} - L_{k,n} - 2 - \sqrt{(L_{k,n+1}-L_{k,n}-2)^2 + 4L_{k,n}(L_{k,n+1}+k)}}{2}.$

Proof. The matrix D can be writen as

$$D = \begin{pmatrix} L_{k,1} & L_{k,2} & \cdots & L_{k,n-1} & L_{k,n} \\ -L_{k,n} & L_{k,1}+L_{k,n} & L_{k,2} & \cdots & L_{k,n-1} \\ \vdots & -L_{k,n}+L_{k,n-1} & \ddots & \ddots & \vdots \\ -L_{k,3} & \vdots & \cdots & \ddots & L_{k,2} \\ -L_{k,2} & -L_{k,3}+L_{k,2} & \cdots & -L_{k,n}+L_{k,n-1} & L_{k,1}+L_{k,n} \end{pmatrix}_{n\times n}$$

From the Lemma 1, the determinant of D is

$$\det D = \prod_{i=1}^{n}(L_{k,1} + L_{k,2}\varepsilon_i + \cdots + L_{k,n}\varepsilon_i^{n-1})$$

$$= \prod_{i=1}^{n}((\alpha+\beta) + (\alpha^2+\beta^2)\varepsilon_i + \cdots + (\alpha^n+\beta^n)\varepsilon_i^{n-1})$$

$$= \prod_{i=1}^{n}\left(\frac{(k+L_{k,n+1}) - \varepsilon_i(L_{k,n+1}-L_{k,n}-2) - L_{k,n}\varepsilon_i^2}{1 - k\varepsilon_i - \varepsilon_i^2}\right).$$

According to Lemma 2, we can get

$$\prod_{i=1}^{n}[(k+L_{k,n+1}) - \varepsilon_i(L_{k,n+1}-L_{k,n}-2) - L_{k,n}\varepsilon_i^2]$$

$$= (k+L_{k,n+1})^n - (k+L_{k,n+1})(g_7^{n-1}+h_7^{n-1}) + (g_7^n+h_7^n) + (k+2)(-L_{k,n})^{n-1},$$

and

$$\prod_{i=1}^{n}(1 - k\varepsilon_i - \varepsilon_i^2) = 1 + (-1)^n k + L_{k,n} - L_{k,n-1}.$$

Hence, we can get

$$\det D = \frac{(k+L_{k,n+1})^n - (k+L_{k,n+1})(g_7^{n-1}+h_7^{n-1}) + (g_7^n+h_7^n) + (k+2)(-L_{k,n})^{n-1}}{1+(-1)^n k + L_{k,n} - L_{k,n-1}},$$

where $g_7 = \dfrac{L_{k,n+1} - L_{k,n} - 2 + \sqrt{(L_{k,n+1}-L_{k,n}-2)^2 + 4L_{k,n}(L_{k,n+1}+k)}}{2}$,

$$h_7 = \frac{L_{k,n+1} - L_{k,n} - 2 - \sqrt{(L_{k,n+1} - L_{k,n} - 2)^2 + 4L_{k,n}(L_{k,n+1} + k)}}{2}.$$

Corollary 4. Let $D' = \text{RSFPLRcircfr}(L_1, L_2, \ldots, L_n)$ be a Lucas RSFPLR circulant matrix, then

$$\det D' = \frac{(1 + L_{n+1})^n - (1 + L_{n+1})(g_3^{n-1} + h_3^{n-1}) + (g_3^n + h_3^n) + 3(-L_n)^{n-1}}{1 + (-1)^n + L_{n-2}},$$

where

$$g_3 = \frac{L_{k,n+1} - L_{k,n} - 2 + \sqrt{(2 + L_{n-2})^2 + 4L_n(L_{n+1} + 1)}}{2},$$

$$h_3 = \frac{L_{k,n+1} - L_{k,n} - 2 - \sqrt{(2 + L_{n-2})^2 + 4L_n(L_{n+1} + 1)}}{2}.$$

Using the method in Theorem 4 similary, we also have

Theorem 5. Let $E = \text{RSFPLRcircfr}(L_{k,n}, L_{k,n-1}, \ldots, L_{k,1})$. Then

$$\det E = \frac{(-L_{k,n} - 2)^n + (L_{k,n} + 2)(g_8^{n-1} + h_8^{n-1}) + (g_8^n + h_8^n) - k^{n-1}(L_{k,n} + L_{k,n+1})}{(-1)^n + L_{k,n} + L_{k,n-1} - k},$$

where

$$g_8 = \frac{L_{k,n+1} + k - 2 + \sqrt{(L_{k,n+1} + k - 2)^2 + 4k(L_{k,n} + 2)}}{2},$$

$$h_8 = \frac{L_{k,n+1} + k - 2 - \sqrt{(L_{k,n+1} + k - 2)^2 + 4k(L_{k,n} + 2)}}{2}.$$

Proof. The matrix E can be writen as

$$E = \begin{pmatrix} L_{k,n} & L_{k,n-1} & \cdots & L_{k,2} & L_{k,1} \\ -L_{k,1} & L_{k,n} + L_{k,1} & L_{k,n-1} & \cdots & L_{k,2} \\ \vdots & -L_{k,n} + L_{k,n-1} & \ddots & \ddots & \vdots \\ -L_{k,n-2} & \vdots & \cdots & \ddots & L_{k,n-1} \\ -L_{k,n-1} & -L_{k,n-2} + L_{k,n-1} & \cdots & -L_{k,1} + L_{k,2} & L_{k,n} + L_{k,1} \end{pmatrix}_{n \times n}$$

From the Lemma 1, the determinant of E is

$$\det E = \prod_{i=1}^{n} (L_{k,n} + L_{k,n-1}\varepsilon_i + \cdots + L_{k,1}\varepsilon_i^{n-1})$$

$$= \prod_{i=1}^{n} [(\alpha^n + \beta^n) + (\alpha^{n-1} + \beta^{n-1})\varepsilon_i + \cdots + (\alpha + \beta)\varepsilon_i^{n-1}]$$

$$= \prod_{i=1}^{n} \frac{k\varepsilon_i^2 + \varepsilon_i(2 - k - L_{k,n+1}) - L_{k,n} - 2}{\varepsilon_i^2 - k\varepsilon_i - 1}.$$

According to Lemma 2, we can get

$$\prod_{i=1}^{n} [k\varepsilon_i^2 - \varepsilon_i(L_{k,n+1} + k - 2) - L_{k,n} - 2]$$

$$= (-L_{k,n} - 2)^n + (L_{k,n} + 2)(g_8^{n-1} + h_8^{n-1}) + (g_8^n + h_8^n) - k^{n-1}(L_{k,n} + L_{k,n+1}),$$

and

$$\prod_{i=1}^{n} (\varepsilon_i^2 - k\varepsilon_i - 1) = (-1)^n + L_{k,n} + L_{k,n-1} - k.$$

Hence, we can get

$$\det E = \frac{(-L_{k,n} - 2)^n + (L_{k,n} + 2)(g_8^{n-1} + h_8^{n-1}) + (g_8^n + h_8^n) - k^{n-1}(L_{k,n} + L_{k,n+1})}{(-1)^n + L_{k,n} + L_{k,n-1} - k},$$

where

$$g_8 = \frac{L_{k,n+1} + k - 2 + \sqrt{(L_{k,n+1} + k - 2)^2 + 4k(L_{k,n} + 2)}}{2},$$

$$h_8 = \frac{L_{k,n+1} + k - 2 - \sqrt{(L_{k,n+1} + k - 2)^2 + 4k(L_{k,n} + 2)}}{2}.$$

Corollary 5. Let $E' = \text{RSFPLRcircfr}(L_n, L_{n-1}, \ldots, L_1)$ be a Lucas RSFPLR circulant matrix, then

$$\det E' = \frac{(-2 - L_n)^n + (2 + L_n)(g_4^{n-1} + h_4^{n-1}) + (g_4^n + h_4^n) - L_{n+2}}{(-1)^n + L_{n+1} - 1},$$

where

$$g_4 = \frac{L_{n+1} - 1 + \sqrt{(L_{n+1} - 1)^2 + 4L_n + 8}}{2}, \quad h_4 = \frac{L_{n+1} - 1 - \sqrt{(L_{n+1} - 1)^2 + 4L_n + 8}}{2}.$$

Theorem 6. Let $F = \text{RSLPFLcircfr}(L_{k,1}, L_{k,2}, \ldots, L_{k,n})$ be a k-Lucas RSLPFL circulant matrix, then

$$\det F = \frac{(-L_{k,n} - 2)^n + (L_{k,n} + 2)(g_8^{n-1} + h_8^{n-1}) + (g_8^n + h_8^n) - k^{n-1}(L_{k,n} + L_{k,n+1})}{(-1)^n + L_{k,n} + L_{k,n-1} - k}(-1)^{\frac{n(n-1)}{2}},$$

where g_8 and h_8 are given in Theorem 5.

Proof. The matrix F can be writen as

$$F = \begin{pmatrix} L_{k,1} & L_{k,2} & \cdots & L_{k,n-1} & L_{k,n} \\ L_{k,2} & L_{k,3} & \iddots & L_{k,n}+L_{k,1} & -L_{k,1} \\ \vdots & \iddots & \iddots & \vdots & \vdots \\ L_{k,n-1} & L_{k,n}+L_{k,1} & \iddots & \vdots & -L_{k,n-2} \\ L_{k,n}+L_{k,1} & -L_{k,1}+L_{k,2} & \cdots & -L_{k,n-2}+L_{k,n-1} & -L_{k,n-1} \end{pmatrix}$$

$$= \begin{pmatrix} L_{k,n} & L_{k,n-1} & \cdots & L_{k,2} & L_{k,1} \\ -L_{k,1} & L_{k,n}+L_{k,1} & \iddots & L_{k,3} & L_{k,2} \\ \vdots & \iddots & \iddots & \iddots & \vdots \\ -L_{k,n-2} & -L_{k,n-3}+L_{k,n-2} & \iddots & \vdots & L_{k,n-1} \\ -L_{k,n-1} & -L_{k,n-2}+L_{k,n-1} & \cdots & -L_{k,1}+L_{k,2} & L_{k,n}+L_{k,1} \end{pmatrix} \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & \cdots & 0 & 1 & 0 \\ \vdots & \iddots & \iddots & \iddots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

$$= E\Gamma .$$

Thus we have $\det F = \det E \det \Gamma$, which matrix $E = \mathrm{RSFPLRcirfr}(L_{k,n}, L_{k,n-1}, \ldots, L_{k,1})$

and its determinant can be obtained from Theorem 5. In addition $\det \Gamma = (-1)^{\frac{n(n-1)}{2}}$. So the determinant of matrix F is

$$\det F = \frac{(-L_{k,n}-2)^n + (L_{k,n}+2)(g_8^{n-1}+h_8^{n-1}) + (g_8^n+h_8^n) - k^{n-1}(L_{k,n}+L_{k,n+1})}{(-1)^n + L_{k,n}+L_{k,n-1}-k}(-1)^{\frac{n(n-1)}{2}},$$

where g_8 and h_8 are given in Theorem 5.

Corollary 6. Let $F' = \mathrm{RSLPFLcircfr}(L_1, L_2, \ldots, L_n)$ be a Lucas RSLPFL circulant matrix, then

$$\det F' = \frac{(-2-L_n)^n + (2+L_n)(g_4^{n-1}+h_4^{n-1}) + (g_4^n+h_4^n) - L_{n+2}}{(-1)^n + L_{n+1}-1}(-1)^{\frac{n(n-1)}{2}},$$

where g_4 and h_4 are given in Corollary 5.

References

1. Yazlik, Y., Taskara, N.: A Note on Generalized k-Horadam Sequence. Comput. Math. Appl. 63, 36–41 (2012)
2. Jaiswal, D.: On Determinants Involving Generalized Fibonacci Numbers. Fibonacci Quart. 7, 319–330 (1969)
3. Lin, D.: Fibonacci-Lucas Quasi-cyclic Matrices. Fibonacci Quart. 40, 280–286 (2002)
4. Lind, D.: A Fibonacci Circulant. Fibonacci Quart. 8, 449–455 (1970)

5. Shen, S.Q., Cen, J.M., Hao, Y.: On the Determinants and Inverses of Circulant Matrices with Fibonacci and Lucas Numbers. Appl. Math. Comput. 217, 9790–9797 (2011)
6. Akbulak, M., Bozkurt, D.: On the Norms of Toeplitz Matrices Involving Fibonacci and Lucas Numbers. Hacet. J. Math. Stat. 37(2), 89–95 (2008)
7. Lee, G.Y., Kim, J.S., Lee, S.G.: Factorizations and Eigenvalues of Fibonacci and Symmetric Fibonacci Matrices. Fibonacci Quart. 40, 203–211 (2002)
8. Miladinovi, M., Stanimirović, P.: Singular Case of Generalized Fibonacci and Lucas Matrices. J. Korean Math. Soc. 48, 33–48 (2011)
9. Stanimirović, P., Nikolov, J., Stanimirović, I.: A Generalization of Fibonacci and Lucas Matrices. Discret. Appl. Math. 156, 2606–2619 (2008)
10. Zhang, Z., Zhang, Y.: The Lucas Matrix and Some Combinatorial Identities. Indian J. Pure Appl. Math. 38, 457–465 (2007)
11. Jiang, Z.L., Xu, Z.B.: Eifficient Algorithm for Finding the Inverse and Group Inverse of FLS r-circulant Matrix. J. Appl. Math. Comput. 18, 45–57 (2005)
12. David, C.: Regular Representations of Semisimple Algebras, Separable Field Extensions, Group Characters, Generalized Circulants, and Generalized Cyclic Codes. Linear Algebra Appl. 218, 147–183 (1995)
13. Jiang, Z.L., Zhou, Z.X.: Circulant Matrices. Chengdu Technology University Publishing Company, Chengdu (1999)

Strong Convergence for Finding Fixed Point
of Multi-valued Mapping

Huancheng Zhang[1], Xinghua Ma[2], Linan Shi[1], Yan Yan[2], and Jingguo Qu[1]

[1] Qinggong College, Hebei United University, Tangshan, 063000, China
`zhanghuancheng@yahoo.cn`
[2] Hebei United University, Tangshan, 063000, China

Abstract. It is known that the Mann iteration is for approximating fixed points of nonexpans-ive single-valued mappings. However, in general the Mann iteration process has only weak convergence. In recent years, Sastry, Babu and Panyanak introduced the Mann and Ishikawa iteration scheme for non-expansive multi-valued mapping and obtained the strong convergen-ce theorems. In this paper, we introduce a new iterative method for quasi-nonexpansive multi-valued maps in Banach spaces, and obtain the strong convergence theorems.

Keywords: quasi-nonexpansive multi-valued, Mann iteration, nonexpansive multi-valued.

1   Introduction

The Banach fixed-point theorem (also known as the Banach contraction mapping theorem or the Banach contraction mapping principle) is an important tool in the theory of metric spaces; it guarantees the existence and uniqueness of fixed points of certain self maps of metric spaces, and provides a constructive method to find those fixed points [1-5].

In 1969, Nadler [6] extended the Banach fixed-point theorem from the single-valued maps to the set-valued contractive maps. Before presenting this important theorem, we start with introducing some notations.

Let X be a Banach space and D a nonempty subset of X. The set D is said to be proximal if for each $x \in X$, there exists $y \in D$ such that $\|x - y\| = d(x, D)$, where $d(x, D) = \inf\{\|x - z\| : z \in D\}$. Let CB(D), P(D) denote the family of nonempty closed bounded subsets and nonempty proximal bounded subsets of D respectively [7-11].

The Hausdorff metric on CB(D) is defined by

$$H(A, B) = \max\left\{\sup_{x \in A} d(x, B), \sup_{y \in B} d(y, A)\right\}$$

for $A, B \in CB(D)$.

Let (X,d) be a complete metric space, an element $x \in X$ is said to be a coincidence point of $T : X - CB(X)$ and $f : X - X$ if $fx \in Tx$. We denote

$C(f,T) = \{x \in X : fx \in Tx\}$,

The set of coincidence points of T and f.

Maps $f : X - X$ and $T : X - CB(X)$ are weakly compatible if they commute at their coincidence points, that is, if $fTx = Tfx$ whenever $fx \in Tx$ [12-14].

Let $T : X - CB(X)$ be a multi-valued map and $f : X - X$ be a single-valued map. The map f is said to be T -weakly commuting at $x \in X$ if $ffx \in Tfx$.

An element $x \in X$ is a common fixed point of $T, S : X - CB(X)$ and $f : X - X$ if $x = fx \in Tx \cap Sx$.

Example. Consider $X = [0, +c)$ equipped with the metric $d(x, y) = |x - y|$ for every $x, y \in X$. Define $f : X - X$ and $T : X - CB(X)$ as

$$fx = \begin{cases} 0, x \in [0,1) \\ 2x, x \in [1,+\infty) \end{cases}, \quad Tx = \begin{cases} \{x\}, x \in [0,1) \\ [1,1+2x], x \in [1,+\infty) \end{cases}.$$

We have

(i) $f1 = 2 \in [1,3] = T1$, that is, $x = 1$ is a coincidence point of f and T ;

(ii) $fT1 = [2,6] \neq [1,5] = Tf1$, that is, f and T are not weakly compatible mappings;

(iii) $ff1 = 4 \in [1,5] = Tf1$, that is, f is T-weakly commuting at 1.

A single-valued mapping $T : D - D$ is called nonexpansive if $\|Tx - Ty\| \leq \|x - y\|$ for $x, y \in D$. A single-valued mapping $T : D - D$ is called contraction mapping [14-18] if $\|Tx - Ty\| \leq \alpha \|x - y\|$, $\alpha \in [0,1)$ for all $x, y \in D$.

A multi-valued mapping $T : D - CB(D)$ is called nonexpansive if $H(Tx, Ty) \leq \|x - y\|$ for $x, y \in D$. An element $p \in D$ is called a fixed point of $T : D - D$ (respectively $T : D - CB(D)$) if $p = Tp$ (respectively $p \in Tp$). Let F(T) denote the set of fixed points of T.

The mapping $T : D - CB(D)$ is called quasi-nonexpansive [1,15-18] if $F(T) \neq \phi$ and for all $x \in D$, $p \in F(T)$, $H(Tx, Tp) \leq \|x - p\|$.

It is clear that every nonexpansive multi-valued mapping T with $F(T) \neq \phi$ is quasi-nonexpansive. But quasi-nonexpansive mappings may be not nonexpan-sive.

Example. Let $K = R$ with the usual metric and $T : K - K$ be defined by

$$T(x) = \begin{cases} 0, x \leq 1 \\ [x - \dfrac{3}{4}, x - \dfrac{1}{3}], x > 1 \end{cases}$$

Then, clearly $F(T) = \{0\}$ and for any x we have

$H(Tx, T0) \leq \|x - 0\| = |x - 0|$,

hence T is quasi-nonexpansive. However, if $x=2$, $y=1$ we get that $H(Tx,Ty) > |x - y| = 1$ and hence not nonexpansive.

The mapping $T : D - CB(D)$ is called hemicompa-ct if for any sequence $\{x_n\}$ in D such that $d(x_n,Tx_n) - 0$ as $n - \infty$, there exists a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ such that $x_{n_k} - p \in D$. We note that if D is compact, then every multi-valued mapping $T : D - CB(D)$ is hemicompact. $T : D - CB(D)$ is called satisfy Condition (I) if there is a nondecreasing function $f : [0,\infty) - [0,\infty)$ with $f(0) = 0$, $f(r) > 0$ for $r \in (0,\infty)$ such that

$d(x,Tx) \ge f(d(x,F(T)))$, $\forall x \in D$.

Two multi-valued maps $S,T : D - CB(D)$ are said to satisfy Condition (II) if there is a nondecreasing function $f : [0,\infty) - [0,\infty)$ with $f(0) = 0$, $f(r) > 0$ for $r \in (0,\infty)$ such that either

$d(x, Sx) \ge f(d(x, F(S) \bigcap F(T)))$ or $d(x,Tx) \ge f(d(x, F(S) \bigcap F(T)))$

for all $x \in D$.

Let $T : D - D$ be a single-valued mapping. The Mann iteration scheme (see [2]) is the sequence $\{x_n\}$ defined by

$x_{n+1} = a_n Tx_n + (1 - a_n)x_n$, $\forall n \ge 0$

where $a_n \in [0,1]$ satisfies certain conditions. The Ishikawa iteration scheme (see[3]) is the sequence $\{x_n\}$ defined by

$$\begin{cases} y_n = \beta_n Tx_n + (1-\beta_n)x_n \\ x_{n+1} = \alpha_n Ty_n + (1-\alpha_n)x_n \end{cases}, \quad \forall n \ge 0$$

where $a_n, \beta_n \in [0,1]$ satisfy certain conditions. Iterative techniques for approximating fixed points of nonexpansive single-valued mappings have been investigated by various authors (see, e.g. [4-7]) using the Mann iteration scheme or the Ishikawa iteration scheme. However, we note that Mann's iteration process (1.1) has only weak convergence, in general.

In 1969, Nadler [19] proved the fixed point theorem of multi-valued mapping.

Theorem N ([19]). Let (X,d) be a complete metric space, and $T : X - CB(X)$ be a multi-valued map satisfying $H(Tx,Ty) \colon qd(x, y)$ for all $x, y \in X$, where q is a constant such that $q \in [0,1)$. Then, T has a fixed point.

In 2010, an extension of Theorem N ([19]) was obtained by Gordji et al [20]. They proved the following result.

Theorem G ([20]) . Let (X,d) be a complete metric space, and $T : X - CB(X)$ be a multi-valued map satisfying

$H(Tx,Ty) \colon ad(x, y) + \beta[D(x,Tx) + D(y,Ty)] + \gamma[D(x,Ty) + D(y,Tx)]$

for all $x, y \in X$, where $a, \beta, \gamma \ge 0$ and $a + 2\beta + 2\gamma < 1$. Then, T has a fixed point.

In 2012, B.Damjanovi' et al [13] establish some results on coincidence and common fixed points for a two-pair of multi-valued and single-valued maps in complete metric spaces. Presented theorems generalize the result given by Gordji and other existing results.

Theorem B1 ([13]). Let (X,d) be a complete metric space. Let $T,S : X - CB(X)$ be a pair of multi-value maps and $f, g : X - X$ a pair of single-valued maps. Suppose that

$H(Sx,Ty) \le \alpha d(fx,gy) + \beta[D(fx,Sx) + D(gy,Ty)] + \gamma[D(x,Ty) + D(y,Tx)]$

for each $x, y \in X$, where $\alpha, \beta, \gamma \ge 0$ and $0 < \alpha + 2\beta + 2\gamma < 1$. Suppose also that

(i) $SX \subseteq gX$, $TX \subseteq fX$, (ii) $f(X)$ and $g(X)$ are closed.

Then, there exist points u and w in X, such that $fu \in Su$, $gw \in Tw$, $fu = gw$ and $Su = Tw$.

Theorem B2 ([13]). Let (X,d) be a complete metric space. Let $T,S : X - CB(X)$ be multi-value maps and $f : X - X$ be a single-valued map satisfying, for each $x, y \in X$,

$H(Sx,Ty) \le \alpha d(fx,fy) + \beta[D(fx,Sx) + D(fy,Ty)] + \gamma[D(fx,Ty) + D(fy,Sx)]$,

where $\alpha, \beta, \gamma \ge 0$ and $0 < \alpha + 2\beta + 2\gamma < 1$. If fX is a closed subset of X and $TX \cup SX \subseteq fX$, then, f, T and S have a coincidence in X. Moreever, if f is both T-weakly commuting and S-weakly commuting at each $z \in C(f,T)$, and $ffz = fz$, then, f, T and S have a common fixed point in x.

In 2005, Sastry and Babu [8] introduced the Mann iteration scheme for nonexpansive multi-valued mapping. Ler $T : D - P(D)$ be a multi-valued mapping and fix $p \in F(T)$. The sequence of Mann iteration is defined by $x_0 \in D$,

$$x_{n+1} = \alpha_n y_n + (1 - \alpha_n)x_n, \quad \alpha_n \in [0,1], \quad n \ge 0 \tag{1}$$

Where $y_n \in Tx_n$ such that $\|y_n - p\| = d(p,Tx_n)$. And they proved that $\{x_n\}$ converges strongly to a fixed point q of T under certain conditions. They also claimed that the fixed point q may be different from p.

In 2007, Panyanak [9] modified the iteration scheme of Sastry and Babu [8]. Let $T : D - P(D)$ be a multi-valued mapping and F(T) be a nonempty proximinal subset of D. The sequence of Mann iteration is defined by $x_0 \in D$,

$$x_{n+1} = \alpha_n y_n + (1 - \alpha_n)x_n, \quad \alpha_n \in [a,b], \quad n \ge 0 \tag{2}$$

Where $a, b \in (0,1)$ and $y_n \in Tx_n$ such that $\|y_n - u_n\| = d(u_n,Tx_n)$ and $u_n \in F(T)$ such that $\|x_n - u_n\| = d(x_n, F(T))$.

Theorem P ([9]). Let E be a uniformly convex Banach space, D a nonempty closed convex bounded subset of E, and $T : D - P(D)$ a nonexpansive multi-valued map

that satisfies condition (A). Assume that $(i) 0 \leq a_n < 1$ and $(ii) \sum_{n=1}^{\infty} \alpha_n = \infty$. Suppose that F(T) is a nonempty proximinal subset of D. Then the Mann iterates $\{x_n\}$ defined by (2) converges to a fixed point of T.

In 2009, Shahzad and Zegeye extended and improved the results of Panyanak, Sastry and Babu and Song and Wang to a quasi-nonexpansive multi-valued map. They also relaxed compactness of the domain of T and constructed an iteration scheme which removes the restriction of T namely $Tp = \{p\}$ for any $p \in F(T)$. They introduced new iterations as follows.

Let D be a nonempty convex subset of a Banach space E and $T : D - CB(D)$ and let $a_n, a'_n \in [0,1]$,

(A) The sequence of Ishikawa iterates is defined by $x_0 \in D$,

$y_n = a'_n z'_n + (1 - a'_n) x_n$

$x_{n+1} = a_n z_n + (1 - a_n) x_n$, $\quad n \geq 0$,

where $z'_n \in Tx_n$ and $z_n \in Ty_n$;

(B) Let $T : D - P(D)$ and $p_T x = \{y \in Tx : \|x - y\| = d(x, Tx)\}$. The sequence of Ishikawa iterates is defined by $x_0 \in D$,

$y_n = a'_n z'_n + (1 - a'_n) x_n$

$x_{n+1} = a_n z_n + (1 - a_n) x_n$, $\quad n \geq 0$,

where $z'_n \in P_T x_n$ and $z_n \in P_T y_n$.

As all known, contraction mappings are nonexpansive mappings. In recent years, many authors proved some fixed point theorems about multi-valued contraction mappings, see, e.g. [14-18].

In 2011, Cholamjiak and Suantai [11] proposed new two-step iterative schemes for finding a common fixed point of two quasi-nonexpansive multi-valued maps in Banach spaces and prove strong convergence theorems of the proposed iterations.

Theorem C1 ([11]). Let E be a uniformly convex Banach space, D a nonempty, closed and convex subset of E. Let T_1 be a quasi-nonexpansive multi-valued map and T_2 a quasi-nonexpansive and L-Lipschitzian multi-valued map from D into CB(D) with $F(T_1) \bigcap F(T_1) \neq \emptyset$ and $T_1 p = \{p\} = T_2 p$ for all $p \in F(T_1) \bigcap F(T_1)$. Assume that (i) T_1, T_2 satisfy Condition: there is a nondecreasing function $f : [0, \infty) - [0, \infty)$ with $f(0) = 0$, $f(r) > 0$ for $r \in (0, \infty)$ such that either

$d(x, T_1 x) \geq f(d(x, F(T_1) \bigcap F(T_2)))$ or $d(x, T_2 x) \geq f(d(x, F(T_1) \bigcap F(T_2)))$ for all $\forall x \in D$;

(ii) $\sum_{n=1}^{\infty} (1 - \alpha_n - \beta_n) < \infty$ and $\sum_{n=1}^{\infty} (1 - \alpha'_n - \beta'_n) < \infty$; (iii) $0 < \ell \leq \alpha_n$, $\alpha'_n \leq k < 1$.

The sequence $\{x_n\}$ generated by

$$y_n = \alpha'_n z'_n + \beta'_n x_n + (1 - \alpha'_n - \beta'_n) u_n$$
$$x_{n+1} = a_n z_n + \beta_n y_n + (1 - a_n - \beta_n) v_n$$

where $z'_n \in T_1 x_n$, $z_n \in T_2 y_n$, $a_n, \beta_n, \alpha'_n, \beta'_n \in [0,1]$ and $\{u_n\}, \{v_n\}$ are bounded sequences in D. Then $\{x_n\}$ converges strongly to an element of $F(T_1) \cap F(T_1)$.

Theorem C2 ([11]). Let E be a uniformly convex Banach space, D a nonempty, closed and convex subset of E. Let T_1 and T_2 be two multi-valued maps from D into P(D) with $F(T_1) \cap F(T_1) \neq \phi$ such that P_{T_1} and P_{T_2} are nonexpansive. Assume that

(i) T_1, T_2 satisfy condition (II)

(ii) $\sum_{n=1}^{\infty} (1 - \alpha_n - \beta_n) < \infty$ and $\sum_{n=1}^{\infty} (1 - \alpha'_n - \beta'_n) < \infty$; (iii) $0 < \ell \leq a_n$,

$\alpha'_n \leq k < 1$.

The sequence $\{x_n\}$ generated by

$$y_n = \alpha'_n z'_n + \beta'_n x_n + (1 - \alpha'_n - \beta'_n) u_n$$
$$x_{n+1} = a_n z_n + \beta_n y_n + (1 - a_n - \beta_n) v_n$$

converges strongly to an element of $F(T_1) \cap F(T_1)$.

In this paper, motivated by Sastry and Babu, Panyanak , Shahzad and Zegeye, we introduce a new iterative method (3) and obtain the strong convergence theorem. And the results obtained in this paper improve and extend the relative author's results.

2   Main Results

We use the following iteration scheme.

Let D be a nonempty convex subset of a Banach space X and $a_n, \beta_n \in [0,1]$. The sequence of Ishikawa iteration is defined by $x_0 \in D$,

$$\begin{cases} y_n = \beta_n x_n + (1 - \beta_n) z_n \\ x_{n+1} = \alpha_n f(x_n) + (1 - \alpha_n) y_n \end{cases}, \quad \forall n \geq 0 \tag{3}$$

where $z_n \in T(x_n)$ and f is a contraction on D.

We shall make use of the following result of Suziiki [10].

Lemma 2.1 ([10]). Let $\{x_n\}$ and $\{y_n\}$ be bounded sequences in a Banach space X and $\{\beta_n\}$ be a sequence in [0,1] with $0 < \liminf_{n \to \infty} \beta_n \leq \limsup_{n \to \infty} \beta_n < 1$. Suppose $x_{n+1} = \beta_n x_n + (1 - \beta_n) y_n$ for all $n \geq 0$ and $\limsup_{n \to \infty} (\|y_{n+1} - y_n\| - \|x_{n+1} - x_n\|) \leq 0$. Then $\lim_{n \to \infty} \|y_n - x_n\| = 0$.

Theorem 2.2. Let X be a uniformly convex Banach space, D a nonempty closed convex subset of X, and $T : D - CB(D)$ a quasi-nonexpansive multi-valued

mapping with $F(T) \neq \phi$ and $Tp = \{p\}$ for each $p \in F(T)$. Let $\{x_n\}$ defined by (3). Assume that T satisfies condition (I) and $\alpha_n, \beta_n \in (0,1)$ satisfy

(i) $\lim_{n\to\infty} \alpha_n = 0$, (ii) $0 < \liminf_{n\to\infty} \beta_n \colon \limsup_{n\to\infty} \beta_n < 1$.

Then $\{x_n\}$ converges strongly to a fixed point of T.

Proof. Let $p \in F(T)$. Then, using (3), we have

$$\|y_n - p\| = \|\beta_n x_n + (1-\beta_n)z_n - p\| \qquad\qquad \leq \beta_n\|x_n - p\| + (1-\beta_n)\|z_n - p\|$$
$$= \beta_n\|x_n - p\| + (1-\beta_n)d(z_n, Tp)$$
$$\leq \beta_n\|x_n - p\| + (1-\beta_n)H(Tx_n, Tp) \quad \leq \beta_n\|x_n - p\| + (1-\beta_n)\|x_n - p\| = \|x_n - p\|$$

and

$$\|x_{n+1} - p\| = \|\alpha_n f(x_n) + (1-\alpha_n)y_n - p\| \leq \alpha_n\|f(x_n) - p\| + (1-\alpha_n)\|y_n - p\|$$

$$\leq \alpha_n\|f(x_n) - f(p)\| + \alpha_n\|f(p) - p\| + (1-\alpha_n)\|x_n - p\|$$
$$\leq \alpha\alpha_n\|x_n - p\| + \alpha_n\|f(p) - p\| + (1-\alpha_n)\|x_n - p\|$$

$$= [1 - (1-\alpha)\alpha_n]\|x_n - p\| + \alpha_n\|f(p) - p\|. \tag{4}$$

By simple inductions, we have

$$\|x_n - p\| \leq \max\{\|x_0 - p\|, \frac{\|f(p) - p\|}{1 - \alpha}\}, \quad \forall n \geq 1$$

which gives that the sequence $\{x_n\}$ is bounded and so is $\{y_n\}$.

Next, we claim that

$$\lim_{n\to\infty} \|x_{n+1} - x_n\| = 0. \tag{5}$$

Put $l_n = \dfrac{x_{n+1} - \beta_n x_n}{1 - \beta_n}$. Then we have

$$x_{n+1} = \beta_n x_n + (1 - \beta_n)l_n, \quad \forall n \geq 0. \tag{6}$$

Observing that

$$l_{n+1} - l_n = \frac{\alpha_{n+1} f(x_{n+1}) + (1-\alpha_{n+1})y_{n+1} - \beta_{n+1}x_{n+1}}{1 - \beta_{n+1}} - \frac{\alpha_n f(x_n) + (1-\alpha_n)y_n - \beta_n x_n}{1 - \beta_n}$$

$$= \frac{\alpha_{n+1}(f(x_{n+1}) - y_{n+1})}{1 - \beta_{n+1}} - \frac{\alpha_n(f(x_n) - y_n)}{1 - \beta_n} + z_{n+1} - z_n$$

We have $\|l_{n+1} - l_n\| \leq \dfrac{\alpha_{n+1}}{1 - \beta_{n+1}}\|f(x_{n+1}) - y_{n+1}\| + \dfrac{\alpha_n}{1 - \beta_n}\|f(x_n) - y_n\| + \|z_{n+1} - z_n\|$.

And we also have $\|z_{n+1} - z_n\| \leq H(Tx_{n+1}, Tx_n) \leq \|x_{n+1} - x_n\|$. So

$$\left\| l_{n+1} - l_n \right\| - \left\| x_{n+1} - x_n \right\| \leq \frac{\alpha_{n+1}}{1-\beta_{n+1}} \left\| f(x_{n+1}) - y_{n+1} \right\| + \frac{\alpha_n}{1-\beta_n} \left\| f(x_n) - y_n \right\|.$$

Observing conditions (i) and (ii), we get $\limsup_{n\to\infty}(\left\| l_{n+1} - l_n \right\| - \left\| x_{n+1} - x_n \right\|) \leq 0$.

From Lemma 2.1, we have $\lim_{n\to\infty} \left\| l_n - x_n \right\| = 0$. Observing (6) yields that

$x_{n+1} - x_n = (1-\beta_n)(l_n - x_n)$. Therefore, (4) holds. Observing that

$x_{n+1} - y_n = \alpha_n(f(x_n) - y_n)$ and condition (i), we can easily get $\lim_{n\to\infty} \left\| y_n - x_{n+1} \right\| = 0$.

On the other hand, we have $\left\| y_n - x_n \right\| \leq \left\| x_n - x_{n+1} \right\| + \left\| x_{n+1} - y_n \right\|$. So

$\lim_{n\to\infty} \left\| y_n - x_n \right\| = 0$. Observing that $y_n - x_n = (1-\beta_n)(z_n - x_n)$ and condition (ii),

we get $\lim_{n\to\infty} \left\| z_n - x_n \right\| = 0$. Also $d(x_n, Tx_n) \leq \left\| x_n - z_n \right\| \to 0$ as $n - \infty$. Since T

satisfies condition (I), we have $\lim_{n\to\infty} d(x_n, F(T)) = 0$. Thus there is a subsequence

$\{x_{n_k}\}$ of $\{x_n\}$ such that $\left\| x_{n_k} - p_k \right\| \leq \frac{1}{2^k}$ for $\{p_k\} \subset F(T)$ and all k. From (4), we

have

$$\left\| x_{n+1} - p \right\| \leq \left\| x_n - p \right\| + \alpha_n \left\| f(p) - p \right\|. \tag{7}$$

Since $\lim_{n\to\infty} \alpha_n = 0$, there exists $K \in N$ such that $\alpha_{n_k} \leq \frac{1}{2^k \left\| f(p) - p \right\|}$ for all

$k \geq K$. From (7), we have $\left\| x_{n_{k+1}} - p_k \right\| \leq \left\| x_{n_k} - p_k \right\| + \alpha_{n_k} \left\| f(p) - p \right\| < \frac{1}{2^{k-1}}$, for all

$k \geq K$. Hence $\left\| x_{n_{k+m}} - p_k \right\| < \frac{1}{2^{k-1}}$ for

all $k, m \geq K$. We now show that $\{p_k\}$ is a Cauchy sequence. For all $k, m \geq K$,

we have

$$\left\| p_{k+m} - p_k \right\| \leq \left\| p_{k+m} - x_{n_{k+m}} \right\| + \left\| x_{n_{k+m}} - p_k \right\| < \frac{1}{2^{k-1}} + \frac{1}{2^{k+m}} < \frac{1}{2^{k-1}}.$$

This shows that $\{p_k\}$ is a Cauchy sequence and thus converges strongly to

$q \in D$. Since $d(p_k, Tq) \leq H(Tp_k, Tq) \leq \left\| q - p_k \right\|$ and $p_k - q$ as $k \to \infty$. So

$d(q, Tq) = 0$. Thus $q \in F(T)$ and $\{x_{n_k}\}$ converges strongly to q. If $\{x_{m_k}\}$ is

another subsequence of $\{x_n\}$ which converges strongly to q^*, we have $q^* \in F(T)$.

Since $Tp = \{p\}$, we have $q^* = q$. So $\{x_n\}$ converges strongly to q. This completes

the proof of Theorem 2.2.

Corollary 2.3. Let X be a uniformly convex Banach space, D a nonempty closed

convex subset of X, and $T : D - CB(D)$ a quasi-nonexpansive multi-valued

mapping with $F(T) \neq \phi$ and $Tp = \{p\}$ for each $p \in F(T)$. Let $\{x_n\}$ defined by (3).

Assume that T is hemicompact and continuous and $\alpha_n, \beta_n \in (0,1)$ satisfy

(i) $\lim_{n\to\infty} \alpha_n = 0$, (ii) $0 < \liminf_{n\to\infty} \beta_n \le \limsup_{n\to\infty} \beta_n < 1$.

Then $\{x_n\}$ converges strongly to a fixed point of T.

Proof. It is clear that every nonexpansive multi-valued mapping T with $F(T) \ne \phi$ is quasi-nonexpansive. So, let $q \in F(T)$. Then, using (3), we have

$$\|x_{n+1} - q\| = \|\alpha_n f(x_n) + (1-\alpha_n)y_n - q\| \le \alpha_n\|f(x_n) - q\| + (1-\alpha_n)\|y_n - q\|$$

$$\le \alpha_n\|f(x_n) - f(q)\| + \alpha_n\|f(q) - q\| + (1-\alpha_n)\|x_n - q\| \tag{8}$$

$$\le \alpha\alpha_n\|x_n - q\| + \alpha_n\|f(q) - q\| + (1-\alpha_n)\|x_n - q\| = [1-(1-\alpha)\alpha_n]\|x_n - q\| + \alpha_n\|f(q) - q\|$$

and

$$\|y_n - q\| = \|\beta_n x_n + (1-\beta_n)z_n - q\| \le \beta_n\|x_n - q\| + (1-\beta_n)\|z_n - q\| = \beta_n\|x_n - q\| + (1-\beta_n)d(z_n, Tq)$$
$$\le \beta_n\|x_n - p\| + (1-\beta_n)H(Tx_n, Tp) \le \beta_n\|x_n - q\| + (1-\beta_n)\|x_n - q\| = \|x_n - q\|.$$

By simple inductions, we have

$$\|x_n - q\| \le \max\{\|x_0 - q\|, \frac{\|f(q) - q\|}{1-\alpha}\}, \quad \forall n \ge 1,$$

Which gives that the sequence $\{x_n\}$ is bounded and so is $\{y_n\}$.

Next, we claim that

$$\lim_{n\to\infty}\|x_{n+1} - x_n\| = 0. \tag{9}$$

Put $l_n = \dfrac{x_{n+1} - \beta_n x_n}{1 - \beta_n}$. Then we have

$$x_{n+1} = \beta_n x_n + (1-\beta_n)l_n, \quad \forall n \ge 0. \tag{10}$$

Observing that

$$l_{n+1} - l_n = \frac{\alpha_{n+1}f(x_{n+1}) + (1-\alpha_{n+1})y_{n+1} - \beta_{n+1}x_{n+1}}{1 - \beta_{n+1}} - \frac{\alpha_n f(x_n) + (1-\alpha_n)y_n - \beta_n x_n}{1 - \beta_n}$$

$$= \frac{\alpha_{n+1}(f(x_{n+1}) - y_{n+1})}{1 - \beta_{n+1}} - \frac{\alpha_n(f(x_n) - y_n)}{1 - \beta_n} + z_{n+1} - z_n$$

We have

$$\|l_{n+1} - l_n\| \le \frac{\alpha_{n+1}}{1 - \beta_{n+1}}\|f(x_{n+1}) - y_{n+1}\| + \frac{\alpha_n}{1 - \beta_n}\|f(x_n) - y_n\| + \|z_{n+1} - z_n\|.$$

And we also have

$$\|z_{n+1} - z_n\| \le H(Tx_{n+1}, Tx_n) \le \|x_{n+1} - x_n\|.$$

So

$$\|l_{n+1} - l_n\| \le \frac{\alpha_{n+1}}{1 - \beta_{n+1}}\|f(x_{n+1}) - y_{n+1}\| + \frac{\alpha_n}{1 - \beta_n}\|f(x_n) - y_n\| + \|x_{n+1} - x_n\|.$$

This implies that

$$\left\| l_{n+1} - l_n \right\| - \left\| x_{n+1} - x_n \right\| \le \frac{\alpha_{n+1}}{1 - \beta_{n+1}} \left\| f(x_{n+1}) - y_{n+1} \right\| + \frac{\alpha_n}{1 - \beta_n} \left\| f(x_n) - y_n \right\|.$$

Observing conditions (i) and (ii), we get

$\limsup_{n \to \infty} (\left\| l_{n+1} - l_n \right\| - \left\| x_{n+1} - x_n \right\|) \le 0$.

From Lemma 2.1, we have $\lim_{n \to \infty} \left\| l_n - x_n \right\| = 0$. Observing (10) yields that

$x_{n+1} - x_n = (1 - \beta_n)(l_n - x_n)$.

Therefore, (8) holds. Observing that

$x_{n+1} - y_n = \alpha_n (f(x_n) - y_n)$

and condition (i), we can easily get $\lim_{n \to \infty} \left\| y_n - x_{n+1} \right\| = 0$.

On the other hand, we have

$\left\| y_n - x_n \right\| \le \left\| x_n - x_{n+1} \right\| + \left\| x_{n+1} - y_n \right\|$.

So $\lim_{n \to \infty} \left\| y_n - x_n \right\| = 0$. Observing that

$y_n - x_n = (1 - \beta_n)(z_n - x_n)$

and condition (ii), we get $\lim_{n \to \infty} \left\| z_n - x_n \right\| = 0$. Also

$d(x_n, Tx_n) \le \left\| x_n - z_n \right\| \to 0$ as $n - \infty$.

Since T satisfies condition (I), we have $\lim_{n \to \infty} d(x_n, F(T)) = 0$. Thus there is a

subsequence $\{x_{n_k}\}$ of $\{x_n\}$ such that $\left\| x_{n_k} - q_k \right\| \le \frac{1}{2^k}$ for $\{q_k\} \subset F(T)$ and all k.

From (8), we have

$$\left\| x_{n+1} - q \right\| \le \left\| x_n - q \right\| + \alpha_n \left\| f(q) - q \right\|. \tag{11}$$

Since $\lim_{n \to \infty} \alpha_n = 0$, there exists $K \in N$ such that $\alpha_{n_k} \le \frac{1}{2^k \left\| f(q) - q \right\|}$ for all

$k \ge K$. From (11), we have

$$\left\| x_{n_{k+1}} - q_k \right\| \le \left\| x_{n_k} - q_k \right\| + \alpha_{n_k} \left\| f(q) - q \right\| < \frac{1}{2^{k-1}}, \text{ for all } k \ge K.$$

Hence $\left\| x_{n_{k+m}} - q_k \right\| < \frac{1}{2^{k-1}}$ for all $k, m \ge K$. We now show that $\{q_k\}$ is a

Cauchy sequence. For all $k, m \ge K$, we have

$$\left\| q_{k+m} - q_k \right\| \le \left\| q_{k+m} - x_{n_{k+m}} \right\| + \left\| x_{n_{k+m}} - q_k \right\| < \frac{1}{2^{k-1}} + \frac{1}{2^{k+m}} < \frac{1}{2^{k-1}}.$$

This shows that $\{q_k\}$ is a Cauchy sequence and thus converges strongly to

$p \in D$. Since

$d(q_k, Tp) \le H(Tq_k, Tp) \le \left\| p - q_k \right\|$

and $q_k - p$ as $k - \infty$. So $d(p,Tp) = 0$. Thus $p \in F(T)$ and $\{x_{n_k}\}$ converges strongly to p. If $\{x_{m_k}\}$ is another subsequence of $\{x_n\}$ which converges strongly to p^* , we have $p^* \in F(T)$. Since $Tp = \{p\}$, we have $p^* = p$. So $\{x_n\}$ converges strongly to p.

Theorem 2.4. Let X be a uniformly convex Banach space, D a nonempty closed convex subset of X, and $T : D - CB(D)$ a quasi-nonexpansive multi-valued mapping with $F(T) \neq \phi$ and $Tp = \{p\}$ for each $p \in F(T)$. Let $\{x_n\}$ defined by (3). Assume that T is hemicompact and continuous and $\alpha_n, \beta_n \in (0,1)$ satisfy

(i) $\lim\limits_{n \to \infty} \alpha_n = 0$ *and* (ii) $0 < \liminf_{n \to \infty} \beta_n \vdots \limsup_{n \to \infty} \beta_n < 1$.

Then $\{x_n\}$ converges strongly to a fixed point of T.

Proof. Let $p \in F(T)$. Then, as in the proof of Theorem 2.2. Using (3), we have

$$\|y_n - p\| = \|\beta_n x_n + (1 - \beta_n)z_n - p\| \leq \beta_n \|x_n - p\| + (1 - \beta_n)\|z_n - p\|$$

$$= \beta_n \|x_n - p\| + (1 - \beta_n)d(z_n, Tp) \leq \beta_n \|x_n - p\| + (1 - \beta_n)H(Tx_n, Tp)$$

$$\leq \beta_n \|x_n - p\| + (1 - \beta_n)\|x_n - p\| = \|x_n - p\|$$

and

$$\|x_{n+1} - p\| = \|\alpha_n f(x_n) + (1 - \alpha_n)y_n - p\| \leq \alpha_n \|f(x_n) - p\| + (1 - \alpha_n)\|y_n - p\|$$

$$\leq \alpha_n \|f(x_n) - f(p)\| + \alpha_n \|f(p) - p\| + (1 - \alpha_n)\|x_n - p\|$$

$$\leq \alpha \alpha_n \|x_n - p\| + \alpha_n \|f(p) - p\| + (1 - \alpha_n)\|x_n - p\|$$

$$= [1 - (1 - \alpha)\alpha_n]\|x_n - p\| + \alpha_n \|f(p) - p\| .$$

By simple inductions, we have

$$\|x_n - p\| \leq \max\{\|x_0 - p\|, \frac{\|f(p) - p\|}{1 - \alpha}\} , \quad \forall n \geq 1 ,$$

which gives that the sequence $\{x_n\}$ is bounded and so is $\{y_n\}$.

Next, we claim that

$$\lim_{n \to \infty} \|x_{n+1} - x_n\| = 0 . \tag{13}$$

Put $l_n = \dfrac{x_{n+1} - \beta_n x_n}{1 - \beta_n}$. Then we have

$$x_{n+1} = \beta_n x_n + (1 - \beta_n)l_n , \quad \forall n \geq 0 . \tag{14}$$

Observing that

$$l_{n+1} - l_n = \frac{\alpha_{n+1}f(x_{n+1}) + (1 - \alpha_{n+1})y_{n+1} - \beta_{n+1}x_{n+1}}{1 - \beta_{n+1}} - \frac{\alpha_n f(x_n) + (1 - \alpha_n)y_n - \beta_n x_n}{1 - \beta_n}$$

$$= \frac{\alpha_{n+1}(f(x_{n+1}) - y_{n+1})}{1 - \beta_{n+1}} - \frac{\alpha_n(f(x_n) - y_n)}{1 - \beta_n} + z_{n+1} - z_n$$

We have $\|l_{n+1} - l_n\| \le \frac{\alpha_{n+1}}{1 - \beta_{n+1}} \|f(x_{n+1}) - y_{n+1}\| + \frac{\alpha_n}{1 - \beta_n} \|f(x_n) - y_n\| + \|z_{n+1} - z_n\|$.

And we also have

$\|z_{n+1} - z_n\| \le H(Tx_{n+1}, Tx_n) \le \|x_{n+1} - x_n\|$. So

$\|l_{n+1} - l_n\| \le \frac{\alpha_{n+1}}{1 - \beta_{n+1}} \|f(x_{n+1}) - y_{n+1}\| + \frac{\alpha_n}{1 - \beta_n} \|f(x_n) - y_n\| + \|x_{n+1} - x_n\|$.

This implies that $\|l_{n+1} - l_n\| - \|x_{n+1} - x_n\| \le \frac{\alpha_{n+1}}{1 - \beta_{n+1}} \|f(x_{n+1}) - y_{n+1}\| + \frac{\alpha_n}{1 - \beta_n}$

$\|f(x_n) - y_n\|$.   Observing   conditions   (i)   and   (ii),   we   get

$\limsup_{n \to \infty}(\|l_{n+1} - l_n\| - \|x_{n+1} - x_n\|) \le 0$. From Lemma 2.1, we have

$\lim_{n \to \infty}\|l_n - x_n\| = 0$. Observing (14) yields that $x_{n+1} - x_n = (1 - \beta_n)(l_n - x_n)$.
Therefore, (12) holds. Observing that $x_{n+1} - y_n = \alpha_n(f(x_n) - y_n)$ and condition (i),
we can easily get $\lim_{n \to \infty}\|y_n - x_{n+1}\| = 0$.

On   the   other   hand,   we   have $\|y_n - x_n\| \le \|x_n - x_{n+1}\| + \|x_{n+1} - y_n\|$.   So

$\lim_{n \to \infty}\|y_n - x_n\| = 0$. Observing that $y_n - x_n = (1 - \beta_n)(z_n - x_n)$ and condition (ii),
we get $\lim_{n \to \infty}\|z_n - x_n\| = 0$.

Since $d(x_n, Tx_n) \le \|x_n - z_n\| \to 0$ as $n - \infty$ and T is hemicompact, there is a
subsequence $\{x_{n_k}\}$ of $\{x_n\}$ such that $x_{n_k} - q$ for some $q \in D$. Since T is
continuous, we have $d(x_{n_k}, Tx_{n_k}) - d(q, Tq)$. So, we have $d(q, Tq) = 0$ and so
$q \in F(T)$. If $\{x_{m_k}\}$ is another subsequence of $\{x_n\}$ which converges strongly to
q^* , we have $q^* \in F(T)$. Since $Tp = \{p\}$, we have $q^* = q$. So $\{x_n\}$ converges
strongly to q. This completes the proof of Theorem 2.4.

Corollary 2.5. Let X be a uniformly convex Banach space, D a nonempty closed
convex subset of X, and $T : D - CB(D)$ a nonexpansive multi-valued mapping with
$F(T) \ne \phi$ and $Tp = \{p\}$ for each $p \in F(T)$. Let $\{x_n\}$ defined by (3). Assume that T
is hemicompact and continuous and $\alpha_n, \beta_n \in (0,1)$ satisfy

   (i) $\lim_{n \to \infty} \alpha_n = 0$, (ii) $0 < \liminf_{n \to \infty} \beta_n \le \limsup_{n \to \infty} \beta_n < 1$.

   Then $\{x_n\}$ converges strongly to a fixed point of T.

References

1. Shiau, C., Tan, K.K., Wong, C.S.: Quasi-nonexpansive multi-valued maps and selections. Fund. Math. 87, 109–119 (1975)
2. Mann, W.R.: Mean value methods in iteration. Proc. Amer. Math. Soc. 4, 506–510 (1953)
3. Ishikawa, S.: Fixed points by a new iteration method. Proc. Amer. Math. Soc. 44, 147–150 (1974)
4. Ishikawa, S.: Fixed point and iteration of a nonexpansive mapping in a Banach space. Proc. Amer. Math. Soc. 59, 65–71 (1976)
5. Nadler Jr., S.B.: Multi-valued contraction mappings. Pacific J. Math. 30, 475–488 (1969)
6. Reich, S.: Weak convergence theorems for nonexpansive mapings in Banach spaces. J. Math. Anal. Appl. 67, 274–276 (1979)
7. Senter, H.F., Dotson, W.G.: Approximating fixed points of nonexpansive mappings. Proc. Amer. Math. Soc. 44, 375–380 (1974)
8. Tan, K.K., Xu, H.K.: Approximating fixed points of nonexpansive mappings by the Ishikawa iteration process. J. Math. Anal. Appl. 178, 301–308 (1993)
9. Sastry, K.P.R., Babu, G.V.R.: Convergence of Ishikawa iterates for a multivalued mappings with a fixed point. Czechoslovak Math. J. 55, 817–826 (2005)
10. Panyanak, B.: Mann and Ishikawa iterative processes for multivalued mappings in Banach spaces. Comput. Math. Appl. 54, 872–877 (2007)
11. Suzuki, T.: Strong convergence of krasnoselskii and manns type sequences for one-parameter nonexpansive semigroups without bochner integrals. J. Math. Anal. Appl. 305, 227–239 (2005)
12. Cholamjiak, W., Suantai, S.: Approximation of common fixed points of two quasi-nonexpansive multi-valued maps in Banach spaces. Computers and Mathematics with Applications 61, 941–949 (2011)
13. Singh, S.L., Mishra, S.N.: Fixed point theorems for single-valued and multi-valued maps. Nonlinear Analysis 74, 2243–2248 (2011)
14. Damjanović, B., Samet, B., Vetro, C.: Common fixed point theorem for multi-valued maps. Acta Mathematica Scientia 32B(2), 818–824 (2012)
15. Sintunavarat, W., Kumam, P.: Common fixed point theorem for cyclic generalized multi-valued contraction mappings. Appl. Math. Lett. (2012), doi:10.1016/j.aml.2012.02.045
16. Sintunavarat, W., Kumam, P.: Common fixed point theorem for hybrid generalized multi-valued contraction mappings. Appl. Math. Lett. 25, 52–57 (2012)
17. Ofoedu, E.U., Zegeye, H.: Iterative algorithm for multi-valued pseudocontractive mappings in Banach spaces. J. Math. Anal. Appl. 372, 68–76 (2010)
18. Khojasteh, F., Rakočević, V.: Some new common fixed point results for generalized contractive multi-valued non-self-mappings. Appl. Math. Lett. 25, 287–293 (2012)
19. Kiran, Q., Kamran, T.: Fixed point theorems for generalized contractive multi-valued maps. Computers and Mathematics with Applications 59, 3813–3823 (2010)
20. Gordji, M.E., Baghani, H., Khodaei, H., Ramezani, M.: A generalization of Nadler's fixed point theorem. J. Nonlinear Sci. Appl. 3(2), 148–151 (2010)

A Hand Model Updating Algorithm
Based on Mean Shift

Xiao Zou, Heng Wang, HongXiang Duan, and QiuYu Zhang

School of Computer and Communication, Lanzhou University of Technology
730050 Lanzhou, China
zouxissao@lut.cn

Abstract. To solve the problem of target model changing and influenced tracking results in gesture target tracking process, this paper proposed a gesture model updating and results forecasting algorithm based on Mean Shift. The algorithm firstly uses the background subtraction and skin color detection methods to detect and obtain gesture modeling, and then uses the Mean Shift algorithm to track gesture and update the gesture model, finally uses the Kalman algorithm to predict the gesture tracking results. The experimental results show that this algorithm reduces influence of surrounding environment in gesture tracking process, tracking effect is good.

Keywords: Mean Shift Algorithm, Hand Tracking, Gesture Model.

1   Introduction

As a preprocessing step for gesture recognition, gesture tracking has a great influence on the gesture recognition. The exact gesture tracking results can plays a positive role in the gesture recognition effect [1-3].

Currently the most popular method is Mean Shift algorithm which was put forward by Faunae and Hostetler in 1975 [2, 3]. The rapid development of Mean Shift algorithm is implementing by Comaniciu ET c. The algorithm is applied to tracking field by them [4, 5-7], and the variable window width Mean Shift algorithm is also studied [8-10]. Peng et al [6, 11] proposed the fore cast Camshift tracking results by using Kalman filter, in order to reduce the interference of large area skin color. Wang et al [7] used K-means clustering algorithm and particle filter method to solve the problem of mutual interference of finger tracking, and achieved good tracking effect. Feng et al [8] proposed a particle filter gestures tracking algorithm based on microscopic structure of state variables, which could describe the microstructure of the hand state variables from the continuous deformation time model, the period of time for deformation model, and mutations strength and self occlusion characteristics. Gan et al [9, 12] proposed a kind of target tracking algorithm based on the Mean Shift algorithm and normalized moment of inertia characteristics, and it could reach real-time and stable tracking when the moving air target was occluded and in large deformation.

However, when the gesture was affected by the changing surrounding environment, such as the light suddenly becomes strong, or become weak or the non target interference became more serious, tracking may be make a fail because of the excursion of the target model [13-16]. This paper proposes a gesture tracking method based on the

updating and forecasting of the target model. Experiments show that the algorithm proposed in the paper is effective when the gesture in a complex environment.

2   Hand Extracting

2.1   Skin Color Detection

In YC_bC_r color system, skin color distributes relatively concentration in C_b, C_r, mainly in the $R_{Cb} = [77,127]$ $R_{Cr} = [133,173]$ [10, 17, 18]. Therefore, judge each pixel of the image as follows:

$$O(x, y) = \begin{cases} 1, \left[C_r(x, y) \in R_{Cr} \right] \cap \left[Cb(x, y) \in R_{Cb} \right] \\ 0, else \end{cases} \quad (1)$$

The judgment result is the skin color detection result.

2.2   Background Subtraction Method

By using the appropriate background subtraction algorithm based on motion information, and combining with skin color detection results to segment the gesture region, in order to obtain the better effect [19-21].

$$D_t(x, y) = \min | I_t(x, y) - B_t(x, y) | \quad (2)$$

Where $I_t(x, y)$ the input is value of pixels (x, y) in t time; $B_t(x, y)$ is the constructing background image of pixels (x, y) in t time, and then, on segment movement part [22, 23]:

$$M_t(x, y) = \begin{cases} 1, D_t(x, y) > \sigma_b \\ 0, else \end{cases} \quad (3)$$

Where σ_b is threshold

Thus, gain the gestural information from skin color information and motion information. This can ensure the gesture segmentation and tracking.

3   Mean Shift Algorithms

Mean Shift algorithm is a common algorithm in the gesture tracking methods based on vision.

The mean shift iterative formula which comes from kernel density estimation theory expressed as [24-26]:

$$y_{j+1} = \frac{\sum_{i=1}^{n} x_i \omega_i g \left(\left\| \frac{y_j - x_i}{h} \right\|^2 \right)}{\sum_{i=1}^{n} g \left(\left\| \frac{y_j - x_i}{h} \right\|^2 \right)} \tag{4}$$

Where $\{y_j \mid j = 1, 2...\}$ is a series of target position, and y_1 was initialized for nuclear density center

Assuming the object model [27-29]:

$$\hat{q} = \{\hat{q}_i \mid i = 1...n, \sum_{i=1}^{n} \hat{q}_i = 1\} \tag{5}$$

Candidate targets:

$$\hat{p}(y) = \{\hat{p}_i(y) \mid i = 1...n, \sum_{i=1}^{n} \hat{p}_i = 1\} \tag{6}$$

The target model and the candidate model can be expressed as [30]:

$$\hat{q}_i = C \sum_{u=1}^{n} k \left(\left\| x_{ui} \right\|^2 \right) \delta \left[b(x_u) - i \right] \tag{7}$$

$$\hat{p}(y) = C_h \sum_{u}^{n_h} k \left(\left\| \frac{y - x_u}{h} \right\|^2 \right) \delta \left[b(x_u) - i \right] \tag{8}$$

Where y is a candidate target position accordingly, the evaluation coefficient Bhattacharyya of the target is defined as follows:

$$\hat{\rho}(y) \equiv \rho[\hat{p}(y), \hat{q}] = \sum_{i=1}^{n} \sqrt{\hat{p}_i(y) \hat{q}_i} \tag{9}$$

The distance between the two distributions is:

$$d(y) = \sqrt{1 - \rho[\hat{p}(y), \hat{q}]} \tag{10}$$

Target tracking is to find the location y with maximum similarity ρ in each frame. If the target position of the first frame is y_0, Taylor expansion them $\rho(y)$ in y_0, and the similar functions are expressed as:

$$\rho(y) \approx \frac{1}{2} \sum_{1}^{m} \sqrt{\hat{p}_i(y_0) \hat{q}_i} + \frac{C_h}{2} \sum_{u=1}^{n_h} \omega_u k \left(\left\| \frac{y - x_u}{h} \right\|^2 \right) \tag{11}$$

Where

$$\omega_u = \sum_{i=1}^{m} \sqrt{\frac{q_u}{p_u(y_0)}} \delta \left[b(x_u) - i \right] \tag{12}$$

4   Target Model Updating

At present the main updating methods of tracking target model are: the updating strategy based on the whole [14, 15] and the updating strategy based on selecting factor [16, 17]. From the perspective of real-time hand tracking, tracking process needs to reduce the calculating amount, and reduce the computational cost. So, this paper take the target gesture model as a whole, make a weighted summation to the current target gesture model and target historical gesture model, and then get the object model in the next frame. The definitions as follow:

$$q^t = (1-\alpha)q^{t-1} + \alpha p^t \qquad (13)$$

where q^t is the updated target gesture model; p^t is the tracked target gesture model; q^{t-1} is the previous frame target gesture model, α is the tracking result weights of the current frame, also known as the speed factor of model updating.

Use coefficient Bhattacharyya on gesture tracking effect evaluation in the updating process. Set update model threshold as ρ_{Th}, and the current tracking results is ρ_t, and the target gesture model is updated as follows:

$$q^t = \begin{cases} q^{t-1}, & \rho_t \geq \rho_{Th} \\ (1-\alpha)q^{t-1} + \alpha p^t, & \rho_t < \rho_{Th} \end{cases} \qquad (14)$$

5   Hand Tracking and Prediction

In a gesture tracking process, the gesture tracking results may appear deviation due to the disturbance of the ambient environment. At short notice, a gesture of motion can be viewed as uniform motion in straight line. This paper uses the Kalman filter to predict the gesture tracking results [18].

Define the observation vector as: $Z_k = (x_k, y_k)$ state vector: $X_k = (x_k, y_k, u_k, v_k)$, then the predicted state vector is: $X_k' = (x_k', y_k', u_k', v_k')$. State transition matrix:

$$A = \begin{bmatrix} 1 & 0 & T & 0 \\ 0 & 1 & 0 & T \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad (15)$$

T is time difference of K moment and $K-1$ moment the observation matrix.

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \qquad (16)$$

Get the system state equation:

$$X_{k+1} = AX_k + W_k \qquad (17)$$

System of observation equation:

$$Z_k = HX_k + V_k \tag{18}$$

Where the measurement noise v_k obey P (V) ~ N (0, R), dynamic noise w_k obey P (W) ~ N (0, Q), and the two noise are uncorrelated white noise. R Is the correlation matrix of measurement noise, Q is the correlation matrix of motion noise.

The prediction equation of the system is:

$$x'_k = A \cdot x_{k-1} + B \cdot u_k \tag{19}$$

The prediction equation of the prior error matrix is:

$$P'_k = A \cdot P_{k-1} \cdot A^T + Q \tag{20}$$

Kalman gain equation:

$$K_k = P'_k \cdot H^T \cdot (H \cdot P'_k \cdot H^T + R)^{-1} \tag{21}$$

State change equation:

$$x_k = x'_k + K_k \cdot (z_k - H \cdot x'_k) \tag{22}$$

Covariance change equation:

$$P_k = (I - K_k \cdot H) \cdot P'_k \tag{23}$$

6   Algorithm Descriptions

The gesture model updating and results forecasting method based on Mean Shift algorithm is described as follows:

Step 1: Judge skin color regions by using thresholding method, and obtain the gesture tracking target preliminarily;

Step 2: Use the background subtraction algorithm which based on motion information, combine with the skin color detection in step 1; get the tracking target accurately;

Step 3: Calculate the candidate target model $p(y_0)$ and the target model of Bhattacharyya coefficient $\rho(y_0)$;

Step 4: Calculate ω_i by using formula (12), calculate the target new position y_1 according to equation (4) and calculate $p(y_1)$ and $\rho(y_1)$;

Step 5: If $\rho(y_1) < \rho(y_0)$, then $y_1 < -(y_0 + y_1)/2$ and go to step 6;

Step 6: If $\|y_1 - y_0\| < \varepsilon$, then iterative end; otherwise make $y_0 = y_1$, and turn to step 3;

Step 7: After the final tracking, calculate the similarity coefficient ρ_t, if $\rho_t < \rho_{th}$, update the target model, otherwise, turn to the next tracking.

Step 8: Predict the tracking results using the last tracking results, and modify tracking results according to the predicted result.

The gesture model updating and results forecasting method based on Mean Shift Code (Section) is described as follows:

```
void CObjectTracker::ObjeckTrackerHandlerByUser(IplImage
*frame)//Tracking function
{
m_cActiveObject = 0;
if (m_sTrackingObjectTable[m_cActiveObject].Status)
{
if
(!m_sTrackingObjectTable[m_cActiveObject].assignedAnObjec
t)
{
FindHistogram(frame,m_sTrackingObjectTable[m_cActiveObjec
t].initHistogram);
m_sTrackingObjectTable[m_cActiveObject].assignedAnObject =
true;
}
…
}
}
void CObjectTracker::FindNextLocation(IplImage *frame)
{
int i, j, opti, optj;
SINT16 scale[3]={-3, 3, 0};
FLOAT32 dist, optdist;
SINT16 h, w, optX, optY;
FindNextFixScale(frame);
optdist=LastDist;
optX=m_sTrackingObjectTable[m_cActiveObject].X;
optY=m_sTrackingObjectTable[m_cActiveObject].Y;
…
{
m_sTrackingObjectTable[m_cActiveObject].H=h+scale[i];
m_sTrackingObjectTable[m_cActiveObject].W=w+scale[j];
FindNextFixScale(frame);
if( (dist=LastDist) < optdist ) //scaling is better
{
optdist=dist;
// printf("Next%f->/n", dist);
}
else //no scaling is better
{
m_sTrackingObjectTable[m_cActiveObject].X=optX;
m_sTrackingObjectTable[m_cActiveObject].H=h;
m_sTrackingObjectTable[m_cActiveObject].W=w;
…
}
};
TotalDist+=optdist; //the latest distance
}
```

7   Simulation Experiment and Results Analysis

In order to verify the accuracy, real-time and continuous of the algorithm, use 30 frames per second, with dimensions of 640×480 sign language video as the contrast material. Take VC++6.0 as the development platform at the Windows XP SP3 environment, with AMD Athlon 7750 Dual-Core Processor 2.70GHz, RAM 2GB.

Compare with the CamShift algorithm, the experiment results is shown in Table 1. The experimental results chart is divided into three layers, where the first layer is a video extracted from the original image, the second layer is tracking results of CamShift algorithm, the third layer is the tracking results of the algorithm proposed in the paper. Seen in the table, in tracking process CamShift tracking algorithm appeared tracking error because of the other color background. Especially when the gesture tracking through the same or similar color region, such as the seventy-seventh frame, eighty-third frame, eighty-fifth frame and ninety-first frame, CamShift algorithm tracked the wrong target. While our approach got better tracking results.

Table 1. Comparison of tracking results
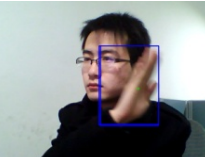
Frame	The Original Image	VideoCamShift Algorithm Tracking Results	This Algorithm Tracking Results
The seventeenth frame			
The thirty-sixth frame			
The fifty-seventh frame			
The seventy-seventh frame			

Table 1. (*Continued.*)

The eighty-third frame			
The eighty-fifth frame			
The ninety-first frame			
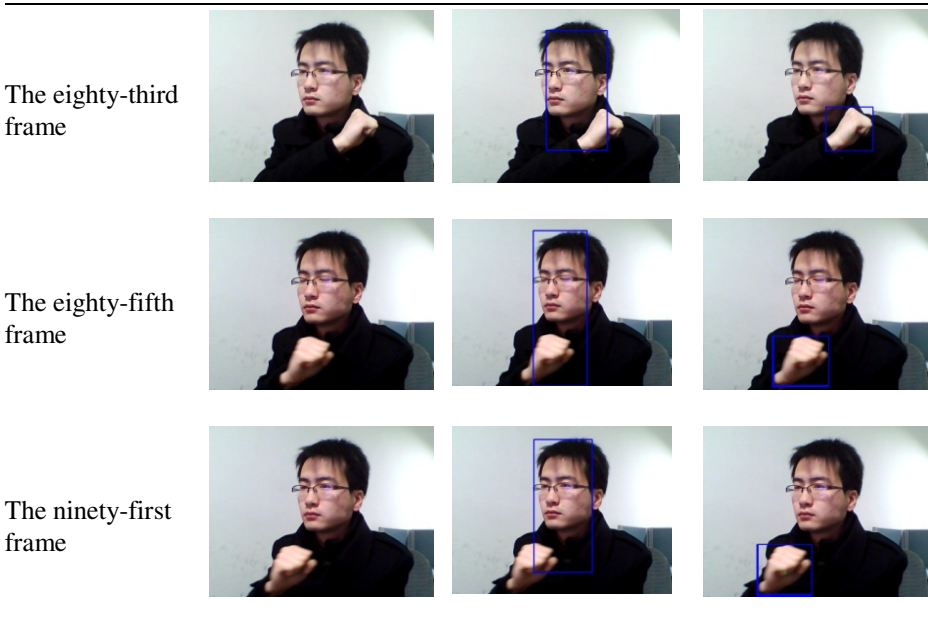
8   Conclusions

To propose a gesture model updating and results forecasting algorithm based on Mean Shift, and to solve the problem of target model changing and influenced tracking results in gesture target tracking process. Firstly, the background difference and skin color detection methods are used to detect and get gesture model, and the Mean Shift algorithm is used to track gesture and update the gesture model, and finally to use the Kalman algorithm to predict the gesture tracking results. The experimental results show that this algorithm reduces the influence of surrounding environment in gesture tracking process, and get better tracking result.

However, the algorithm also has its flaws. For example in the shade and shelter, the hand gesture tracking is inaccurate. And also the contour description problem is still not solved. We will study it in future.

References

1. Ren, H.B., Zhu, Y.X., Xu, G.Y., et al.: Vision-Based Recognition of Hand Gestures: A Survey. Acta Electronica Sinica 2(28), 118–122 (2000)
2. Comaniciu, D., Meer, P.: Mean shift analysis and applications. In: Proceedings of IEEE International Conference on Computer Vision, Kerkira, Greece, vol. 52(11), pp. 1197–1203 (1999)

3. Hou, Z.Q., Han, C.Z.: A Survey of Visual Tracking. Acta Automatica Sinica 32(48), 603–617 (2006)
4. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-Based Object Tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(5), 564–577 (2003)
5. Comaniciu, D., Ramesh, V., Meer, P.: The variable bandwidth Mean Shift and data-driven scale selection. In: Proceedings of IEEE International Conference on Computer Vision, Vancouver, BC, vol. 30(6), pp. 438–445 (2001)
6. Peng, J.C., Gu, L.Z., Su, J.B.: The Hand Tracking for Humanoid Robot Using Camshift Algorithm and Kalman Filter. Journal of Shanghai Jiaotong University 40(7), 1161–1165 (2006)
7. Wang, X.Y., Zhang, X.W., Dai, G.Z.: An Approach to Tracking Deformable Hand Gesture for Real-Time Interaction. Journal of Software 18(10), 2423–2433 (2007)
8. Feng, Z.Q., Yang, B., Li, Y., et al.: Research on Hand Gestures Tracking Based on Particle Filtering Aiming at Optimizing Time Cost. Acta Electronica Sinica 37(9), 1989–1995 (2009)
9. Gan, M.G., Chen, J., Wang, Y.L., et al.: A Target Tracking Algorithm Based on Mean Shift and Normalized Moment of Inertia Feature. Acta Automatica Sinica 36(9), 1332–1336 (2010)
10. Chen, D.S., Liu, Z.K.: A Survey of Skin Color Detection. Chinese Journal of Computers 29(2), 194–207 (2006)
11. Chai, D., Gan, K.: Locating Facial Region of a Head-and-Shoulders Color Image. In: Proceedings of the 3rd International Conference on Automatic Face and Gesture Recognition, Nara, Japan, vol. 34(5), pp. 124–129 (1998)
12. Xiao, M., Han, C.Z., Zhang, L.: Moving Object Detection Algorithm Based on Space-Time Background Difference. Journal of Computer-Aided Design & Computer Graphics 7(18), 1044–1048 (2006)
13. Comaniciu, D., Ramesh, V., Meer, P.: Real-Time Tracking of Non-Rigid Objects using Mean Shift. In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 12(34), pp. 142–149 (2000)
14. Nummiaro, K., Koller-Meier, E., Van Gool, L.: An adaptive color-based particle filter. Image and Vision Computing 21(1), 99–110 (2002)
15. Nummiaro, K., Koller-Meier, E., Van Gool, L.: Object Tracking with an Adaptive Color-Based Particle Filter. In: Van Gool, L. (ed.) DAGM 2002. LNCS, vol. 2449, pp. 353–360. Springer, Heidelberg (2002)
16. Shen, Z.X., Yang, X., Hang, X.Y.: Study on Target Model Update Method in Mean Shift Algorithm. Acta Automatica Sinica 35(5), 478–483 (2009)
17. Venkatesh Babu, R., Perez, P., Bouthemy, P.: Robust tracking with motion estimation and local Kernel-based color modeling. Image Vision Computing 25(8), 1205–1215 (2007)
18. Luo, Y., Li, L., Zhang, B.S., Yang, H.M.: Video hand tracking algorithm based on hybrid CamShift and Kalman filter. Application Research of Computers 26(3), 1163–1165 (2009)
19. Aggarwal, G., Ghosal, S., Dubey, P.: Efficient Query Modification for Image Retrieval. In: Proc. 2000 IEEE Conf. Computer Vision and Pattern Recognition, vol. II, pp. 255–261 (June 2000)
20. Bajcsy, R., Lee, S.W., Leonardis, A.: Detection of Diffuse and Specular Interface Reflections and Inter-Reflections by Color Image Segmentation. Int'l J. Computer Vision 17, 241–272 (1996)
21. Barash, D.: Bilateral Filtering and Anisotropic Diffusion: Towards a Unified Viewpoint. IEEE Trans. Pattern Analysis and Pattern Analysis (to appear)
22. Bertsekas, D.P.: Nonlinear Programming. Athena Scientific (1995)

23. Black, M.J., Sapiro, G., Marimont, D.H., Heeger, D.: Robust Anisotropic Diffusion. IEEE Trans. Image Processing 7, 421–432 (1998)
24. Bradski, G.R.: Computer Vision Face Tracking as a Component of a Perceptual User Interface. In: Proc. IEEE Workshop Applications of Computer Vision, pp. 214–219 (October 1998)
25. Cheng, Y.: Mean Shift, Mode Seeking, and Clustering. IEEE Trans. Pattern Analysis and Machine Intelligence 17(8), 790–799 (1995)
26. Choi, E., Hall, P.: Data Sharpening as a Prelude to Density Estimation. Biometrika 86, 941–947 (1999)
27. Chu, C.K., Glad, I.K., Godtliebsen, F., Maron, J.S.: Edge-Preserving Smoothers for Image Processing. J. Am. Statistical Assoc. 93, 526–541 (1998)
28. Comaniciu, D.: Nonparametric Robust Methods for Computer Vision, PhD thesis, Dept. of Electrical and Computer Eng., Rutgers Univ. (1999), `http://www.caip.rutgers.edu/riul/research/theses.html`
29. Comaniciu, D., Meer, P.: Robust Analysis of Feature Spaces: Color Image Segmentation. In: Proc. 1997 IEEE Conf. Computer Vision and Pattern Recognition, pp. 750–755 (June 1997)
30. Comaniciu, D., Meer, P.: Distribution Free Decomposition of Multivariate Data. Pattern Analysis and Applications 2, 22–30 (1999)

Retraction Note to: New Algorithms for Solving RFML*r*R Circulant Linear Systems in Information

Jianhua Liu, Dengjie Wu, and Zhao-Lin Jiang

Department of Mathematics, Linyi University, Linyi 276005, P.R. China
jhliu@lyu.edu.cn, wdj201303l3@l63.com,
jzhl208@sina.com

Retraction Note to:
Chapter "New Algorithms for Solving RFML*r*R
Circulant Linear Systems in Information" in:
Y. Yang, M. Ma, and B. Liu (Eds.):
Information Computing and Applications, CCIS,
DOI: 10.1007/978-3-642-53932-9_17

The paper starting on page 169 of this publication has been retracted because it does not differ sufficiently from the paper starting on page 59 of the same publication. One of the authors is the same.

Author Index