Lecture Notes in Electrical Engineering 236

W. Eric Wong
Tinghuai Ma
Editors

Emerging Technologies for Information Systems, Computing, and Management

Springer

Lecture Notes in Electrical Engineering

Volume 236

For further volumes:
http://www.springer.com/series/7818

W. Eric Wong · Tinghuai Ma
Editors

Emerging Technologies for Information Systems, Computing, and Management

Springer

Editors
W. Eric Wong
Department of Computer Science
University of Texas at Dallas
Richardson, TX
USA

Tinghuai Ma
College of Computer and Software
Nanjing University of Information Science and Technology
Nanjing, Jiangsu
People's Republic of China

ISSN 1876-1100 ISSN 1876-1119 (electronic)
ISBN 978-1-4614-7009-0 ISBN 978-1-4614-7010-6 (eBook)
DOI 10.1007/978-1-4614-7010-6
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013938268

Printed on acid-free paper

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

Preface

This book contains papers accepted by the 2012 *International Conference on Emerging Technologies for Information Systems, Computing, and Management* (ICM), which was held from 8 to 9 December, 2012 in Hangzhou, China.

It covers emerging topics in a timely manner for information systems, computing, and management. In particular, it helps researchers and students improve their knowledge of state-of-art in related areas, and it provides engineers and practitioners with useful information on how to improve their productivity by using the latest advanced technologies.

The book has two volumes with the first focusing on Information Systems, Algorithms and Applications, and Pattern Recognition, whereas the second includes papers in the areas of Data Processing, System Identification, and Management Science. Both volumes can be used as excellent references by industry practitioners, faculty, and students who must get up to speed on the most recent technology development and current state-of-practice for using computing services efficiently and effectively to produce, maintain, and manage large complicated trustworthy systems, which have profound impacts on everyone's daily life.

We would like to thank all the authors for sharing their ideas, research results, authentic industry experiences and professional best practices, as well as all the reviewers for their help in evaluating and selecting high quality papers. Without their participation, the publication of this book would not be possible. Special thanks also go to Mr. Brett Kurzman at Springer US for all his valuable assistance with other publication-related logistics.

We hope you enjoy reading the book.

W. Eric Wong
Tinghuai Ma

ICM2012 Committee

General Chairs

W. Eric Wong	University of Texas at Dallas, USA
Qingyuan Tian	Beijing Language and Culture University, China

Program Chairs

Shuangqin Liu	Hohai University, China
Tinghuai Ma	Nanjing University of Information Science and Technology, China

Program Committee

Akshay Girdhar	Guru Nanak Dev Engineering College, India
Artur Opalinski	Gdansk University of Technology, Poland
Chang Kuan-Tsung	Ming-hsin University of Science and Technology, Taiwan, China
Chao Wang	Oregon Health & Science University, USA
DaeHee Seo	Electronics and Telecommunication Research Institute, Korea
Dehong Li	National Institute of Metrology, China
Qingchao Dong	PLA University of Science and Technology, China
Fa Zhang	Chinese Academy of Sciences, China
Fangqiang Chang	Huaqiao University, China
Fufang Li	Guangzhou University, China
Gang Hu	Xi'an University of Technology, China
Gang Zhang	Guangdong University of Technology, China
Hasim Altam	University of Sheffield, England
Hemanga Krishna Borah	Tata Consultancy Services, England
Heng Luo	Suzhou University, China
Huizhu Ma	Harbin Engineering University, China

Jing Chen	University of Macau, China
Jingshan Li	Heilongjiang Institute of Science and Technology, China
Jingyan Wang	King Abdullah University of Science and Technology, Saudi Arabia
Jisheng xia	Yunnan University, China
Kriti Srivastava	Dwarkadas J. Sanghvi College of Engineering, Mumbai
Lei Sun	China University of Petroleum, China
Lijia Xu	Sichuan Agricultural University, China
Lin He	Dalian Maritime University, China
Linhua Jiang	Leiden University, Holland
Liu peng	Chong Qing University, China
Marko Jäntti	University of Eastern Finland, Finland
Mehmet Celenk	Ohio University, USA
Mengshi Li	South China University of Technology, China
Michael Collier	Shandong University of Science and Technology, China
Niu Dan	Waseda University, Japan
Qingli Li	East China Normal University, China
Ru Guo	Tongji University, China
Ruihua Ding	University of Washington, USA
Tao Li	Nankai University, China
Tse-Chen Yeh	Academia Sinica, Taiwan, China
Wancai Li	The Third Research of Ministry of Public Security, China
Wang donglei	Chinese Academy of Engineering Physics, China
Wei Sun	Shanghai Maritime University, China
Weidong Sun	Tsinghua University, China
Wenju Liu	Chinese Academy of Sciences, China
Wenpeng Wang	Qingdao University of Science and Technology, China
Xianfang Tan	University of Nevada, Las Vegas, USA
Xiangzhong Xu	Armored Force Engineering College English, China
Xianling Mao	Peking University, China
Xiao Cheng	Beijing Normal University, China
Xiaoguang hu	Wuhan University, China
Xin He	Henan University, China
Xuefang Zhu	Nanjing University, China
Xueyi Fang	Zhejiang University, China
Yudong Zhang	Columbia University, USA
Yuliang Ma	Hangzhou Dianzi University, China
Yuqing Zhang	China University of Geosciences (Beijing), China

Zheng Liang	National University of Defense Technology, China
Zhenhai Wang	Linyi University, China
Zhiqun Li	Southeast University, China
Zhuangzhi Guo	Guangdong University of Technology, China
Ziwei Dong	Wuhan Ordnance Petty Officer School, China

Contents

Part IV Data Processing

Part V System Identification

Part I
Information Systems

Chapter 1
The Design for High Dynamic GPS Receiver in a Combinated Method of FLL and PLL

Na Shen and Xiangjin Zhang

Abstract To solve the problems that the large Doppler shift makes satellite signals difficult to be captured and the high dynamic stress damages the health of the tracking loop, a novel design is proposed for high dynamic GPS receiver, which can rapidly capture signals against the large Doppler shift and has high tolerance of the high dynamic stress in its tracking loop. The satellite signals are captured by means of linear search method. The carrier wave is tracked by carrier tracking loop using Frequency-locked loop (FLL) assisted by phase-locked loop (PLL). The form of Carrier wave aided is adopted in the pseudo-code tracking loop. The experiment shows the receiver prototype can track 100 g/s high dynamic signal, which indicates that the receiver prototype could satisfy the error limits of the tracking loop, and it also can capture the visible satellite signals with a good effect of real-time tracking lock.

Keywords GPS receiver · Frequency-locked loop (FLL) · Phase-locked loop (PLL)

1.1 Introduction

Global Positioning System (GPS) has been developed for almost 30 years. However the study of high dynamic GPS receiver starts relatively late, it still needs further research. The difficulties in the design of high dynamic GPS receiver are listed as following [1]: (1) the satellite signals are difficult to be captured because

N. Shen (✉)
The National key lab of instant physics, Nanjing University of Science and Technology, Nanjing, China
e-mail: snbox@163.com

X. Zhang
ZNDY of Ministerial Key Laboratory, Nanjing University of Science and Technology, Nanjing, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_1,

the carrier waves received with high dynamic have a large Doppler shift; (2) The tracking loop must be designed rationally to the satellite signals with high tolerance of the high dynamic stress. In order to track the carrier wave stably, Literature [2] proposes a method using INS auxiliary carrier loop, Literature [3] develop a carrier tracking algorithm used in high dynamic and static environments based on the nature of adaptive bandwidth adjusting of IMM. However, both of the two methods have complicated structure and cost highly.

In this paper, a novel design program is proposed for high dynamic GPS receiver. It uses the method of frequency-locked loop (FLL) assisted by phase-locked loop (PLL), which has a good effect and lower cost compared with the above two methods. The system uses the frequency (RF) chip GP2015 as the RF front-end. And the baseband part consists of digital signal processor DSP6713 and the 12-channels correlator GP2021 which form the desired digital tracking loop.

1.2 Hardware Components

Figure 1.1 shows the hardware system of the whole GPS receiver. The GPS RF signals are received by the antenna, and then amplified by the low noise preamp with 2.4 MHz bandwidth. After filtering the signals are amplified by a passive band pass amplifier between the antenna and preamp to minimize out-of-band RF interference. The RF front-end is composed of the GP2015 and peripheral circuits, it is used to convert those RF signals to an intermediate frequency (IF). Due to the Doppler shift, the frequency of the received signal would not be 1575.42 MHz accurately, there will be some deviation. The GP2015 degrades the signal frequency from 1575.42 to 4.309 MHz by three-stage down-conversion [4]. The first level and second level of the filter must be designed separately, and the third level of the filter has been integrated within the chip. After down-conversion, the IF analog signal is discrete into two bits digital signal after sampling, one is amplitude bit, and another one is sign bit. The sampling frequency is 5.714 MHz and the digital IF signal after sampling is 1.405 MHz. The baseband signal processing part is complemented

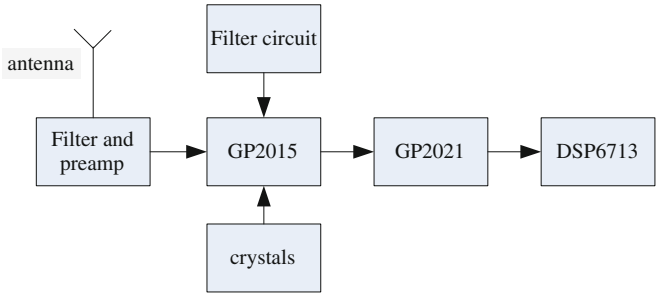


Fig. 1.1 GPS receiver hardware system block diagram

through the combination of hardware and software, which includes the correlator and the DSP, playing a role of capturing and tracking the satellite signals. The correlator controlled by the DSP has integrated 12 correlation channels, as well as Carrier generator, the pseudo-code generator and the points clear.

1.3 Satellite Signals Acquisition

The algorithm of satellite signal acquisition is mainly linear search, parallel frequency search and parallel code phase search method. In this paper, a linear search method to search for satellite signals with the multi-channels correlator because this method is relatively simple.

Figure 1.2 shows the block diagram of the linear search for satellite signal. Firstly, the receiver searches the satellite signal with step of 500 Hz. When the satellite signals are searched, a 100 Hz frequency is adopted to refine the Doppler shift and the estimated value of the code phase. After refining, the capture should be verified, and then the satellite signals are captured for certain (Fig. 1.3).

At the beginning of search, the eudipleural search is executed with the search step-length of 500 Hz and centered on 1.405 Hz. When the signal amplitude calculated in a search unit is greater than the predefined threshold, there are may be some signals in the unit. Then make the unit as a search centre, with the search step 100 Hz, search seven times symmetrically to find out the unit which has the largest non-coherent integral amplitude as the capture result. By doing this, we can control the frequency error within the range of 50 Hz. After captured the satellite

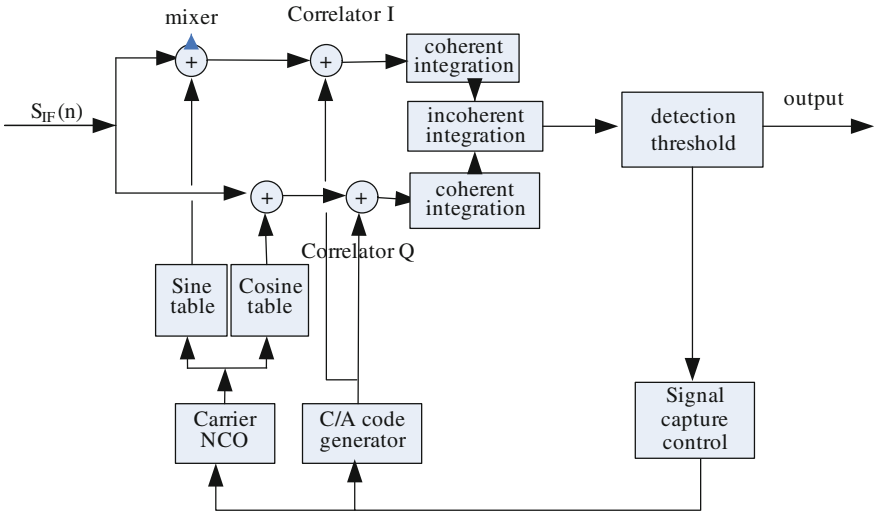


Fig. 1.2 The block of linear search method

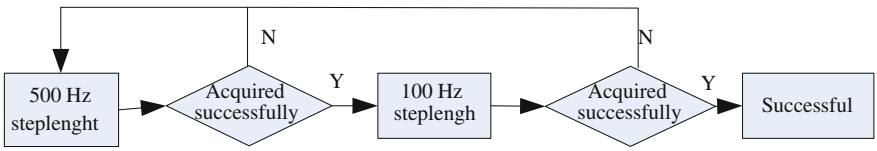


Fig. 1.3 The search flowchart

signals, we use Tong Search Recognition Act to make sure the unit does exist in the visible satellite signal. In the search process, the pseudo-code search step is 0.5 code chip, and with the error of 0.25 code chip.

1.4 Satellite Signals Tracking

After the satellite signals have been captured, the satellite signals should be tracked in real time. A combination method of FLL, the PLL and the code loop is used in here. When designing the tracking loop, the order of the tracking loop should be set firstly, and then the bandwidth of the tracking loop is selected, which is the most important parameter in the design of the tracking loop.

1.4.1 Carrier Tracking Loop

A combination of PLL and FLL is used for Carrier tracking loop. The PLL uses a narrower noise bandwidth, which can track the signals more closely. The output of the carrier phase measurement are more accurately, and the demodulated data bit error rate is low, but it has a poor performance under the dynamic stress; While FLL can adopt a wide noise bandwidth, it has a good dynamic tracking performance while tracking the signals with the low signal to noise ratio. However, the FLL signal tracking is not close enough and the demodulation of the bit error rate is high [5].

Consider the relative motion between the satellite and user, the GPS receiver will be interfered by the excitation signal of frequency ramp. Under this interference, only the third-order or more than third-order phase-locked loop can track the signal accurately. However the second-order FLL can achieve the performance of third-order PLL, therefore, a 3rd-order PLL assisted by 2nd-order FLL will adopted to track the signals.

The following table lists the characteristics of the phase detector and the frequency detector in this paper (Table 1.1).

When the phase difference or frequency difference are detected, a loop filter is used to filter high frequency signal and noise. The block diagram of the loop filter is shown in Fig. 1.4.

Table 1.1 The Discriminator characteristics

	Calculation method	Advantage	Disadvantage
Phase detection algorithms	$\Phi_e = \arctan(Q/I)$	Phase accurate, insensitive to the data transition	Calculation large
Frequency detection algorithms	$w_e = \frac{P_{cross} \times sign(P_{dot})}{t(n)-t(n-1)}$	Calculation small, insensitive to the data transition	Frequency pull-in range of small

It can be seen from the figure that the loop filter of the FLL is one integrator ahead than the PLL filter. Since the output of the frequency discriminator is frequency difference, it changes to be a phase difference after passing away the integrator, and then the local carrier frequency is adjusted by adjusting the digital controlled oscillator (DCO).

The process of loop filter which is implemented by software is shown as follows.

$$X_1(n) = \Delta f(n)\omega_{nf}^2 T + \Delta p(n)\omega_{np}^3 T + X_1(n-1) \tag{1.1}$$

$$\begin{aligned} X_2(n) = {} & +\Delta p(n)a_3\omega_{np}^2 T + X_2(n-1) + [\Delta f(n)\omega_{nf}^2 T + \Delta p(n)\omega_{np}^3 T \\ & + 2X_1(n-1)]T/2 + \Delta f(n)a_2\omega_{nf} T \end{aligned} \tag{1.2}$$

$$\begin{aligned} u_f(n) = {} & [\Delta f(n)a_2\omega_{nf} + \Delta p(n)a_3\omega_{np}^2]T/2 + [\Delta f(n)\omega_{nf}^2 T + \Delta p(n)\omega_{np}^3 T \\ & + 2X_1(n-1)]T/4 + X_2(n-1) + \Delta p(n)b_3\omega_{np} \end{aligned} \tag{1.3}$$

where $X_1(n)$ and $X_2(n)$ are the intermediate variables, $\Delta f(n)$ and $\Delta p(n)$ are the output of the frequency discriminator and the phase detector, respectively. $u_f(n)$ is the output of the loop filter which controls the frequency of the carrier digital controlled oscillator to adjust the output signal frequency.

The parameters of loop filter are shown in Table 1.2 [6].

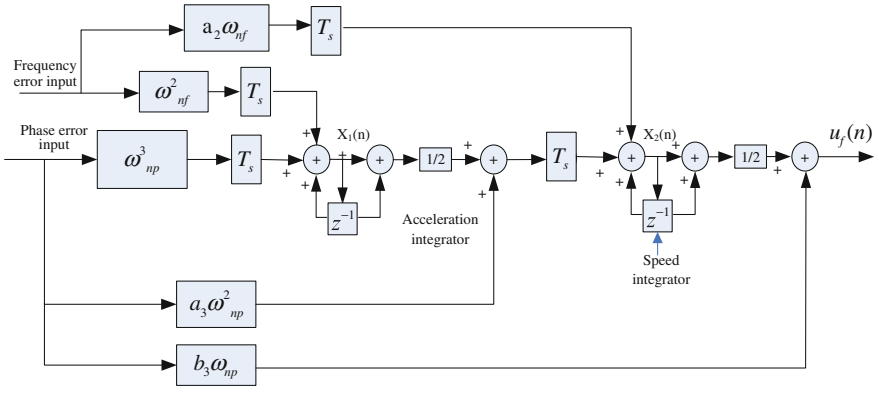


Fig. 1.4 The carrier tracking loop

1.4.2 Pseudo-code Tracking Loop

The code loop assisted by carrier is adopted in this system. The Doppler shift measured values of the carrier loop can be used to assist the code loop for adjusting the bit rate. Because the Doppler frequency shift of the pseudo-code is much smaller than the carrier Doppler frequency shift, the carrier aiding is 1/1540 of carrier wave. Figure 1.5 shows the block diagram of code tracking loop.

With the aid of the carrier loop, second-order code loop do not need high order filter any more. So a two-order code loop is adopted here. The non-coherent early minus late method is used while detecting the phase difference in the code loop. Calculation method is shown as follows [7].

$$\frac{1}{2}\frac{\sqrt{\mathrm{I}_E^2+\mathrm{Q}_E^2}-\sqrt{I_p^2+Q_P^2}}{\sqrt{\mathrm{I}_E^2+\mathrm{Q}_E^2}+\sqrt{I_p^2+Q_P^2}}$$

This phase detection algorithm has a large computing and needs two pairs of correlators, but it has a small errors.

1.5 Experiment Results

Based on the design method of the loop, we developed a high-dynamic GPS receiver principle prototype as Fig. 1.6 shows.

The experiments include static and kinematic. Figure 1.7 shows the non-coherent integration amplitude of the promote branch and the promote branch while searching the seventh satellite while the receiver is in static state. While the non-coherent integration amplitude of the two branches exceeds the preset threshold, the system considers that the coarse acquisition is successful, and signals may exist in the current unit.

Because the vehicle-mounted experiment is unable to get an enough high speed, so here we used the high dynamic satellite signal simulator to test the tracking performance of the tracking loop. When the FLL bandwidth is set to 18 Hz, the PLL bandwidth is set to 10 Hz, the receiver principle prototype can track 100 g/s high dynamic signal. The test results show that the receiver principle prototype have achieved the desired indicators, it can search the satellites signal and Complete Satellite Positioning under the high dynamic environment.

Table 1.2 The Loop filter characteristics

Loop order	BW of noise	Filter parameters	Steady state error	Characteristics
Second order	$\frac{\omega_0(1+a_2^2)}{4a_2}$	$a_2\omega_{nf} = 1.414\omega_{nf}$ $B_n = 0.53\omega_{nf}$	$\frac{d^2R/dt^2}{\omega_{nf}^2}$	Sensitive to stress of acceleration. Unconditionally stable for all the noise bandwidth
Third order	$\frac{\omega_0(a_3b_3^2+a_3^2-b_3)}{4(a_3b_3-1)}$	$a_3\omega_{np}^2 = 1.1\omega_{np}^2$ $b_3\omega_{np} = 2.4\omega_{np}$ $B_n = 0.7845\omega_{np}$	$\frac{d^3R/dt^3}{\omega_{np}^3}$	Sensitive stress jerk. When the bandwidth is less than or equal to 18 Hz it can remain stable

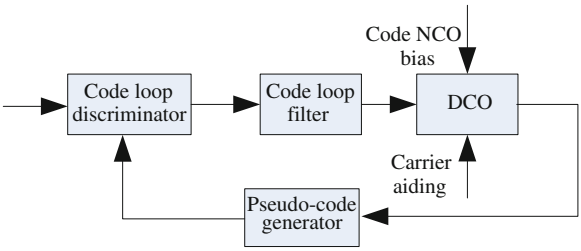


Fig. 1.5 Code loop structure

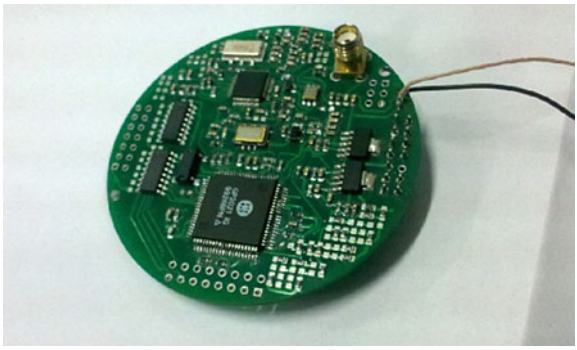


Fig. 1.6 High dynamic satellite positioning receivers principle prototype

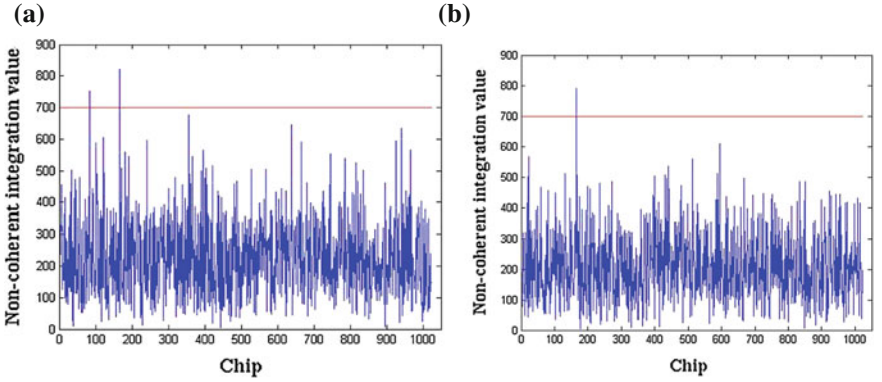


Fig. 1.7 Non-coherent integration values of the No. 7 satellite (Band 0) **a** the early branch **b** the promote branch

1.6 Conclusion

This paper proposed a novel design program of high dynamic GPS receiver, whose phase-locked loop bandwidth is 10 Hz, frequency-locked loop bandwidth 18 Hz, and code loop bandwidth 2 Hz. The experiment shows that the receiver prototype could satisfy the error limits of the tracking loop, and it also can capture the visible satellite signals with a good effect of real-time tracking lock.

References

1. Tian, M., Shao, D., Cheng, N., Xue, W.: A scheme for high dynamic GPS receiver. Telemetry & Telecontrol **23**(2), 15–20 (2002)
2. Ma, X., Liu, B.: Design of an INS aided high dynamic GPS receiver. Electronics, Communication and Control (ICECC), pp. 1404–1407. Ningbo, China (2011)
3. Li, J., Ba, X., Chen, J.: New high dynamic GPS receiver carrier tracking algorithm based on IMM. J. Syst. Simul. **20**(9), 2483–2486 (2008)
4. Xin, Z., Li, J., Liang, H., Wang, F., Yan, Y.: Design of a GPS receiver based on ARM and FPGA. Transducer Microsyst. Technol. **30**(7):108–110 (2011)
5. Xie, G.: Principle of GPS Receiver Design. pp. 78–85, Electronic Industry Press, Beijing (2009)
6. Kaplan, E.: Understanding GPS: Principles and Applications. pp. 180–181, Artech House, Inc., Norwood (2006)
7. Parkinson, B.: Global Positioning System: Theory and Applications. pp. 245–325, American Institute of Aeronautics and Astronautics, Reston (1996)

Chapter 2
The Modeling of Grid Workflow Service Based on GSCPN

Yaqian Yang, Jun Zheng and Wenxin Hu

Abstract In order to evaluate and optimize the performance of service composition, a formal description of Grid workflow activity and a model of grid workflow service based on Generalized Stochastic Colored Petri Net are presented in this paper. The token of color of GSCPN can be used to stimulate the different events of message types of business processes. Moreover, Grid workflow service composition algorithm and reduction algorithm based on GSCPN are proposed. Empirical results show that the reduced model reduces the complexity of the qualitative and quantitative analysis of the generated grid workflow service model. The grid workflow service model based on GSCPN can be used for performance prediction and to guide the optimization.

Keywords Generalized Stochastic Petri net · Grid workflow service · Service composition and reduction · Modeling

2.1 Introduction

Service-based application systems are often not based on a simple service, and a service is not based on an activity. As a qualitative and quantitative analysis of formal method, GSCPN model has advantages in the description of the service composition and its interactive behavior. Therefore, for scheduling and integration

Y. Yang (✉) · J. Zheng · W. Hu
Computer Center, East China Normal University, Shanghai, China
e-mail: yangyq1988@hotmail.com

J. Zheng
e-mail: jzheng@cc.ecnu.edu.cn

W. Hu
e-mail: wxhu@cc.ecnu.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_2,

of these services, this paper presents a service modeling algorithm based on GSCPN for performance prediction and guiding optimization, using which it's able to model both service and service composition, and analysis some attributes.

Workflow management systems and workflow modeling and processing techniques have been studied and used for complex business processes [1]. Many related works have been reported in the modeling of web service and grid service composition based on Petri nets. Han et al. model web service using Petri nets, and give the algorithm of combination and simplification of the model, which is analyzed through chart [2]. Yang et al. presents the model of web service and web service composition operation based on generalized stochastic colored Petri net (GSCPN), realizes QoS (Quality of Service) and data representations [3]. Xiong et al. discusses the model of grid workflow service composition using colored Petri nets and examples are described [4]. But those documents are lack of descriptions of details about modeling.

The definition of GSCPN and the activity model based on GSCPN are presented in Sect. 2.2. Section 2.3 provides an algorithm for service modeling and Sect. 2.4 gives an algorithm for reducing service model. Section 2.5 presents an example of grid workflow service composition and Sect. 2.6 provides some concluding remarks.

2.2 Basic Activity Based on GSCPN

Activities constitute a service and the activity is modeled using GSCPN. This model concludes an input token and an output token for interface information storage and some transitions which present the internal logic of GSCPN.

Definition 1 Generalized Stochastic Colored Petri net is a 9-tuple GSCPN $= (\sum$, P, T, F, C, G, E, λ, I), where:

(1) $\sum = \{ps\}$ is a set of colors, where ps represents a parameter of atomic activity;
(2) P = {pi, po} is a set of tokens, where pi represents the input of atomic activity and po represents its output;
(3) $T = T_i \cup T_t$ is a set of transitions, where T_i represents immediate transition in which the implementation time can be negligible, T_t represents time transition in which the implementation requires a certain time and timed transition obeys negative exponential distribution;
(4) F = {(pi,T), (T,po)} is a set of arcs from P to T and T to P, where input arcs represent arcs from P to T, and output arcs represent arcs from T to P;
(5) C = {C (pi), C (po)} is a set of color functions representing a mapping from P to$\sum$;
(6) G is a set of guard functions, G:T → BoolExpression, $\forall_t \in$ T:Type (G(t)) = Boolean ∧ Type (G(t)) = Boolean∧ type $(\text{var}(G(t))) \subseteq \sum$, where var (G(t)) represents the variable set of G(t);

(7) $E = \{E(pi, T), E(T, po)\}$ is a set of arc expression functions, $E:F \rightarrow$ Expression, $\forall f \in F:Type(E(f)) = C(p(f))_{MS} \wedge Type(var(E(f))) \subseteq \sum$, where p(f) is the tokens associated with f, and $C(p(f))_{MS}$ represents the multi-set of C(p);
(8) λ is a priority collection in the average implementation rates of timed transition or instantaneous conflict transition, where timed transition obeys negative exponential distribution;
(9) I is a set of initialization functions, $I: P \rightarrow \sum$, which assign initial color for each token, generating initial marking M_0.

Definition 2 An atomic Grid Workflow activity is a tuple, $A = (\sum, P, T, F, C, G, E, \prod, I)$(see Fig. 2.1), where $\sum$, P, T, F, C, G, E, $\prod$, I have the same definition as definition 1. $pi \in P$ represents the input token, of which the set of predecessor nodes is empty, and $po \in P$ represents the output token, of which the set of successor nodes is empty.

Figure 2.1 shows a model of an atomic Grid Workflow activity. This paper assumes that the transition T will occur when pi conditions are met, and generating po.

2.3 Service Modelling Composition Algorithm Based on GSCPN

Many simple activities combine to a service having some certain function. Atomic activities constitute a service through sequence, concurrent, selection and iteration operation, according to which, the activities composition algorithm of grid service is given. The composition and reduction of activities is shown in Fig. 2.2.

Algorithm 1 Construction of GSCPN model for an service composed of a collection of activities:

Input: Activity A_1, A_2, A_3… A_n, including their parameters and predecessor and successor activities.

Output: GSCPN model for a service composed of a collection of atomic activities.

(1) $A = \{A_1(i, 2), A_2(1, 3), \ldots, An(m, o)\}$, $P = \{p_i, p_o\}$, $T = \{t_i\}$, $E = \{E(p_o, t_i), E(t_i, p_i)\}$, where i and o represent the input and output place, GSCPN = Φ.
(2) For each activity $A_i(h, j)$ Do

IF $\exists_t \in T$, $E(t, p_i) \neq \phi$, THEN activity A_i is iterative, which performs for n times(shown in Fig. 2.3), this iterative composition of activities is modified by

Fig. 2.1 Atomic Grid Workflow activity A

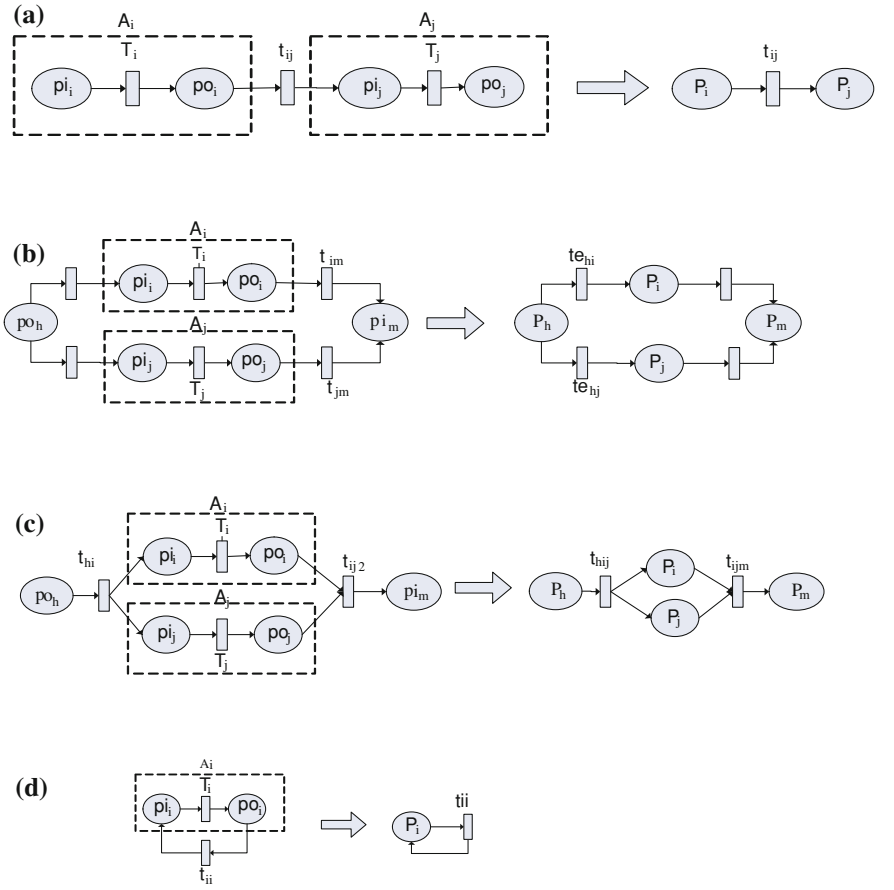


Fig. 2.2 Four kinds of operations of the composition and reduction of activities **a** Sequence composition and reduction of activity A_i and A_j **b** selection composition and reduction of activity A_i and A_j **c** concurrent composition and reduction of activity A_i and A_j **d** iteration composition and reduction of activity A_i

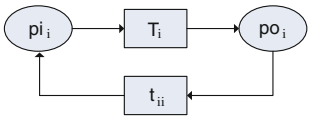


Fig. 2.3 Iteration of activity A_i based on GSCPN for n times

following rules: A_i: $\sum = \sum_i$, $P_i = P_i$, $T_i = T_i \cup \{t_{ii}\}$, $F_i = F_i \cup \{(po_i, t_{ii}), (t_{ii}, pi_i)\}$, $C = C_i$, $E_i = E_i \cup \{E(po_i, t_{ii}), E(t_{ii}, pi_i)\}$, $G_i = G_i \cup \{G(t_{ii})\}$, $I_i = I(pi_i)$, $\prod_i = \prod_i \lambda_i = \lambda_i$.

II IF $i = n + 1$, then go to (3), else $i + = 1$, A_i (h, j), then go to III.

III Activity A_h and A_i compose by following rules as shown in Fig. 2.4, between two activities add an immediate transition t_{hi} and two arcs, of which one is from po_h to t_{hi}, the other is from t_{hi} to pi_i: IF j >= i THEN $\sum_{=} \sum_{h} \cup \sum_{i,P=P_h} \cup P_i$, $T = T_h \cup T_i \cup \{t_{hi}\} \cup \{t_{ij}\}$, $F = F_h \cup F_i \cup \{(po_h, t_{hi}), (t_{hi}, pi_i)\}$, $C = C_h \cup C_i$, $E = E_h \cup E_i \cup \{E(po_h, t_{hi}), E(t_{hi}, pi_i)\}$, $G = G_h \cup G_i \cup \{G(t_i)\}$, $I = I_h \cup I_i$, $\prod = \prod_h \cup \prod_i$, $\lambda = \lambda \cup \lambda_i$, ELSE IF j < i THEN $\sum_{=} \sum_{h} \cup \sum_{i,P=P_h} \cup P_i$, $T = T_h \cup T_i \cup \{t_{hi}\} \cup \{t_{ij}\}$, $F = F_h \cup F_i \cup \{(po_h, t_{hi}), (t_{hi}, pi_i), (po_i, t_{ij}), (t_{ij}, pi_j)\}$, $C = C_h \cup C_i$, $E = E_h \cup E_i \cup \{E(po_h, t_{hi}), E(t_{hi}, pi_i), E(po_i, t_{ij}), E(t_{ij}, pi_j)\}$, $G = G_h \cup G_i \cup \{G(t_i)\}$, $I = I_h \cup I_i$, $\prod = \prod_h \cup \prod_i$, $\lambda = \lambda \cup \lambda_i$, return II.

(3) Find out all output tokens po_i which have two or more successor transition nodes, in other words, from which there are two or more arcs, and also the collection of these arcs as $\{(po_i, t_{ij})\}$. IF the collection is empty, THEN return (4), ELSE by successively traversing, for the C (po) of related set of activities $\{A_j\}$ to every element (po_i, t_{ij}) in the collection do pairwise intersect operation.

IF $C(po_{j1}) \cap C(po_{j2}) \cap \ldots \cap C(po_{jm}) = \varnothing$, and the intersection of the successor activities collection of A_{j1}, $A_{j2} \ldots A_{jm}$ is not empty THEN get the first element a_p, the collection of transitions pointed to this element as $\{t_{rp}\}(r = r1, r2 \ldots rn)$ and the intersection of the successor activities collection of S_{j1}, $S_{j2} \ldots S_{jm}$ is empty, where S_{jk} is a subset of A_{jk} and a set of activities from activity a_{j1} to a_p, THEN A_{j1}, $A_{j2} \ldots A_{jm}$ is concurrent. Activities A_{j1}, $A_{j2} \ldots A_{jm}$ are composed by following rules: delete transitions t_{ik} (k = j1, j2,…,jm) and t_{rp}(r = r1, r2…rn) and add transitions t_{ij1jm} and t_{r1rnp}, where t_{ij1jm} produces proper input for activities A_{c1}, $A_{c2} \ldots A_{cn}$ and t_{r1rnp} realize omposes the output of activities A_{c1}, $A_{c2} \ldots A_{cn}$ as the output of the activities composition. Practically, (k = c1, c2,…,cn,r = r1,r2…rn) (shown in Fig. 2.5):

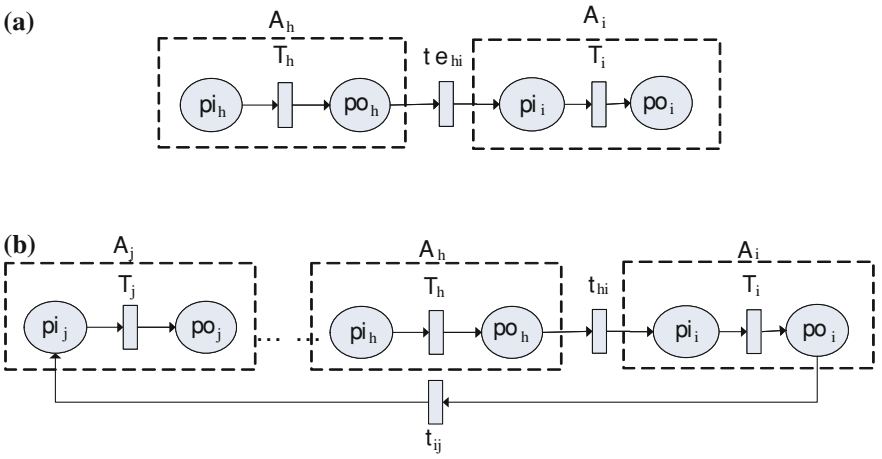


Fig. 2.4 The composition of activity A_h, A_i and A_j based on GSCPN **a** If j > i, the composition of activity A_h, A_i and A_j based on GSCPN **b** If j < i, the composition of activity A_h, A_i and A_j based on GSCPN

$T = (T-\{t_{ik}\}-\{t_{rp}\})\cup\{t_{ij1jm}\}\cup\{t_{r1rnp}\}, F = (F-\{(po_i, t_{ik}),(t_{ik}, pi_k)\} -\{(po_r, t_{rp}), (t_{rp}, pi_p)\})\cup \{(po_i, t_{ij1jm})\}\cup\{(t_{ij1jm}, pi_k)\}\cup\{(po_r, t_{r1rnp})\}\cup\{(t_{r1rnp}, pi_p)\}$, $E = (E-\{E(po_i, t_{ik}), E(t_{ik}, pi_k)\} -\{E(po_r, t_{rp}), E(t_{rp}, pi_p)\})\cup\{E(po_i, t_{ij1jm})\}\cup\{E(t_{ij1jm}, pi_k)\}\{E(po_r, t_{r1rnp})\}\cup\{E(t_{r1rnp}, pi_p)\}$, $G = (G-\{G(t_{ik})\}-\{G(t_{rp})\})\cup\{G(t_{ij1jm})\cup\{G(t_{r1rnp})\}$;

(4) The algorithm ends, generating GSCPN model of the service.

2.4 Service Modelling Reduction Algorithm Based on GSCPN

To reduce the complexity of qualitative and quantitative analysis of the service model and retain the previous relations between activities, the service model generated from algorithm 1 will be reduced by following algorithm. Replace an atomic activity model with a node to reduce the number of nodes (Shown in Fig. 2.2).

Algorithm 2: (1) Replace every activity A_i with a node P_i, Concretely, $P = (P - P_i) \cup \{P_i\} F = (F - \{(pi_i, T_i), (T_i, po_i)\}), E = (E - \{E(pi_i, T_i), (T_i, po_i)\}, T = (T - \{T_i\}),\ i = 1, 2...n.$

(2) Replace pi_i and po_i in the relevant collection of arcs and functions with P_i, Concretely, $F = (F - \{(po_i, t_{ij}), (t_{ij}, pi_j)\})\cup\{(P_i, t_{ij})\}\cup\{(t_{ij}, P_j)\}$, $E = (E-\{E(po_i, t_{ij}), E(t_{ij}, pi_j)\}) \cup\{E(P_i, t_{ij})\}\cup\{E(t_{ij}, P_j)\}$.

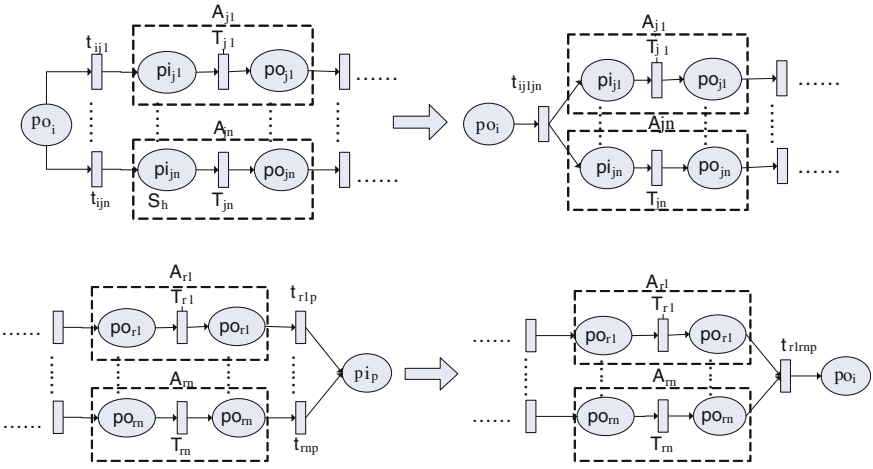


Fig. 2.5 The parallel operation of the composition of activities

2.5 Case Study

2.5.1 Modeling

This paper will build GSCPN model for Business Travel, a business process of the example, Travel. bpel given by [5].

After receiving a travel plan from the client, Business Travel calls Employee Travel Status activity to check job level of the employee and select corresponding cabin class. Then parallel call American Airlines activity and Delta Airlines activity to check the price of two airline tickets and its availability and choose the lower-priced airline. Finally, inform client of the result of travel plan consultation about the available ticket or not. According to above algorithms, the model will be given as follows (Shown in Fig. 2.6).

2.5.2 Analysis

Yang et al. have analyzed and proved the arithmetic properties [3]. The collection of service operations is closed. In other words, a service model starts from the input, and are always able to produce output token mark.

The performance analysis of GSCPN model is usually completed by constructing isomorphic Markov chain to the model, depending on the meaning of expressed in the model, analyze the model of survival, security and make performance prediction, thereby find the performance bottleneck of the portfolio of composite service [6, 7, 8]. The steps are as follows.

$U = F + EG^{\infty,}$ where U represents the transition probability matrix of the reduction Markov chain, F represents the matrix from tangible state to tangible state, E represents the matrix from tangible state to vanishing state, and $g_{ij} = P(r \rightarrow j)$ represents the probability from given vanishing state r to the first accessible tangible state j. A tangible state is a set of states merely viable for time transitions, and a vanishing state is a set of states accessible for immediate transitions.

$Y = YU$, where Y is a unit row vector, and represents steady state probability.

Set a state i as a reference, then the times for visiting the state j during continuous visiting the state i, $V_{ij} = Y_j/Y_i$. Average residence time ST_i, is zero if state i is a vanishing state and is $[\sum(\lambda_k)]^{-1}$, $T_k \in E(M_i)$, which is a set of viable transitions in the state of i, if state i is a tangible state. Then the average cycle time back to i, $W_i = \sum V_{ij}STj$ and the steady state probability Pj is 0 if state j is a vanishing state and is $V_{ij}STj/W_i$ if state j is a tangible state.

Analyze above model by the above steps under the condition given by [9], and then find the same bottleneck activity, invoking and receiving airlines, after obtaining the average excitation time and the timed transition utilization.

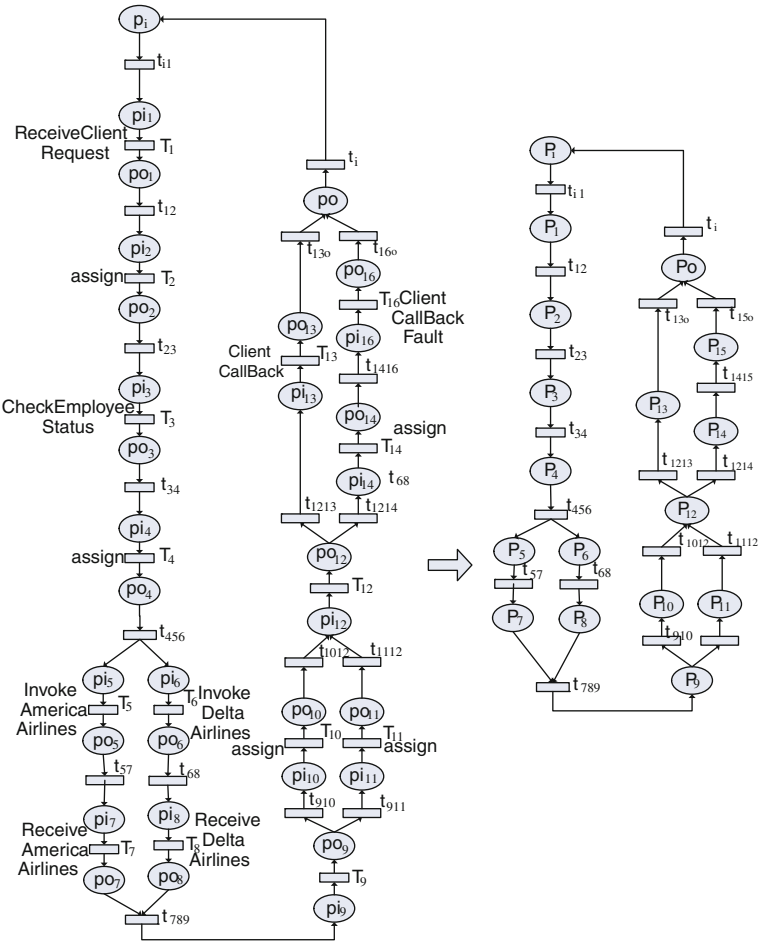


Fig. 2.6 Modeling of business travel service based on GSCPN

2.6 Conclusion

For formal verification and qualitative and quantitative analysis of grid services composite, this paper presents the composition and reduction algorithms of service model based on GSCPN, which enable it to quickly get its service model for complex service. Since reduced model reduces almost half of the nodes, it simplifies the analysis, and provides preconditions for making performance prediction before releasing the composite service and finding performance bottlenecks.

References

1. Pinar, S.: Composite web service construction by using a logical formalism. In: Proceedings of International Conference on Data Engineering Workshops, pp. 331–336 (2006)
2. Han, Y., Luo X.: Composition and reduction of web service based on dynamic timed colored petri nets. International Parallel and Distributed Processing Symposium, pp. 659–663 (2009)
3. Yang, N., Yu, H., Guo, X.: Web service composition model based on generalized stochastic colored petri nets. Comput. Sci. **39**(4), 142–144 (2012)
4. Xiong, Z., Zhai, Z., et al.: Grid workflow service composition based on colored petri net. Int. J. Digit. Content Technol. Appl. **5**(5), 125–131 (2011)
5. Matjaz, J., et al.: Business Process Execution Language for Web Services Second Edition. Packt Publishing, Birmingham (2006)
6. Men, P., Duan, Z.: Modeling and evaluation of composite web services based on generalized stochastic petri net. J. Xi'an Jiaotong Univ. **42**(8), 967–971 (2008)
7. Lin, C.: Stochastic Petri Nets and System Performance Evaluation. Tsinghua University Press, Beijing (2005)
8. Chi, W., Zhu, Y.: Modeling and evaluation of web service composition based on GSPN. Microprocessors **5**, 26–29 (2011)
9. Zhu, J., Guo, C., Wu, Q.: GSPN based web service composition performance prediction model. Computer Science **38**(8), 125–129 (2011)

Chapter 3
A Voice Conversion Method Based on the Separation of Speaker-Specific Characteristics

Zhen Ma, Xiongwei Zhang and Jibin Yang

Abstract This paper aims to study independent and complete characterization of speaker-specific voice characteristics. Thus, the authors conduct a method on the separation between voice characteristics and linguistic content in speech and carry out voice conversion from the point of information separation. In this paper, authors take full account of the K-means singular value decomposition (K-SVD) algorithm which can train the dictionary to contain the personal characteristics and inter-frame correlation of voice. With this feature, the dictionary which contains the personal characteristics is extracted from training data through the K-SVD algorithm. Then the authors use the trained dictionary and other content information to reconstruct the target speech. Compared to traditional methods, the personal characteristics can be better preserved based on the proposed method through the sparse nature of voice and can easily solve the problems encountered in feature mapping methods and the voice conversion improvements are to be expected. Experimental results using objective evaluations show that the proposed method outperforms the Gaussian Mixture Model and Artificial Neural Network based methods in the view of both speech quality and conversion similarity to the target.

Keywords Voice conversion · Speaker-specific characteristics · Information separation · K-SVD

Z. Ma (✉) · X. Zhang · J. Yang
Institute of Communication Engineering, PLA University of Science and Technology, Nanjing, China
e-mail: mazhen1989@126.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_3,

3.1 Introduction

Voice conversion systems modify a speaker's voice to be perceived as uttered by another speaker [1]. Voice conversion technology can be applied to many areas. Definitely, one of the most important applications is the use in the area of text-to-speech synthesis (TTS) where it hopes to synthesize different speaking style voices without a large utterance corpus. Besides speech synthesis, however, voice conversion has other potential applications in areas like entertainment, security and speaker individuality for interpreting telephony and voice restoration.

Many conversion methods have been proposed since the problem was first formulated in 1988 by Abe et al. [2]. Generally speaking, the main methods such as Gaussian Mixture Model (GMM) [3], Hidden Markov Model (HMM) [4], Frequency Warping (FW) [5] and Artificial Neural Network (ANN) [6] based methods has gained similarity in converted speech together with quality declines dramatically. Though many improved methods like bring Global Variance and Parameter Trajectory in GMM based method has been put forward over the last few years, and all upgraded in quality to some extent, the conversion results is not satisfactory for practical application. Therefore, we need to pursue new conversion methods to achieving high-quality converted speech.

It's clear that the information separation of speech signals is going to be a new approach for speech processing. Among the information, the meaning of message being uttered is of prime importance, and the information of speaker identity also plays an important role in oral communication. We define these two types of information as content and style of speech respectively. General knowledge tells us that human auditory perception system has the ability to identify the meaning of message and the speaker identity simultaneously, i.e. to separate content and style factors sophisticatedly from speech signals.

In Popa Victor's work, the bilinear model was used to model speech, and two sets of speech parameters, indicating style and content respectively, are achieved by singular value decomposition (SVD) on speech observations [7]. With the separation result, voice conversion is realized by replacing the source style with the target one, while preserving the initial content, i.e. the converted speech is reproduced based on source content and target style. Based on this literature, in this paper we use a new method to separate the information of speaker identity.

K-means singular value decomposition (K-SVD) is a signal decomposition method aims at sparse representation of signals. This method has been successfully used in image denoising, character extracting and so on. Because of the sparsity nature and other characteristics, we can use K-SVD to decomposing the vocal tract spectrum into a dictionary which conveys the personal identity and corresponding sparse matrix which contains the content information. Taking this into account, we can use the dictionary to achieve the separation of speaker-specific characteristics and speech conversion based on substitution of the dictionary.

Combining the identity of speech with K-SVD, we implement voice conversion using style replacing technique as proposed in Popa Victor's work. Experimental

results show that the proposed voice conversion approach outperforms the traditional GMM and ANN method in terms of both similarity and naturalness, especially in the case of small size of training dataset.

The paper is organized as follows. In the next section, we briefly introduce the K-SVD theory. In Sect. 3.3, we describe the voice conversion scheme based on K-SVD. Next, we make experiments to evaluate the proposed method, and the results demonstrate the benefits of K-SVD comparing with GMM and ANN in voice conversion. Finally, we make remarks on the proposed method, and some potential future research directions are also presented in section V.

3.2 The K-SVD Algorithm

We have witnessed a growing interest in the use of sparse representations for signals in recent years. Using an overcomplete dictionary matrix $D \in IR^{n\times K}$ that contains K atoms, $\{d_j\}_{j=1}^{K}$, as its columns, it is assumed that a signal $Y \in IR^n (n \ll K)$ can be represented as a sparse linear combination of these atoms. The representation of Y may be approximate, $Y \approx DX$, satisfying $||Y - DX||_2 \leq \varepsilon$. The vector $X \in IR^K$ contains the representation coefficients of the signal Y. This sparsest representation is

$$(P_{0,\varepsilon}) \ \min_x ||x||_0 \text{ subject to } ||Y - DX||_2 \leq \varepsilon \tag{3.1}$$

In the K-SVD algorithm [8] we solve (1) iteratively, using two stages, parallel to those in K-Means. In the sparse coding stage, we compute the coefficients matrix X, using any pursuit method, and allowing each coefficient vector to have no more than T_0 non-zero elements. Then, we update each dictionary element sequentially, changing its content, and the values of its coefficients, to better represent the signals that use it. This is markedly different from the K-Means generalizations that were proposed previously, e.g., since these methods freeze X while finding a better D, while we change the columns of D sequentially, and allow changing the relevant coefficients as well. This difference results in a Gauss–Seidel-like acceleration, since the subsequent columns to consider for updating are based on more relevant coefficients. We also note that the computational complexity of the K-SVD is equal to the previously reported algorithms for this task.

We now describe the process of updating each atom d_k and its corresponding coefficients, which are located in the k-th row of the coefficient matrix X, denoted as x_k. We first find the matrix of residuals, and restrict this matrix only to the columns that correspond to the signals that initially use the currently improved atom. Let r be the set of indices of these signals, similarly, denote E_k^r as the restricted residual matrix, which we would now like to approximate using a multiplication of the two updated vectors d_k and x_k, i.e. The equation converted to:

$$||Y - DX||_F^2 = ||y - \sum_{j=1}^{N} d_j x_j^r|| = ||(y - d_j x_j^r) - d_k x_k^r||_F^2 = ||E_k - d_k x_k^r||_F^2 \quad (3.2)$$

We seek for a rank-one approximation. Clearly, this approximation is based on the singular value decomposition (SVD), taking the first left and right singular vectors, together with the first singular value. So we can get the sparse representation of speech signals as well as the dictionary.

We can suppose that the personal characteristics are contained in the dictionary while the content is conveyed in coding matrix due to the dictionary open out the main identity and inner structure of speech. And for this hypothesis, we will demonstrate it at the experiment phase.

3.3 Voice Conversion Based On K-SVD

So far, we know that a speech signal can be decomposed to a dictionary conveys the information related to personal identity and a corresponding sparse matrix contains the information related to content. Naturally we can replace the source speaker's dictionary by the target speaker's dictionary and then multiply with the source sparse matrix to implement voice conversion. While this scheme is easy to realize, there is a cute problem that the result trained by K-SVD is of multiplicity which means that a vocal track spectral matrix may have different combination of dictionary and sparse matrix. To settle this problem, we excerpt the means put forward by XU [9] to introduce partial limitation for the vocal track spectral conversion which is shown in Fig. 3.1.

There are two basic modes in the voice conversion system. In training mode, we process the data of source and target speaker (part 1 in Fig. 3.1) and extract the dictionary based on K-SVD (part 2 in Fig. 3.1). In converting mode, we implement voice conversion based on dictionary replacement (part 3 in Fig. 3.1). Next, we will introduce the process of part 1, 2 and 3 concretely.

3.3.1 Data Processing

The source and target speech to be trained are denoted as x and y, in which the contents are the same. As for the extraction of vocal tract spectral, we take use of STRAIGHT model which can separate the excitation to vocal tract better to analyze the speech. The trained speech data analyzed by the STRAIGHT model and we get the STRAIGHT spectrum of source and target speech are denoted as S_x and S_y.

Before training, the S_x and S_y must be aligned in time to ensure the coherence of the coding matrix decomposed at the next step. While the common method

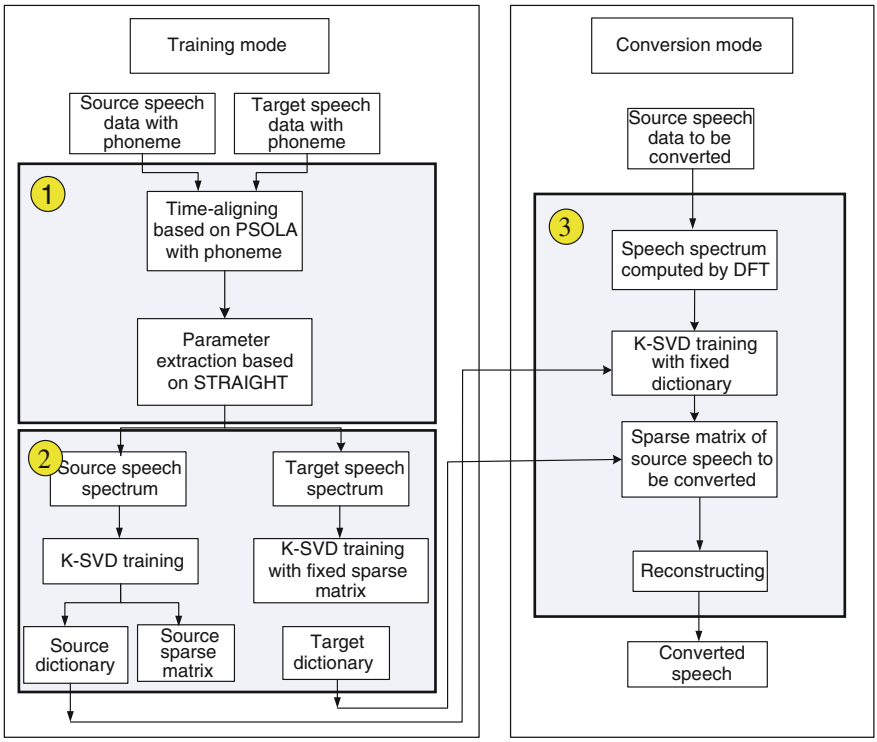


Fig. 3.1 Voice conversion system based on K-SVD

Dynamic Time Warping (DTW) [2] may lead to the quality declines dramatically because of having not considered the continuity of the inter-frame when inserting or removing the sequential frames, we adopt the time-aligned method with phoneme label information put forward in [10]. $S_x^{'}$ and $S_y^{'}$ denote time-aligned source and target STRAIGHT spectrum respectively.

3.3.2 Dictionary Extraction

So far, we have got the time-aligned STRAIGHT spectrum, and then we will extract the dictionary used for conversion.

As illustrated in part 2 of Fig. 3.1, using K-SVD to decomposing $S_x^{'}$, we can get the dictionary D_x and sparse matrix H_x of source speech. If the target speaker's STRAIGHT spectrum is analyzed in the same way, it is difficult to ensure the coherence of the dictionary of the target with the source. Considering the hypothesis that the sparse matrix is determined by the content information, it is clearly that the sparse matrix H_x of $S_y^{'}$ is the same with H_x. Therefore, set $H_y = H_x$

when decomposing $S_y^{'}$ to get corresponding dictionary of the target speaker denoted as D_y.

3.3.3 Voice Conversion Based on Dictionary Replacement

As shown in part 3 of Fig. 3.1, in converting mode we should first extract the STRAIGHT spectrum of the source speech. And then we shall use OMP algorithm to decompose this spectrum with D_x fixed to get the corresponding sparse matrix $H_x^{convert}$. So we can achieve the converted STRAIGHT spectrum based on $H_x^{convert}$ and the dictionary D_y accepted at the training mode according to the Eq. (3.3).

$$S_y^{convert} = D_y \cdot H_x^{convert} \tag{3.3}$$

3.4 Experimental Evaluations

3.4.1 Experimental Conditions

We conducted two experimental evaluations. Firstly, in order to demonstrate the dictionary conveys the speaker-specific characteristics and the sparse matrix conveys the content information, we use K-SVD to separate the speaker-specific characteristics with content information. Secondly, in order to demonstrate the effectiveness of the proposed conversion method, we compared the method against the GMM and ANN based methods, which are the most popular conversion techniques in the past few years.

In our experiments, all the data are selected from the CMU ARCTIC databases. It consists of four different database named SLT, BDL, JMK and AWB uttered by four different speakers.

The two parameters may affect the conversion result for the K-SVD; they are the numbers of atoms (K) in the dictionary and the sparsity (t) of the sparse matrix. It is clear that the reconstruction error can be decreased with the increasing of the two parameters. However, the result is not always better as the parameters increase too much. We have found that when the two parameters fill Eq. (3.4)

$$K = 40,\ t = 12 \tag{3.4}$$

The converted voice is the best in quality as well as similarity through many experiments. Therefore, we set the two parameters as Eq. (3.4) at the next experiments.

3.4.2 Speaker-Specific Characteristics Separation

To demonstrating the speaker-specific characteristics are included in the dictionary, we do the work as follow.

Firstly, we select the preceding 10 sentences in BDL and SLT and compute the STRAIGHT spectrum respectively, the dimensionality of each row vector is 256. Secondly, getting the dictionary D_{BDL} and D_{SLT} through K-SVD. Then, we choose the succedent 5 sentences in BDL and SLT to calculate the STRAIGHT spectrum. At last, computing the reconstructed error with the dictionary settled as D_{BDL} and D_{SLT}. The reconstructed error is calculated using,

$$e = 10\log_{10}\left(\frac{1}{N_S}\sum_{i=1}^{N_S}\|\mathbf{s}_i - \hat{\mathbf{s}}_i\|^2 \Big/ \frac{1}{N_S}\sum_{t=1}^{N_S}\|\mathbf{s}_i\|^2\right) \tag{3.5}$$

where N_S denotes the number of rows in STRAIGHT spectrum, $\mathbf{s}_i$ denotes the i-th row of the source spectrum, $\hat{\mathbf{s}}_i$ denotes the i-th row of the reconstructed spectrum. The results are summarized in Table 3.1.

It is clear that the reconstructed error of dictionary suited to the voice is much less than the other when analyzing by K-SVD. Therefore, we can assure that the dictionary trained by K-SVD is certain to containing the speaker-specific characteristics, and we can separate the speaker-specific characteristics effectively. And according to this, we can implement voice conversion by replacing the dictionary.

3.4.3 Effectiveness of the Proposed Method

In Fig. 3.2, we can see the source, target and converted speech directly. To contrasting the converted result, two subjective listening tests were carried out. One test is Mean Opinion Score (MOS), which aims to give a score between 1 and 5 to evaluate speech naturalness, and the other ABX which uses the correct rate of speaker recognition to evaluate the successfulness of individuality conversion. All tests were performed by five listeners who have been engaged on speech processing research for many years. The results are given in Table 3.2.

From the experimental results shown in Table 3.2, it is easy to see that the proposed method for voice conversion based on style and content separation clearly outperform traditional GMM and ANN algorithm based on parameter

Table 3.1 Results of reconstructed error (db)

	D_{BDL}	D_{SLT}
BDL	−18.02	−15.65
SLT	−13.23	−19.21

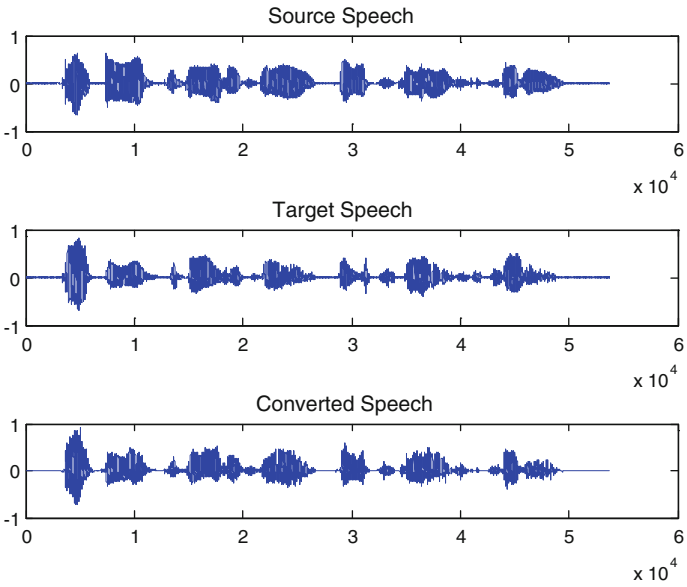


Fig. 3.2 Source, target and converted speech

Table 3.2 Results of subjective listening tests

Subjective test	Voice conversion methods		
	Proposed method	GMM	ANN
Naturalness (MOS)	2.96	2.81	2.77
Individuality (ABX)	0.88	0.74	0.80

extraction and mapping. So we can make conclusion that the method would work with only small size of training data, while GMM and ANN methods does not because of unreliable estimation of GMM parameters due to insufficient data.

The result also shows that although the proposed method is able to implement voice conversion in the framework of style and content separation, and has superior performance. We may conclude that K-SVD has the advantage of preserving the speaker-specific characteristics and taking the inter-frame correlation into account for separation and conversion.

3.5 Conclusion

This paper presents a voice conversion method based on style and content separation, which is solved by K-SVD. And on the basis of successful separating speaker-specific characteristics, the authors invent a novel method for voice

conversion using style replacement. And they also have solved the multiplicity of the decomposition result of K-SVD in the conversion method. Experimental results show that the proposed method for voice conversion has superior performance than traditional GMM and ANN based algorithm in view of both speech quality and conversion similarity and naturalness.

The authors present a novel way to realize voice conversion and improve the conversion performance dramatically which is still a linear combination of style and content while the complex relationship between observed data and style and content is not able to be fully described only by linearity. Therefore, in future studies, researchers will extend the K-SVD to non-linearity to improve the performance of speech style and content separation technique. In addition, the style and content separation technique will immensely promote the development of technique in other speech signal processing domain, such as speech recognition, speaker identification, and low-rate speech coding, etc.

References

1. Stylianou, Y.: Voice transformation: a survey. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3585–3588 (2009)
2. Abe, M., Nakamura, S., Shikano, K., Kuwabara, H.: Voice conversion through vector quantization. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 655–658 (1988)
3. Stylianou, Y., Cappe, O., Moulines, E.: Continuous probabilistic transform for voice conversion. IEEE Trans. Speech Audio Process. **6**(2), 131–142 (1998)
4. Yamagishi, J., Kobayashi, T., Nakano, Y., et al.: Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans. Audio Speech Lang. Process. **17**(1), 66–83 (2009)
5. Erro, D., Moreno, A., Bonafonte, A.: Voice conversion based on weighted frequency warping. IEEE Trans. Audio Speech Lang. Process. **18**(5), 922–931 (2010)
6. Desai, S., Black, A.W., Yegnanarayana, B., et al.: Spectral mapping using artificial neural networks for voice conversion. IEEE Trans. Audio Speech Lang. Process. **18**(5), 954–964 (2010)
7. Popa, V., Nurminen, J., Gabbouj M.: A Novel Technique for Voice Conversion Based on Style and Content Decomposition with Bilinear Models. Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009), pp. 6–10. Brighton, U.K. (2009)
8. Michal, A., Michael, E., Alfred, B.: K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process. **54**(11), 4311–4322 (2006)
9. Xu, N., Yang, Z., Zhang, L.H., et al.: Voice conversion based on state-space model for modelling spectral trajectory. Electron. Lett. **45**(14), 763–764 (2009)
10. Jian, S., Xiongwei, Z., Tieyong, C. et al. Voice conversion based on convolutive non negative matrix factorization. Data Collect. Process. **28**(3), 285–390 (2012)

Chapter 4
The Application of Information Security on the Computer Terminals of 3rd-Generation Nuclear Power Plant

Zhiping Song and Yi Luo

Abstract The paper aims to develop a computer terminal solution which suits with the status of the company to eliminate risks of information security caused by the huge amount of computer terminals. The research carries out analysis on Intranet access, standard configuration, information data management, etc. through introduction of standards fit with national information security requirements. To make the analysis smooth, the author uses a real example of computer terminal configuration in a nuclear power plant. Results of implementing the solution in the real example show that security risk of computer terminals are eliminated, and the information security level of the company is significantly improved. The configuration solution is scientific and effective. It is capable of reducing risks from computer terminals to enterprise information security, and it can provide reference to information security construction in companies of the similar situations.

Keywords Information security · 3rd-generation nuclear power · Computer terminal · Network admittance

Z. Song (✉) · Y. Luo
China Nuclear Power Technology Research Instiute,
China Guangdong Nuclear Power Holding Co.Ltd,
Shenzhen, China
e-mail: songzhiping@cgnpc.com.cn

Y. Luo
e-mail: luoyi08@cgnpc.com.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_4,

4.1 Introduction

4.1.1 Status of Information Security of Computer Terminals in 3rd-Generation Nuclear Power

The Taishan Nuclear Power Plant (TSNP) is co-invested by the China Guangdong Nuclear Power Group and the Electric De France, and the company is responsible for the phase I project of constructing and operating the Taishan nuclear power plant. The power units applied are the CEPR (Chinese EPR) units designed and constructed with co-efforts of China and France based on the EPR technology (3rd-generation European Pressed-Water Reactor). The CEPR technology integrates the mature experience of N4 from France and KONVOI from Germany in engineering, construction and operations, while referring to the domestic system of standards and the leading practices of CPR1000 nuclear power construction. At present, TSNP possesses a total of approximately 1,500 computer terminals, maintained by the Taishan Branch of Center of Information & Technology (CIT) of the China Guangdong Nuclear Power Group (CGNPG).With the rapid progress of information construction, computer becomes a necessary tool in daily work and coming along were unprecedented challenges of information security including virus, Trojans, hacker attacks, etc. Analysis to TSNP information system shows that there are typical issues of security including easy access to internet, inconsistent configuration, substandard governance and limited guarantee to data security (see Fig. 4.1). TSNP urgently needs a complete set of security insurance

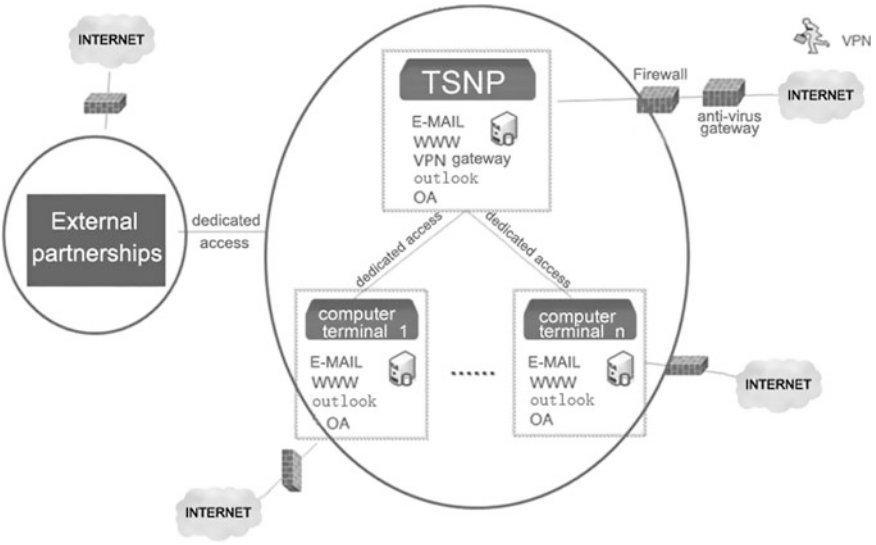


Fig. 4.1 Status of TSNP network

solution for computer terminals that fits with the national information security standards [1, 2, 3].

4.1.2 Risks in Security of Computer Terminals

As the Intranet of the TSNP expands, the amount of computer terminals grows as well. As a result, the Intranet suffers from increasing information security hazards from the Internet, e.g. network worms, internet attacks, junk mails, etc. Due to lack of consistent security management, the virus database could not be renewed on time. As security patches for operation systems of computer terminals are not updated on time, the accumulation of risks would finally damage the information security construction of TSNP.

4.2 Overall Targets and Overall Requirements of Information Security of Computer Terminals

4.2.1 Overall Targets

Overall targets of the computer terminals information security in TSNP can be concluded as: construct a scientific, efficient and effective protection system to address the special requirements of information security in TSNP, and fulfill the information security objectives through both regulations and technologies. Finally, we would achieve the target of "no external access, could not see it after entering the internal network, could not take it away even if they see it, could not use it after taking it away, and the operations could be traced."

4.2.2 Overall Requirements

Management measures and technological methods are equally important and both necessary for the information security of computer terminals of TSNP. If the focus is put on and only on management measures, and advanced technologies are not adopted, we would rely purely on the conscience of the practitioners. If the staff is not self-restricted, the management measures will not really work. On the contrary, if we use technological methods without sound management measures, information security will not receive as much attention as necessary, and the technological methods would block the construction of information security as a result.

4.3 Overall Solution of Information Security of Computer Terminals

4.3.1 Management Policy

TSNP shall establish an information security management organization to prepare the regulations and promote the importance of information security construction. Responsibilities of this organization shall include:

1. Establish regulations and operation systems suitable for the information security control of computer terminals in the special environment of TSNP.
2. Classify the departments in TSNP into different groups and apply stricter security configuration to computer terminals.
3. The CIT provides suggestions to TSNP over information security, technical improvement, and other optimization measures.
4. Carry out information security trainings to the staff of TSNP.
5. Monitor the communications of confidential information.
6. Work with the Audit Dept. to carry out inspections over information security in TSNP.
7. Report the progress of information security construction to the executives of TSNP and the confidentiality office of CGNPG.

4.3.2 Management of Security of Computer Terminals

Management of security of computer terminals is an important security system, and will directly affect the success of the construction of information security.

4.3.2.1 Network Admission of Computer Terminals

Purpose: control the access to internet from the intranet, avoid illegal access, and reduce the risks of virus and hacker attacks to the computer terminals [4, 5].

Measures: Carry out registration management to computer terminals. Computers linked with the intranet of TSNP shall be double certificated for equipment and user ID. (See Fig. 4.2).

1. Use back office servers to verify if the computer terminals are registered with TSNP, and satisfy the security standards (computers shall be installed with network access control software and special certification software), and the computer will not be linked to the intranet until verified to be qualified.

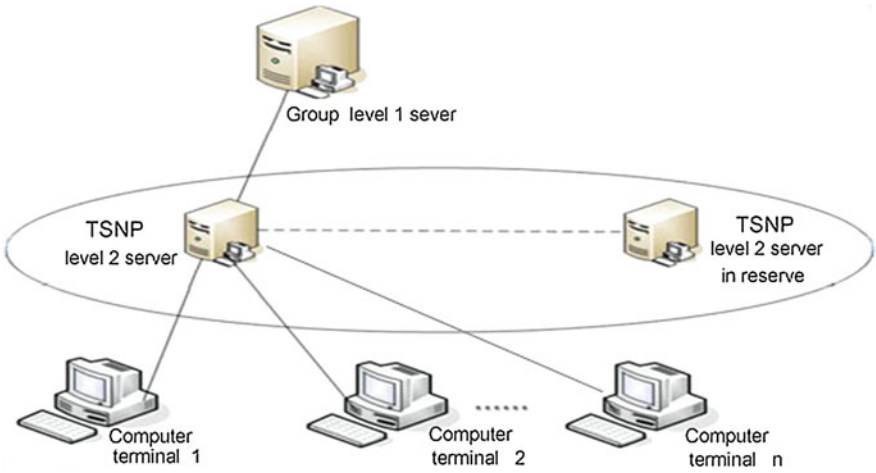


Fig. 4.2 Network admission of computer terminals in TSNP

2. The users shall put in the domain ID, pass word, and be verified to use internal information resources. Unregistered or unauthorized accounts will not be able to link to the intranet. Unsafe computer terminals (include terminal with virus) will have limited authority in the intranet.

Effects: After applying the Network Admission verification system, we have gained control over the linkage from external computers to the intranet, and the risks from external network are reduced. This is the first layer of protection to information security of the company.

4.3.2.2 Information and Data Management

Purpose: avoid leakage of information and data from various channels.

Measures: manage the means of information and data transferring.

1. Wi-Fi accesses on computer terminals include Wi-Fi network, infrared connections, blue tooth, etc., could be effectively controlled via waterproof strategies, and the utility of Wi-Fi facilities can be counted and reviewed.
2. Printing via computer terminals can be controlled with waterproof strategies, and the content printed can be traced. Audit measures will be able to check if the printing involves confidential information.
3. Management of Mobile Storage Mediums.

Only mobile storage mediums registered with the TSNP waterproof system can be used on a company computer [6].

Use automatic encryption measures to protect data in the mobile mediums, so that the information saved in the mediums would not be leaked when the mobile medium is lost.

Data exchange: documents on the computer terminals cannot be copied into a mobile storage medium that is not registered and the data on a registered medium could not be visited on a computer which is not certified. The confidential agent can be set up as the only exit of document-exchange. If an ordinary employee would like to pass a document to a client outside the company, authorization from the confidential agency is needed.

Effects: Through measures of controlling data and information transferring in TSNP, we well managed the channels of information and data transferring, and reduced the risk of information leakage.

4.3.2.3 Inspections and Audit

Target: identify illegal operations on computer terminals through inspections and audit, and effectively control the behaviors of the terminals.

Content: use waterproof software to record the operations of users, include linking to the internet, receiving and sending emails, copying documents, and other computer activities that pass on information, monitor all operations on computer terminals, and audit the information as necessary. Build up the inspection and audit architecture of TSNP information security through establishing the four roles of administrators, auditors, confidentiality agent, and users.

1. Administrators manage the operations of the security system and deploy security strategies.
2. Auditors are responsible for the monitory, include: monitor the operations of administrators, check the operation log of computer terminals, manage log information, release log reports, etc. Auditors carry out audit on operations of users including creating, deleting, copying and renaming of documents. When a computer terminal is offline (e.g. notebooks), the security settings are still valid, ports still under control, and any offline operations will be recorded for potential audits. Log information provided by auditors would be the evidence that information security dept. can rely on to deal with illegal operations on computer terminals.
3. Confidentiality agents are the channel of data between inside and outside the company. The agents' review and control data flowing outward assist with information security affairs and act as the interface of information security.
4. Users are the objects of inspection and audits. Users follow the security strategies deployed by administrators and coordinate with auditors with review of logs of computer terminals.

Effects: Ever since the inspection and audit solution was carried out in TSNP, we have effectively controlled the operations on computer terminals, and set up the second layer of protection to information security of the company.

4.3.3 Standard Settings of Computer Terminals

Computer terminals are the direct objects of daily operations of users in daily work. They are a key part of information security and a weak point of defending risks.

4.3.3.1 Permission of Users

The CIT establishes unique user ID and pass word for each user in TSNP to visit the internet and there are strict rules over the composition, length, and valid span of pass words. The rules keep in consistency with the information security standards of CGNPG. Users have limited user accounts, which could not install or unload software or modify registries and key settings. Such strict control measures are adopted so that mis-operations and systems of computer terminals will not used by malicious software or Trojans [7].

4.3.3.2 Setup of Software on Computer Terminals

All computer terminals of TSNP are installed with genuine software that satisfies security standards. Consistent technologies are adopted in preparing cloning packages, so that all computer terminals use the same software. If a user has special requirements for software, customer service engineers will install the software after the requirements are approved.

4.3.3.3 Security Setup of Computer Terminals

Using the SEP software, virus with on-line computer terminals are checked weekly. Virus library of this software is forced to be upgraded by the server, and risks of virus can be detected in real-time manner through the monitory from servers. If the virus library of the SEP software on a computer terminal is not upgraded on time, the terminal would be automatically isolated in the intranet.

Waterproof software is used on computer terminals, and users' ID and pass words are used to register into this software so as to identify the users' identifications. Only users in the CGNPG group can use resources in the intranet, when being verified by the waterproof wall.

Necessary patches are forced to be updated on each on-line computer terminal through background pushing according to the server strategies. This practice effectively avoids spreading of hole type virus in the intranet, improves security of the system, and is the 3rd layer of protection to information security of the company. [8].

4.4 Conclusion

During the construction of information security of computer terminals of TSNP, the CIT has kept the management principles advance with time. Continuous learning and innovation have preserved in scientific and reasonable standards of information security construction. The researchers firmly believe that the execution of the solution stated above will help to improve obviously the capability of information security protection of computer terminals in TSNP. Besides, this solution will improve the sense of confidentiality of TSNP staff, and would benefit the information security work. The CIT has kept on working on the construction of information security of computer terminals, and providing TSNP a safe, reliable, fast and effective information security platform.

References

1. Objectives, measures & strategies of information security (CGN-IT-C4-A01). The standard program of CGNPG (2009)
2. Scope of the information security management system (CGN-IT-C4-A02), The standard program of CGNPG (2009)
3. Regulations of information assets' security management (CGN-IT-C4-D01), The standard program of CGNPG (2009)
4. Symantec. Symantec endpoint security solution—a reliable foundation of enterprise information protection. Inform. Netw. Secur. **3**, 28 (2007)
5. Zhang, T.: Specialist Clinic of Network Security technology. Tsinghua University Publishing House, Tsinghua (2005)
6. Chi, T., Chen, P.: The discussion of information security of removable storage medium. Secur. Inform. Netw. **10**, 62 (2008)
7. Harold, F. Krause, T.M.: Administrative manual of information security (2004)
8. Sun, Q., Chen, W.: Information security management. Tsinghua University Publishing House, Tsinghua (2004)

Chapter 5
Comprehensive Feature Index for Meridian Information Based on Principal Component Projection

Jianhua Qin and Chongxiu Yu

Abstract Specifically for quantify and extract meridian information, a comprehensive multi-acupoint feature index was given. The feature parameters of single acupoint were extracted and reconstructed based on AR parameter model. Then feature weight was obtained by objective weighting method and feature matrix was weighted. The ideal feature vector was built based on orthogonal transformation of eigenvalues in meridian feature space. Based on PCP, the distance between each feature vector and the ideal model vector was calculated, and the projection value of fixed-weighted feature matrix on ideal feature vector was obtained. The simulation results show that the method can be more stability and higher around 3 % in the recognition rate than the main acupoint in human multi-acupoint system. The same results also show that the recognition rates can be coincided with sort results.

Keywords Human Meridian · PCP · Orthogonal transformation · Feature extraction

5.1 Introduction

Life activities of the human body are an extremely complex process, and somatic information are transferred and communicated through meridian systems. "Biological Cybernetics" studies show that, the so-called "gas, blood" in the medicine meridian theory means "information carrier", "channels, collaterals" corresponds to "information channel", and "acupoints" corresponds "information input or output". The majority experts and scholars at present study the relationship between meridian and human physiology changes based on single acupoint [1].

J. Qin (✉) · C. Yu
School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing, China
e-mail: jianhua7@sina.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_5,

The studies show that acupoint has some unique signal features such as high complexity, uncertainty, multi-level and multi-development. However, the studies also show that some acupoints are closely related to body functions, and others are very low or even irrelevant to physiological functions. Furthermore, the acupoints are affected by the internal rules, the external stimuli, and other factors. So the single acupoint features are not very good at reflecting the whole meridian.

Statistics believe that everything has its particularity, contingency, and randomness, but not chaotic, not rules. Learn from extraction idea for multi-lead EEG feature [2, 3], and cluster and discriminant analysis [4, 5], this paper constructs a new comprehensive multi-acupoints feature index for channels and collaterals that identifies the physiological changes of the human body based on the principal component projection method (PCP).

5.2 Extraction and Reconstruction to Single Acupoint

According to the traditional Chinese medicine theory, each channel and collateral line has a certain number of acupoints which are the response points of human organ and physiological state and play a important role to adjust channels-and-collaterals and viscera-and-blood. Therefore, the acupoint feature parameters are established.

The time sequence parameters model method is a mature method in physiological signals of meridian and acupoint, and especially the AR parameter model is commonly used to extract the feature parameters of acupoint [6, 7]. The formula of the AR parameter model is given by:

$$s(n) = -\sum_{m=1}^{p} a_m s(n-m) + u(n) + w(n) \tag{5.1}$$

where $u(n)$, $s(n)$, $w(n)$ denote the input excitation signal, the output impedance signal, and the white noise sequence. Here, p is the model order, and a_m is the AR model parameter with the order for p.

Order p is a key problem to accurately reflect somatic state. When order number is very low, the AR spectrum is too smooth to reflect the spectrum peak. And when order number is very large, the AR spectrum is instability and easily produce false peak. In this paper, the AR model order is obtained by AIC criterion, and optimal order estimation is 39 based on a large number of meridian impedance samples.

The AR model parameters in formula (1) are directly used as the signal feature in the traditional feature extraction method. But the number of model parameters are very large, and at the same time each parameter only express the partial information of the system, which inevitably leads to reduce the classification capacity. So the model parameters are not very suitable for channels and collaterals diagnostic, and need to be reconstructed. The reconstructed feature is obtained by AR spectrum. Figure 5.1 shows the AR spectrum for three acupoint impedance samples.

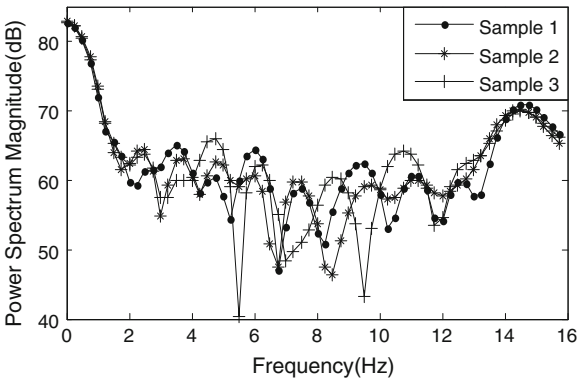


Fig. 5.1 AR spectrum for three meridian impedance samples of a tester

As can be seen from Fig. 5.1, AR model spectrum line is relatively smooth with multiple peaks in the frequency domain. The peaks are prominent and accurate with overcoming spectrum lines leak, emergence of side-lobe, low-resolution, and submerged by weak signal. It shows that AR model spectrum is conducive to the automatic computer extraction of feature parameters. And under the frequency less than 2 Hz, the spectrum line also decreases monotonically and is not well in reflecting acupoint differences. So amplitude maximum peak between 2–16 Hz is selected as feature peaks, and the frequency and the amplitude of the center cite in 2–16 Hz are extracted and named as center frequency and center peak. The energy of the frequency part, in which the frequency is higher than the frequency in the feature peaks, is named for high frequency energy, and then the high frequency energy is represented as high frequency percentage in the total frequency energy. Thus the feature vectors of acupoint impedance signal are composed of the five AR spectral features. Experimental results show that the feature vector can be very well in reflecting the signal features, and reduce the calculation and classification for the next step.

5.3 Comprehensive Feature Extraction for Channels

Firstly, the number of the feature-extracted acupoints in human channel line is set to m, and each acupoint has been described to the AR spectral feature vector ($\vec{a}' = \{a'_1, a'_2, \cdots, a'_n\}$). Thus sample matrix for the channels features may be written as:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} = (x_{ij})_{m \times n} \tag{5.2}$$

where i, j are the vector number and the acupoints number. Here, the eigenvector of the i-th acupoint corresponds to the i-th row of X (x_{i1}, $x_{i2}\cdots x_{in}$), and the j-th column acupoint feature corresponds to the j-th column of X (x_{1j}, $x_{2j}\cdots x_{mj}$). The j-th column acupoint feature means the evaluation index value of the j-th feature acupoints.

In the sample matrix composed of the multi-acupoints and multi-features, feature types are very big difference, and thus the features are standardized. Based on the linear function, the formula (2) is normalized and transformed to the new feature sample matrix, which is expressed as:

$$y_{ij} = \begin{cases} \dfrac{x_{ij} - \min\limits_{1 \le i \le m}(x_{ij})}{\max\limits_{1 \le i \le m}(x_{ij}) - \min\limits_{1 \le i \le m}(x_{ij})} \\ \dfrac{\max\limits_{1 \le i \le m}(x_{ij}) - x_{ij}}{\max\limits_{1 \le i \le m}(x_{ij}) - \min\limits_{1 \le i \le m}(x_{ij})} \end{cases}$$

where y_{ij} meet to $y_{ij} \in (0,1)$.

When the proportion of the single-acupoint features in multi-acupoints comprehensive features is larger, the single-acupoint features contain more information, and are stronger to be identified. So feature weight is introduced to denote the attention degree of the single acupoint features in the multi-acupoints. In the channels and multiple acupoints system, feature weight (λ_{ij}) of the single-acupoint (x_{ij}) is:

$$\lambda_{ij} = x_{ij} / \sum_{i=1}^{m} x_{ij} \tag{5.3}$$

The discrepancy between each acupoint is directly reflected by the difference degree of information entropy in the meridian system. In order to facilitate comparison and analysis, the information entropies for each acupoint are normalized and the result is:

$$H_j = -\left(\sum_{i=1}^{m} \lambda_{ij} \operatorname{In} \lambda_{ij} \right) / \operatorname{In} m \tag{5.4}$$

Then, feature weights of the j-th features are obtained based on objective weighting method and expressed as:

$$w_j = (1 - H_j) / \sum_{j=1}^{m} (1 - H_j) \tag{5.5}$$

Features matrix Y is weighted by using the feature weights (w_j). Let $z_{ij} = w_{ij} y_{ij}$. Where $Z = (z_{ij})_{n \times m}$ is weighted feature matrix and the feature vector for acupoints is:

$$\vec{d}_i = (z_{i1}, z_{i2}, \cdots, z_{im}), (i = 1, 2, \cdots, n) \tag{5.6}$$

The correlation between acupoints in the meridian system often causes mutual interference and overlapping to feature information, and thus the relative position of feature vector is difficult to analyze objectively. This may be solved by the orthogonal transform method that can filter duplicate information of the acupoints.

Set up: Feature values of weighted feature matrix are expressed as $\lambda_1, \lambda_2, \ldots, \lambda_m$ (It satisfies the inequality: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$), and correspond to the flat feature vector for $\alpha_1, \alpha_2, \ldots, \alpha_m$. Let $A = \{\alpha_1, \alpha_2, \ldots, \alpha_m\}$ and the weighted feature matrix are orthogonal transformed by Z. Thus the modified weighted feature matrix is obtained by equation $U = ZA = (u)_{m \times n}$, and denoted as:

$$\bar{d}_i' = (u_{i1}, u_{i2}, \cdots, u_{in}), (i = 1, 2, \cdots, m) \tag{5.7}$$

where $\bar{d}_i'$ is the influence value of the i-th acupoint in the meridian syndrome.

Here, if the influence value is higher, the influence of acupoint is stronger. Conversely, if the influence value is smaller, the influence of acupoint is weak.

First of all, take the optimal features as reference features, and maximum feature parameters in the channels system are used to construct the optimal feature vector. The feature vector is named to the ideal acupoint feature vector, and expressed as

$$\bar{Z} = (\bar{z}_1, \bar{z}_2, \cdots, \bar{z}_n)' \tag{5.8}$$

The expression (8) is united and then the following equation is obtained.

$$\bar{Z}_d = \frac{1}{\left|\bar{Z}\right|}\bar{Z} = \frac{1}{\sqrt{\bar{z}_1^2 + \bar{z}_2^2 + \cdots + \bar{z}_n^2}}\bar{Z} \tag{5.9}$$

Secondly, the projection value of modified weighted feature matrix in the reference feature vector is obtained by the PCP method and expressed as:

$$D_i = \bar{d}_i'\bar{Z}_d = \frac{1}{\sqrt{\bar{z}_1^2 + \bar{z}_2^2 + \cdots + \bar{z}_n^2}}\sum_{j=1}^{n}\bar{Z}_j u_{ij} \tag{5.10}$$

where D_i is the projection value set that is the channels comprehensive features.

5.4 Experimental Simulation and Analysis

In the experiment, the experimental objects are 25 normal volunteers with 25–35 years age and under the before and after strenuous exercise state. The measurement for each volunteer is repeated by 50 times in each state. In the experiment, the stimulation point and the reference electrode point are placed in Daling acupoint of Jueyin Pericardium Channel and Tianquan acupoint. And the received electrodes are placed in Shaoshang, Yuji, Taiyuan, Jingqu, Lieque,

Table 5.1 Features sample of Taiyin Lung channel of hand before and after standardization under normal human condition

Acupiont	Feature parameter before standardization					Feature parameter after standardization				
	Spectral peak	Center frequency	Center peak	Total frequency energy	High frequency energy	Spectral peak	Center frequency	Center peak	Total frequency energy	High frequency energy
Shaoshang	59.4818	5.4375	52.6634	363.8079	0.7113	0.5229	1.0000	0.1421	0.1333	0.3240
Yuji	61.1296	5.3594	58.9543	373.5321	0.7700	0.8968	0.6480	0.890	0.3433	0.5377
Taiyuan	61.5845	5.2969	59.8814	403.9406	0.6384	1.0000	0.3664	1.0000	1.0000	0.0586
Jingqu	60.1877	5.2156	54.9233	369.0267	0.7607	0.6831	0.0000	0.4107	0.246	0.5038
Lieque	58.6158	5.3000	55.3983	361.2367	0.8970	0.3264	0.3806	0.4671	0.0778	1.0000
Kongzui	57.7861	5.2163	51.4682	361.0062	0.7059	0.1381	0.0032	0.0000	0.0728	0.3043
Chize	58.0686	5.2844	53.5268	357.6359	0.7180	0.2022	0.3100	0.2447	0.0000	0.3484
Xiabai	59.2371	5.3172	54.6212	362.123	0.6223	0.4674	0.4579	0.3748	0.0970	0.0000
Tianfu	60.8983	5.4063	55.6170	387.6739	0.7850	0.8443	0.8594	0.4931	0.6487	0.5923
Yunmen	57.1773	5.4267	55.2457	364.4686	0.8317	0.0000	0.9510	0.4490	0.1476	0.7623
Zhongfu	59.4088	5.4375	53.8136	360.1794	0.7695	0.5063	1.0000	0.2788	0.0550	0.5359

Table 5.2 Comprehensive information feature of Taiyin Lung channel of hand for 6 testers before and after exercise

State	Tester	Shaoshang	Yuji	Taiyuan	Jingqu	Lieque	Kongzui	Chize	Xiabai	Tianfu	Yunmen	Zhongfu
Before strenuous exercise	1	0.2328	0.3980	0.0826	0.3634	0.7037	0.2131	0.2463	0.0102	0.4368	0.5398	0.3800
	2	0.2287	0.4012	0.0876	0.3872	0.7015	0.2212	0.2378	0.0200	0.3998	0.5421	0.3978
	3	0.2382	0.3887	0.0872	0.3321	0.6993	0.2229	0.2147	0.0172	0.4207	0.5212	0.4123
	4	0.2169	0.4109	0.0852	0.3492	0.7102	0.2082	0.2064	0.0178	0.4380	0.5683	0.4001
	5	0.2465	0.4321	0.7911	0.3725	0.7077	0.2180	0.2300	0.0203	0.4365	0.5676	0.3965
	6	0.2301	0.3771	0.8198	0.3612	0.6890	0.2008	0.2289	0.0117	0.4074	0.5307	0.3910
After strenuous exercise	1	0.1367	0.0709	0.5783	0.1344	0.2578	0.4573	0.4507	0.6624	0.1032	0.2280	0.1324
	2	0.1207	0.0723	0.6523	0.1723	0.2000	0.4726	0.4821	0.5842	0.1313	0.1001	0.1238
	3	0.1233	0.1201	0.6876	0.1721	0.2372	0.4321	0.3902	0.7432	0.2283	0.2342	0.1231
	4	0.1326	0.1231	0.6232	0.1800	0.2438	0.4721	0.4793	0.5832	0.0991	0.2543	0.1782
	5	0.1351	0.1483	0.6327	0.1743	0.2265	0.4962	0.4392	0.6666	0.1413	0.1867	0.1896
	6	0.1372	0.1821	0.6789	0.1657	0.2012	0.4583	0.4203	0.7029	0.1772	0.1938	0.1628

Kongzui, Chize, Xiabai, Tianfu, Yunmen and Zhongfu acupoint of Taiyin Lung Channel of Hand. The excitation signal in the experiment is the multisine (it is the periodic current signal superimposed by multiple positive (I) sine wave), in which sampling frequency for the excitation signal is 1 kHz, and the received signal is voltage signal [8]. The part result was given by Tables 5.1 and 5.2. (Due to limited space, only the results of six testers were listed).

In order to analyze the multi-acupoints feature index, the Elman Neural Network method is used to recognize the comprehensive feature vector of multi-acupoints and the feature vector of Shaoshang, Chize and Tianfu acupoint. In the simulation experiment, status output is set up for 0 with the before-exercise state and for 1 with the after-exercise state, and permissible error of status output is 0.2. It means that the output result in 1 ± 0.2 is considered to be the moving state. In addition, the first 10 sets of data in each target are as learning samples and the last 40 sets are testing samples. The recognition results are shown in Figs. 5.2, 5.3. In Figs. 5.2, 5.3, the ordinate and the abscissa are the recognition rate and the test personnel number.

As can be seen from Figs. 5.2 and 5.3, the average recognition rate of the multi-acupoints feature that is above 95 % under the before-exercise state and above 94 % under the after-exercise state, is about 3 % higher and more stable than the single acupoint in the same state. It can also be seen from Figs. 5.2 and 5.3, the recognition rate in Tianfu acupoint is higher than Shaoshang acupoint and Chize

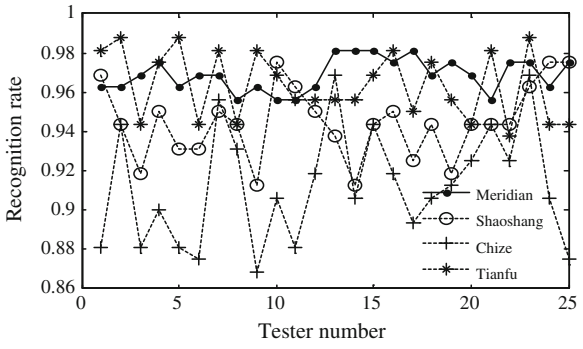


Fig. 5.2 Recognition results based on neural network before exercise

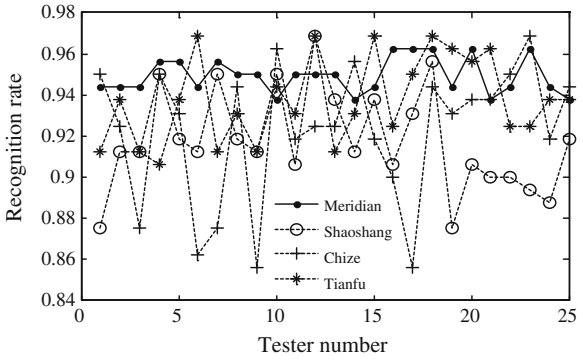


Fig. 5.3 Recognition results based on neural network after exercise

acupoint. It indicates that the feature weight of Tianfu acupoint in the channel (Taiyin Lung Channel of Hand) system is higher than other two acupoint, which is coincide with the PCP sort of the whole channel. The same result is also proven in other channel.

5.5 Conclusion

Inspired by cluster and discriminant analysis, a comprehensive multi-acupoint feature index method is established based on PCP method in this paper. Feature weight and orthogonal transformation are introduced to this method, and then principal component analysis method and projection method are merged organically to solve acupoint difference-degree in channels and collaterals system, mutual interference and overlapping of information feature and unit schedule for acupoints feature in a different time or space. The recognition results based on Elman Neural Network show that the recognition rate of this method is more stable and about 3 % higher than the single acupoint. And at the same time, the recognition rate of single acupoint coincides with the PCP sort. It sets up the foundation for further identification of human disease states based on the meridian signal.

References

1. Mo, F.: Multi-discipline research in traditional Chinese medicese-an exploration. Bulletin of National Science Foundation of China **13**(2), 45-53 (2005)
2. Cui, J., Wang, X., Li Z., et al.: Method of surface EMG pattern recognition based on AR parameter model and clustering analysis. Acta Metrologica Sigica, **27**(3), 286–289 (2006)
3. Kalil, M., Duchene, J.: Uterine EMG analysis: a dynamic approach for change detection and classification. IEEE Trans. Biomed. Eng. **47**(6), 748–756 (2000)
4. Wang, H., Zhou, Y., Yang, J., et al.: PPCA in DCT domain for face recognition using embedded HMM. J. Shanghai Jiaotong Univ. **41**(6), 885–888 (2007)
5. Zhou, D., Yang, X., Peng, N.: A modified linear discriminant analysis and its application to face recognition. J. Shanghai Jiaotong Univ. **39**(4), 527–530 (2005)
6. Ma, L., Yang, Y.: A study of the bio-impedance measurement system and extraction of the impedance characteristic parameters. Space Med. Med. Eng. **15**(3), 199–202 (2002)
7. Cover, T.M., Thoms, J.A.: Elements of Information Theory. John Wiley and Sons, New York (1999)
8. Van der, O., Edwin, S.J., Rnneboog J.: Peak factor minimization of input and output signals of linear systems. IEEE Trans. Instrum. Meas. **37**(2), 207–212 (1988)

Chapter 6
The Centralized Maintenance Mode of SAP System Based on Finance Shared Service Center

Heng Cheng and Ye Wang

Abstract In order to support those companies which use SAP as their ERP management software to develop the Finance Share Service Center (SSC), the centralized maintenance mode is strongly recommended. In this paper, by introducing the disadvantage of common decentralized maintenance mode, elaborating the structure of the SAP centralized maintenance mode and analyzing this mode's effectiveness, the SAP centralized maintenance mode is proved to be a better solution to resolve the conflicts and difficulties between SSC and the companies accepting SSC services. Thus, SAP centralized maintenance mode is able to provide a better and safer SAP maintenance service to SSC.

Keywords Finance shared service center (SSC) · The centralized SAP maintenance mode · The organizational system · The maintenance tool platform · The maintenance management regulation system

6.1 Introduction

With companies expanding constantly and business environment becoming more complex day by day, the management of company faces greater challenges. In order to control operating cost, enhance financial control intensity, prevent risks both in finance and business more effectively and obtain more long-term competitive advantages, Finance Shared Service Center (hereinafter referred to as "SSC") makes the debut as a new management mode.

In this paper, the SSC means such a distributed management mode: it will utilize the information technology and integrate relevant business processes; its

H. Cheng (✉) · Y. Wang
Information Technology Center, China Nuclear Power Technology Research Institute, Shenzhen, China
e-mail: chengheng@cgnpc.com.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_6,

purpose is to optimize organizational structure, standardize business processes, improve operational efficiency, cut down operating cost or create values; it can provide professional service with the market perspective for internal and external customers [1].

Obviously, SSC's dependency on information systems and requirements for maintenance supports are very high. In addition, while SSC is expanding, the integration of information systems will also increase and the application of the information technology will be wide and deep. The relation between SAP system and other systems is shown as Fig. 6.1. All these will put great challenges on the maintenance of SAP system for those companies which use SAP system as their ERP management software. If these companies take the usual way of decentralized mode as SAP system maintenance, the difficulties listed below will be inevitable for SSC and the companies accepting SSC services, and even for the maintenance of SAP system.

Firstly, according to the statistic data, as the number of companies accepting SSC service increases, about 80 % of the SAP system maintenance tasks are accomplished by the SAP maintenance group in the SSC, the rest are accomplished by the maintenance group in the companies accepting SSC services. Therefore, it will cause unbalance in task distribution between the SAP maintenance group in SSC and that in the companies accepting SSC services. This will lead to the difficulty in allocation of consultants to each maintenance group when the maintenance tasks are uneven and heavy. In addition, when the consultant in

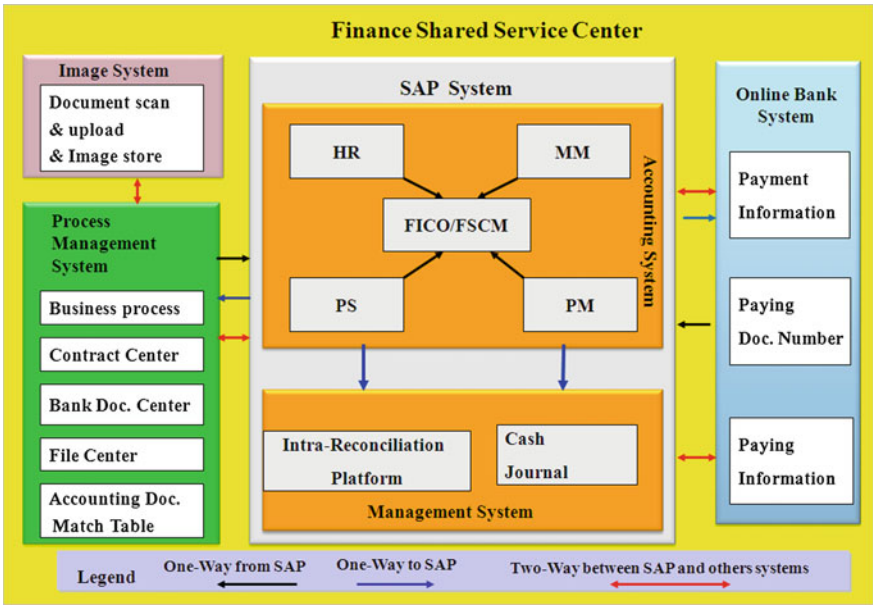


Fig. 6.1 The relation between SAP system and other systems based on SSC

one group is changed, if the successor can't be competent for the job immediately, it will cause serious effect both in the service quality and user's satisfaction.

Secondly, because maintenance consultants are dispersed in SSC and the companies accepting SSC services, it is easy to cause the lack of necessary experience sharing, knowledge exchange and information feedback in the maintenance group, then finally cause information isolation.

Thirdly, when the number of the companies accepts SSC services increases, decentralized maintenance mode will increase maintenance costs continuously [2].

Furthermore, because each maintenance group pays attention only to the maintenance tasks of themselves, they will lack of the macro and global consciousness. It is likely to affect SAP system both in the security and the global controlling.

Therefore, for the company that has equipped SSC, it is inevitable to adopt the centralized SAP maintenance mode when developing finance shared business.

6.2 The Centralized SAP Maintenance Mode Based on Finance Shared Service Center

What kind of centralized maintenance mode will meet the needs to SSC?

In general, there are four types of maintenance tasks in SSC: dealing with daily problems, doing user authorizations, fixing program bugs and realizing user's demands. Therefore, the centralized SAP maintenance mode that meets SSC's requirements should fulfill the following needs [3]: completing tasks according to schedule [4], completing works in more efficiency and with high satisfaction from user. That means it is necessary to take three aspects into account when building the centralized SAP maintenance mode, setting up the organizational system, building the maintenance tool platform [5], and establishing the maintenance management regulation system.

6.2.1 Setting Up the Organizational System

The setting up of the organizational system includes two aspects: design of organization and personnel structure, design of relevant post responsibilities.

(1) For the organization and personnel structure design, and according to SAP maintenance tasks, the centralized SAP maintenance group should be designed to include the foreground maintenance, the background supporting, the expert group and the dispatcher who allocate the tasks. The personnel structure is shown as Fig. 6.2.
(2) For the settings of post responsibilities, as each maintenance consultant has different abilities to resolve different requirements, they should be allocated maintenance tasks according to their abilities. According to this principle, the

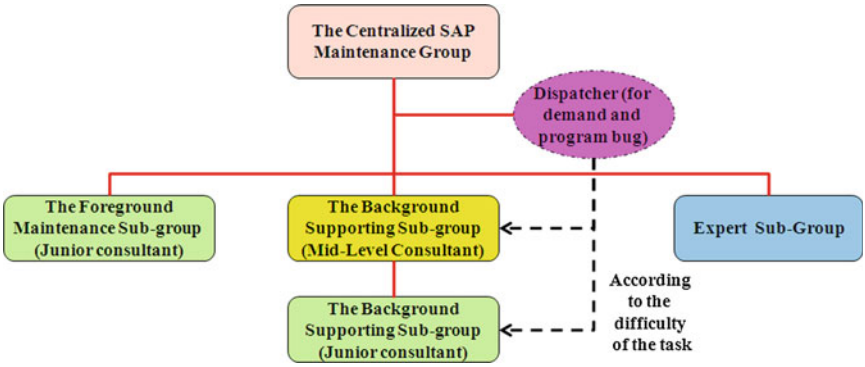


Fig. 6.2 The organization and personnel structure of the centralized SAP maintenance mode

maintenance consultant should be divided into three levels from junior, mid-level to expert.

(a) The junior consultants usually deal with daily business problems and user authorizations by telephone, the maintenance platform and e-mail. According to dispatcher's arrangement, the junior consultants will complete the diagnosis and fixing of program bugs, and realizing demands on schedule.

(b) The mid-level consultants make the diagnosis and fixing of program bugs, evaluate and realize the more difficult demands on SAP System, especially those maintenance tasks that are inter-module or inter-system. Of course, when the junior consultants can't handle the case, the mid-level consultants are obligated to assist or guide the juniors to do it.

(c) The experts must ensure that SAP system is safe and reliable, control the risks towards SAP from configuration changes and customized programs, prevent from effect the multi-system operation of SSC and improve the overall maintenance ability of the centralized SAP maintenance group. Therefore, the experts should focus on evaluating the SAP system solutions comprehensively from the technical point, reviewing adequately configurations and the programs of SAP to find out any conflicts or errors. When necessary, they need organize some kinds of the system tests, such as the regression testing. In addition, the experts have the responsibility to organize regular experience sharing and knowledge exchange, to build up knowledge repository of maintenance problem handling solutions [6] and improve the maintenance management regulation system. Furthermore, when lower level and mid-level consultants can't deal with some problems, experts need to organize discussion with consultants involved and decide the final solutions.

(d) The main work of dispatcher is to evaluate the difficulty of the maintenance tasks received from users, and determine the task schedule, then assign tasks to the appropriate consultants. Certainly, dispatcher also needs to analyze the processing of tasks and trace the progresses. Thus, it is more suitable to choose dispatcher from mid-level consultants.

6.2.2 Building the Maintenance Tool Platform

For the purpose of utilizing resource sufficiently and completing the maintenance tasks of SSC efficiently and timely, it is necessary to build a maintenance tool platform so that maintenance tasks can be dealt with in workflow mode. The workflows can be created in accordance with maintenance task classification. When users initiate relevant workflows, the consultants in the centralized SAP maintenance group start to work according to the corresponding workflow steps. The platform ensure that each maintenance task can be recorded and traced from initiation, confirming, distribution, processing, testing to going-live, shown as Fig. 6.3. Furthermore, this ensures tasks will be completed by consultants according to schedule, and all kinds of relevant reports can be generated about maintenance task statistical data. In this way, knowledge repository will be created gradually by accumulating problem handling solutions continuously.

In the structure of maintenance tool platform is shown as Fig. 6.4.

6.2.3 Establishing the Maintenance Management Regulation System

In order to fulfill the needs of SSC's development and strengthen the foundation of the organizational system and the maintenance tool platform as well as guarantee each maintenance task which completed by the rules, the maintenance

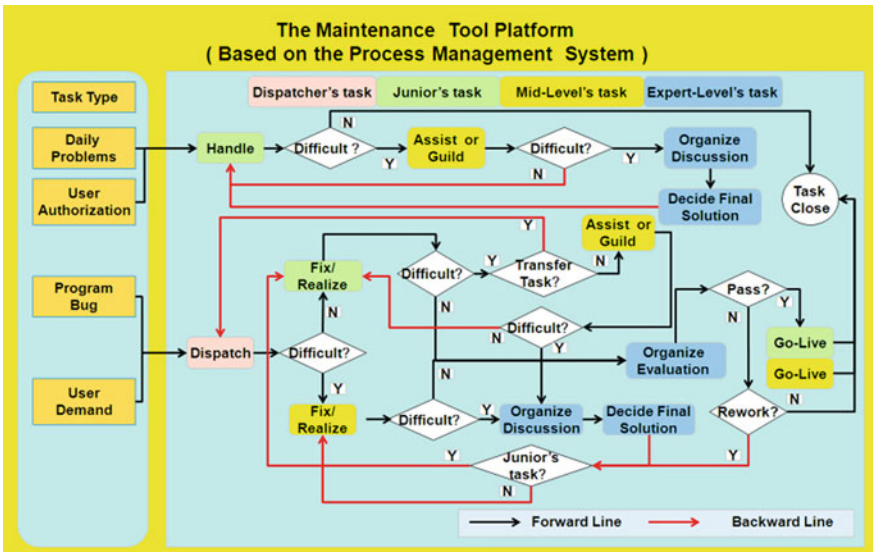


Fig. 6.3 Diagram of post responsibility and task execution

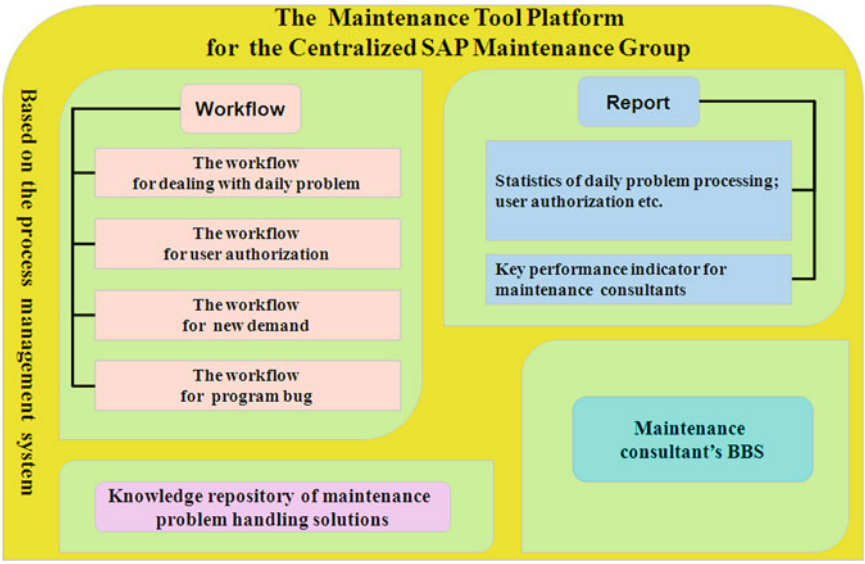


Fig. 6.4 The structure of maintenance tool platform

management regulation system must be established. In the regulation system, the organization and personnel structure of centralized SAP maintenance group, the post responsibility and the maintenance service catalog should be well defined. Regulation should be set up for user authorization of SAP system, for SAP system go-live management of program bug fixing and demand realization, for the transition from implementation stage to maintenance stage, for knowledge exchange management, and for the management of knowledge repository of maintenance problem handling solutions, etc.

6.3 Effectiveness Analysis

By optimizing the organization and personnel structure of centralized SAP maintenance group, establishing the maintenance tool platform and improving the maintenance management regulation system, the centralized SAP maintenance mode brings more benefits to the SSC' s development in the aspect of information system building.

First, it helps improving maintenance work quality and user's satisfaction [7]. By deploying consultants and arranging maintenance tasks reasonably, it ensures the rational utilization of maintenance human resources, improves the response speed and increases the user satisfaction to SSC.

Second, it helps strengthen maintenance capability. Decentralized maintenance mode easily leads to information isolation. The centralized SAP maintenance

mode can be a good way to avoid such a situation because it can decrease the waste of resources in configuration or program development to SAP system. In addition, by training the consultants regularly and comprehensively, organizing experience exchange and knowledge sharing constantly, improving the knowledge repository of maintenance problems handling solutions, it can improve consultants' ability in analyzing and resolving problems and eventually lift the overall maintenance service level.

Third, it helps cut down the maintenance cost continuously [5]. Because the centralized SAP maintenance mode makes more reasonable arrangement for maintenance consultants, it can ensure that the maintenance tasks will be completed on schedule, in more efficiency and with higher success rate. It also enables consultants to have more spare time and effort to learn deeper knowledge, this will final bring the enhancement in human resource efficiency [8] and cut down maintenance cost.

The fourth, it is good to strengthen global control intensity to SAP system [9]. Because the centralized SAP maintenance group resolves problems or decides solutions in a macro and global perspective, they can analyze demands comprehensively and unify system solutions perfectly, prevent cross influence between SAP system and other systems or between different SAP modules. So it can ensure SAP system's safety and prevent risks to the maximum extent.

6.4 Conclusion

As the concept and content of SSC are changing constantly, the centralized SAP maintenance mode is subject to change as well. It should be optimized in mode, workflow management and structure design continuously. In addition, it is necessary to create and optimize training and promotion mechanism, strengthen consultant echelon and establish the knowledge repository of maintenance problems handling solutions.

If we firmly push forward the optimization of the centralized SAP maintenance mode, it is sure that the centralized SAP maintenance group will provide the best and safest service to SSC and obtain the double win with SSC' s development.

References

1. Chen, H., Dong, H.: Shared service of finance and accounting. China Financial and Economic Publishing House, pp. 25 (2008)
2. Gao, Y.F., Wang, Y.W., Liu, J.: Research on maintenance modes of group organizations' ERP System. Sci. Technol. Rev. **9**(23), 36–39 (2005)
3. IT Service Business Division in Neusoft Group. IT Maintenance System Building and Implementing Resolve Solution in Power Industry. Unpublished, p. 26 (2012)
4. Doc Palmer.: Centralized versus decentralized maintenance. Reliable Plant p. 1 (2008)

5. Brown, C.V.: The IT organization of the future. In: Competing in the Information Age: Align in the Sand, 2nd ed. Oxford University Press (2003)
6. Xu, H.: Study of power grid's information system maintenance management in data centralized mode. Electr Power IT. **7**(3), 19–20 (2009)
7. Dai, W.Z., Li, K.T., Wu, Y.M.: How to manage the complicated tax system-studying the building of centralized maintenance management system of Shenzhen Municipal Office of SAT. Comput. IT Week. **16**, 1 (2006)
8. Kris, K.: http://www.linkedin.com/answers/management/planning/MGMPLN/538263-11098632 (2009)
9. Rochte, T.: IT service management: centralize or decentralize IT operations? Nimsoft Modern IT blog, p. 1 (2012)

Chapter 7
An Improved Method of Polyphase Filter Banks Channelization

Min Li and Fengming Bai

Abstract In order to solve the large amount of data of polyphase filter banks channelization, a new method is used in this paper. The method greatly simplifies the complexity of the channel filter banks channelization, in the extraction rate to the original 2 times and reduces the need for processing data transmission rate. This method has a better signal real-time processing ability and good application prospect. Through the simulation experiment, this method proves to have a good feasibility.

Keywords Polyphase filter · Filter banks · Channelized receiver

7.1 Introduction

Current channelization is usually used in the If or RF. From the existing engineering examples, we often rely on parallel filter banks to realize channel not more than 6 [1], but through the use of software to achieve the channelized receiver, the command to deal with the channel is often more than 6 [2]. In this case, if we still use the traditional parallel multichannel structure, it will inevitably lead to the receiver hardware platform too large [3], decreased the stability. Due to the limit of Nyquist sampling theorem, receiver high speed A/D conversion data rate is at least two times for processing bandwidth, so directly use DSP to process such data flow is very difficult, we should reduce the data rate. Based on the current DFT

M. Li
College of Humanities and Information, Changchun University of Technology, Changchun University of Science and Technology, Changchun, China

F. Bai (✉)
Changchun University of Science and Technology, Changchun, China
e-mail: 82587739@qq.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_7,

filter banks and polyphase channelization method, we can temporarily alleviate the problem [4], because in the actual process, the signal we encountered in general is real signal. On this premise, this paper combines the real signal frequency spectrum characteristics, according to the given channel frequency division scheme, based on the concept of polyphase filter, derivate and set up a channelized model. Finally, a simulation system is established, it shows that the scheme is correct under the simulation.

7.2 Multiple Sub-Band Channelized Model

As we know, the spectrum of real signal is symmetrical, in order to reduce the use of bandwidth, the real signal channel can be redevided. The spectrum expression is given as follows:

$$\omega_k = \left(k - \frac{2A-1}{4}\right) \cdot \frac{2\pi}{A} \tag{7.1}$$

ω_k is 0 to K−1 sub-band normalized angular frequency, A is the data extraction ratio.

After channel partition, we can put the whole frequency band into a plurality of sections, then move to zero frequency nearby, this can be achieved by low-pass filter [5]. Because after the shift, the signal is a complex signal, we can extract the signal of sub-filter 2A times. Combined with the concept of polyphase filter, we can get new polyphase filter channel structure, allow each subband signal parallel output. The amount of calculation is reduced to the original One Ath, it also greatly improves the ability of real time signal processing.

7.3 Polyphase Decomposition of Filter

From the Fig. 7.1 we can get that channelized receiver extraction is after filtering, so the data retained only the 1/A when A is big enough, the computational efficiency is the main problem of the hardware implementation.

Here we introduce the polyphase decomposition of the filter. We suppose a lowpass filter transfer function $H_0(z)$, It's A with polyphase form can be expressed as

$$H_0(z) = \sum_{l-0}^{A-1} z^{-l} E_l\left(z^A\right), \tag{7.2}$$

$E_l(z)$ is the lth polyphase components;

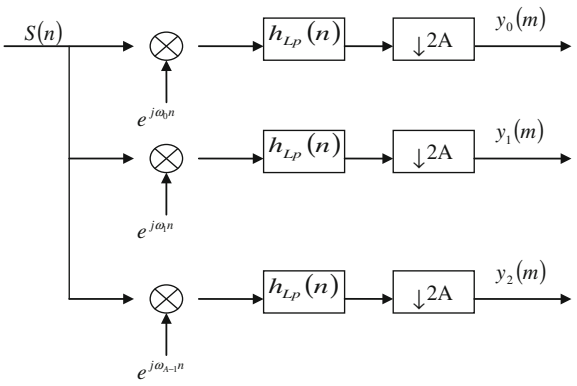


Fig. 7.1 Real signal filter low pass implementation

$$E_l(z) = \sum_{n=0}^{\infty} e_l[n]z^{-n} = \sum_{n=0}^{\infty} h_0[l + nA]z^{-n}, \qquad 0 \le l \le A - 1 \tag{7.3}$$

In order to simplify, we use $W_A^{kA} = 1$ instead of Z can express a moved subfilter frequency response, then we can get the Kth sub filters corresponding to the frequency of the A band polyphase decomposition:

$$H_k(z) = \sum_{l=0}^{A-1} z^{-l} W_A^{-kl} E_l\left(z^A W_A^{kA}\right) = \sum_{l=0}^{A-1} z^{-l} W_A^{-kl} E_l\left(z^A\right) \tag{7.4}$$

If $W_A^{KA} = 1$ is used in the equation, then last expression can be used in matrix form:

$$H_k(z) = \begin{bmatrix} 1 & W_A^{-k} & W_A^{-2k} & \cdots & W_A^{-(A-1)k} \end{bmatrix} \begin{bmatrix} E_0\left(z^A\right) \\ z^{-1}E_1\left(z^A\right) \\ z^{-2}E_2\left(z^A\right) \\ \vdots \\ z^{-(A-1)}E_{A-1}\left(z^A\right) \end{bmatrix} \tag{7.5}$$

7.4 Real Signal Completely Polyphase Decomposition Filtering Banks Channelized Structure

Through the above analysis, because of the introduction of the polyphase decomposition, we can implement this structure before A times extracted device to improve the efficiency of the operation. But two times decimation filter is still at the back of the first level multiplier. So from the hardware implementation point of

view, it will increase the complexity. If some processing makes 2 times decimation can also work before the first multiplier, it can get more efficient polyphase filter structure. Here we only analysis from the filter by a branch. By polyphase decomposition, we can get Fig. 7.2.

Implementation of polyphase filter design steps: first, according to the input data rate and the actual filter indicator, we can design prototype low-pass filter; second, calculate the prototype filter orders and coefficients. Finally, use the following formula calculated for each subset of polyphase filter coefficient

$$h_k(m) = h(m \cdot 2A + k), m = 0, 1, 2, \cdots, \tag{7.6}$$

The improved structure is mainly used for baseband processing module, the channelized structure of each branch has an alternate symbol converter, and it can be collected after the data alternation symbol, achieved after the fixed-point data bitwise which can save a large number of multipliers.

7.5 Complexity Analysis

In order to illustrate the effectiveness of this method better, we aim at the existing methods of contrast complexity analysis. Firstly, we suppose the entire filter banks cost function for M, which the output of each channel is a data by multiplication times, then, direct to achieve a uniform filter banks complexity cost function $M = \frac{4A^2K}{\Delta f} + 4A^2$. The expression of A is for channel partition number, Δf is the

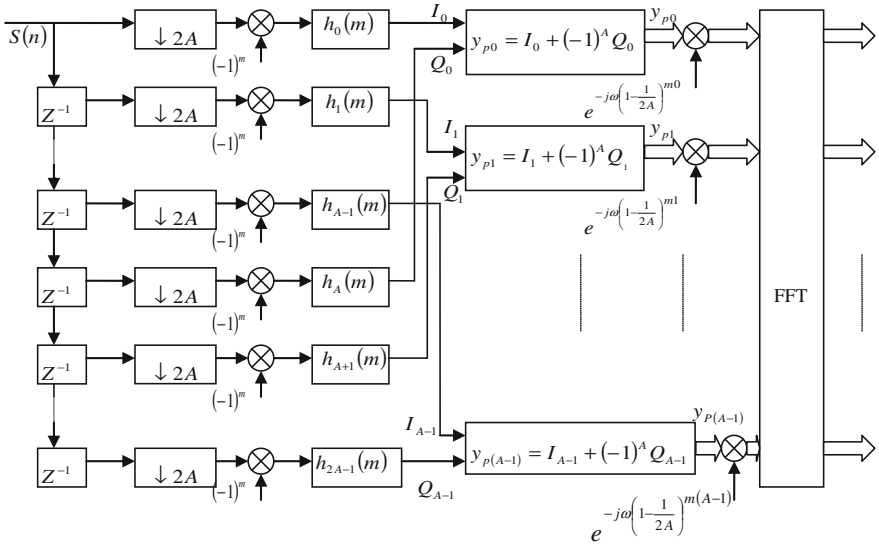


Fig. 7.2 Real signals completely polyphase decomposition filter banks structure diagram

prototype filter transition bandwidth, and because the transition bandwidth digital filters and filter order inversely, here we make K as the filter order and transition bandwidth reciprocal correspondence factor, while using the traditional multiphase filter banks cost function: $M_1 = A\log_2 A + \frac{4K}{\Delta f} + 8A$. According to the above analysis, the structure of the cost function: $M = A\log_2 A + \frac{K}{\Delta f} + 7A$. In this paper, by comparing the new design structure, the computational complexity significantly lower, so it is more suitable for hardware implementation.

7.6 Simulation Results

According to the above analysis, we can establish a 32 channel analog system. Samples of the input signal frequency F_S is 1, Samples are 8192, normalized bandwidth is 0.5, the prototype filter design using MATLAB FIRPMORD and FIRPM function determined by the FIR filter, order number is 1024, these parameters are used to determine the future of simulation experiments. At first, we produce a raised cosine pulse, and then it is modulated separately to each channel. In order to analyze conveniently, here we simplify the processing. The raised cosine pulse signal is modulated into four signals,then added to form the multiphase filter banks input signal, a four modulated signal normalized frequencies are 0.0625, 0.125, 0.1875, 0.25, through the real signal channel division we find that they should be observed in 9th, 11th, 13th, 15th channel. The experimental results are shown below: (Fig. 7.3).

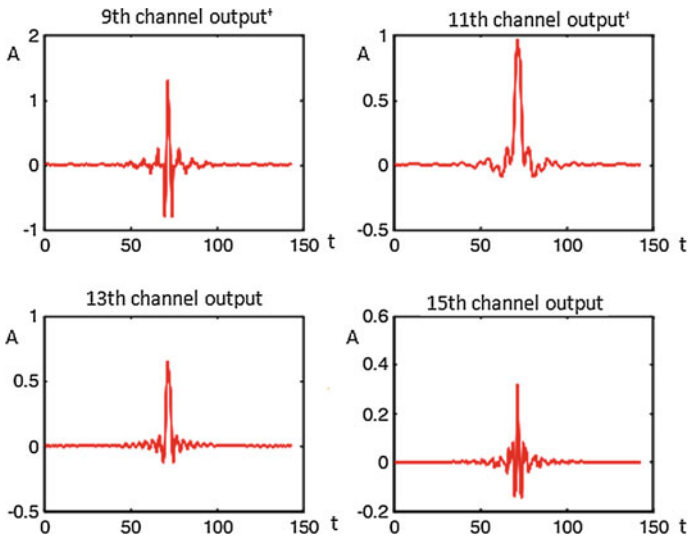


Fig. 7.3 Output waveform

We can find from the figure, the design of structure function is correct, if not simplified, the processing in accordance with 32 channels, each channel outputs much data. Direct to achieve a uniform filter banks need the number multiplication $M_1 = \frac{4A^2K}{\Delta f} + 4 \cdot A^2 = 135168$ Real signal polyphase filter banks need the number of multiplication $M_2 = A\log_2 A + \frac{4K}{\Delta f} + 8A = 4512$.

The design of multiphase filter banks need for multiplication times $M = A\log_2 A + \frac{K}{\Delta f} + 7A = 1408$. From the calculation amount, it is just real signal multiphase filter banks 1/3 only, so, this design structure is more convenient for hardware implementation.

7.7 Conclusion

In this paper, through theoretical analysis and simulation experiments, researchers present a signal complete decomposition of the polyphase filters channelized method. It has lower computational complexity and higher efficiency, also greatly reduces the complexity. It will be widely used in the sub-band separation system.

References

1. Henstchel, T.: Channelization for software defined base-stations, annals communications (2001)
2. Harris, F.J., Dick, C., Rice, M.: Digital receivers and transmitters using polyphases filter banks for wireless communications. IEEE Trans. Microwave Theor. Tech. **51**(4), 1395–1412 (2003)
3. Harris, F.J., Rice, M.: Multi-rate digital filters for symbol timing synchronization in software defined radios, IEEE J. Sel. Areas Commun. **19**, 2346–2357 (2001)
4. Abu-al-saud, W.A., Studer, G.L.: Efficient Wideband Channelizer for Software Radio Systems Using Modulated pr Filter Bands. IEEE Trans. Sig. Process. **52**(7), 1954–1962 (2004)
5. Zangi, K.C., Koilpillai, R.D.: Software radio issues in celluar base stations. IEEE J. Sel. Areas Commun. **17**(4), 561–573 (1999)

Chapter 8
An Approach for Large Scale Retrieval Using Peer-to-Peer Network Based on Interest Community

Shuang Feng, Shouxun Liu and Yongbin Wang

Abstract Conventional multimedia information retrieval systems use a central system to store and index multimedia data. Inherent limitations of such a central approach surface many problems, such as insufficient bandwidth, server overloading and failures. In order to retain the original system and control the cost, the paper shares the access pressure of central servers by constructing a peer-to-peer network based on interests. A user's interest is computed by mining this user's search behaviour periodically. Then researcher form a peer-to-peer network based on interests by clustering peers with similar interests. Centralized server will push relevant multimedia information to certain communities on time. By this way, uses can further clear what they want, and useless retrievals on servers will be dramatically decreased. The experimental results evaluate the average search path length of unstructured P2P network, semi-distribution P2P network and the P2P network based on interest community and demonstrate the efficiency of the approach.

Keywords Interest community · Peer to peer · Multimedia retrieval · User profile

S. Feng (✉) · Y. Wang
School of Computer Science, Communication University of China, Beijng, China
e-mail: fengshuang@cuc.edu.cn

Y. Wang
e-mail: ybwang@cuc.edu.cn

S. Liu
The Graduate School of Communication University of China, Beijng, China
e-mail: sxliu@cuc.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_8,

8.1 Introduction

With the rapid development of information technology, multimedia applications have been widely used in people's daily lives. The need of developing effective and efficient multimedia information retrieval technologies has been identified in recent years. Due to the large amount of computational power needed for the searching and processing of multimedia data, distributed multimedia information retrieval has attracted researchers' attentions [1]. But there are still many original retrieval systems with central servers. As more and more multimedia objects are collected and the scale of the applications grows, the inherent limitations of such a central approach surface, such as insufficient bandwidth, server overloading and failures [2].

The rapid development of P2P technology made it as one of the most disruptive tools for the construction of large-scale distributed system over Internet. P2P adopts a distributed and decentralized architecture, and each peer of P2P is equal and acts as both a server and a client, so P2P removes the drawback of the structure of central server.

Compared with breadth-first search in the network, retrieval efficiency can be greatly improved if peers which most probably contain relevant data are visited first. The assumption is that every peer has its own topics of interest. These interested topics are the reflection of the interests of the user behind the peer. The user is also more likely to query multimedia data on the topics that he/she is interested in. The multimedia information retrieval process can be used to facilitate relation establishment between peers in the network. The relevancy judgment of the results returned made by the querying peer can be seen as an assessment to the information retrieval performance of their corresponding data sources. In this way, we model the multimedia information retrieval network as a social network and propose a multimedia retrieval model based on P2P and a recommendation system based on interests to decrease the access of central servers.

The remainder of this paper consists of five parts. In Sect. 8.2, related work is introduced. In Sect. 8.3, an interest-based P2P system architecture is presented. Section 8.4 describes the calculation of user profile. Section 8.5 shows algorithm of community evolution. Section 8.6 is the evaluation of our approach. Finally, conclusions are presented in Sect. 8.7.

8.2 Related Work

To reduce the number of query search messages in the P2P network, many algorithms propose the concept of groups or clusters in P2P networks [3–5]. The nodes with more powerful resources (in term of processing power, memory and bandwidth) are the suitable candidates for the role of server, whereas, less powerful nodes become clients [4]. A new protocol was proposed for building and

repairing of overlay topologies based on the formation of interest-based superpeers [5]. An interest-based superpeer algorithm creates groups or societies that have common interests. A semantic community establishing method based on consideration of semantic similarity degree is proposed, trust degree and active strength, which improves the structural and resource-located veracity of P2P resource organization [6]. SOSPNET maintains a superpeer network topology that reflects the semantic similarity of peers sharing content interests [7]. Superpeers maintain semantic caches of pointers to files, which are requested by peers with similar interests. Client peers, on the other hand, dynamically select superpeers offering the best search performance. There has been work in community construction in P2P networks [8, 9]. Four sources of self-construction of P2P communities, namely, ontology matching, attribute similarity, trust, and link analysis were introduced to form community [8]. In communities for sharing academic papers, each peer computes its trust in the peers with whom it interacts [9].

8.3 Superpeer Overlay Network

To take full advantage of the client's ability, we build a P2P overlay network on internet to alleviate the pressure on servers. Relevant peers are clustered based on their historical behavior on the retrieval server, called interest community. By this way, the network is divided into different interest communities. Each of community elects a superpeer and a backup superpeer, they are responsible for cooperative management of their own community. IS will sent heartbeat information on time so that it knows weather communities work well. SP will also sent heatbeat information to IS if there have some changes in the community. IS will push recommendation multimedia information to certain communities on time. By this way, uses can further clear what they want and useless retrievals on servers will be dramatically decreased. The Architecture is shown in Fig. 8.1:

There are five entities in our system, they are:

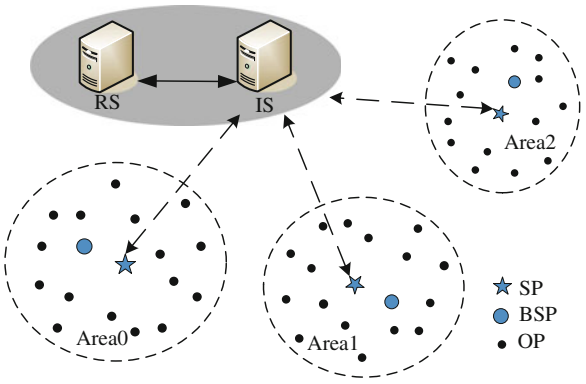


Fig. 8.1 Overall system architecture of the proposed scheme

- RS (Retrieval Server): Store all the information of multimedia materials, responsible for generating commendation file and responding client requests for the retrieval.
- IS (Index Server): Create and maintain community information dynamically.
- SP (Superpeer): Each community in P2P overlay network elected a superpeer based on the capacity of the node dynamically. SP is responsible for getting commendation file from IS and maintaining its community work well.
- BSP (Backup Superpeer): In order to avoid the problem of single point failure, SP designated a backup superpeer based on the capability and reputation of nodes in its community, BSP maintained the community information with SP cooperatively. When SP can't serve the community, BSP will take over the community and start updating mechanism.
- OP (Ordinary Peer): OP can join or exit the retrieval network. Each OP stores the information of RS, IS and community details such as SP and BSP.

8.4 User Profile Construction

Personalized social search can be achieved by utilizing historical query data from people in a community with similar interests.

Definition 1 Assuming each multimedia file has a topic and the number of topics is limited, the similarity of peers is defined as the similarity of a set of weighted topics.

$$\begin{aligned} Sim(P_1, P_2) &= Sim(<T_i, \lambda_i>, <T_j, \lambda_j>), i = 1, 2, \ldots\ldots m, j = 1, 2, \ldots\ldots n \\ &= \sum_{j=1}^{n} \sum_{i=1}^{m} Sim(T_i, T_j) \times (\lambda_i, \lambda_j) \end{aligned} \tag{8.1}$$

where P_i is the peer, T_i is the topic, λ_i is the weight of T_i, it can be calculated as follows:

$$\lambda_i = \frac{N_i}{\sum_{j=1}^{n} N_j} \tag{8.2}$$

where N_i is the number files belongs to T_i among all the files.

Many methods have been developed to measure the similarity of concepts. Our content summary of a peer is calculated based on the method proposed by Yuhua Li et al. [10], as follows:

$$Sim(T_i, T_j) = f_1(l) f_2(h) = e^{-\alpha} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \tag{8.3}$$

where l is the shortest path length between topics, h is the depth of subsumer in the hierarchy semantic nets, α and β are parameters scaling the contribution of shortest path length and depth, respectively.

8.5 Algorithm of Interest-Based Community Evolution

The algorithm, which is used to build the relationship between interest-based superpeer and clients, is shown in the Fig. 8.2. When the user searches multimedia on central servers for the first time, user's interaction will be recorded. By analyzing these, we can calculate the user's profile P_i. According to P_i, IS will sort all the communities through formula (3), mentioned in user profile construction. Compared P_i with SP_0, if there is no change, it means community SP_0 is still the most suitable one for P_i. Otherwise, we select SP_j where SP_j is most suitable for P_i. P_i asks for joining community SP_j. If it is OK, the community SP_j will check if P_i is able to be new SP according to its capacity. If so, both the community and IS status will be updated. If P_i can't join SP_j due to connection problems, IS will choose another SP for P_i until SP is NULL. If SP is NULL, a new community will be created.

Fig. 8.2 Algorithm of interest-based community evolution

```
FormCommunity ()
{
  p_i =CalUserProfile ();
  SP = Sort Community ( p_i );
  if (!CompareComm( p_i , SP_0 ))
  {
    while ( SP !=NULL)
    {
      select SP_j where SP_j is most suitable for p_i ;
      if ( !JoinComm( p_i , SP_j ))
      { SP = SP - SP_j ; continue ;}
      else
      {
        if ( CanBeSP( p_i ))
        {UpdateComm (); UpdateIS () ;}
        UpdateSP ();
      }
    }
    FormNewComm();
}
```

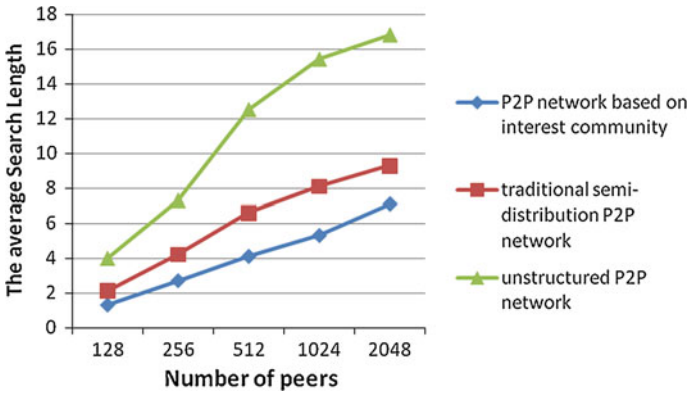


Fig. 8.3 The average search length of three kinds of P2P networks

When the system extends, the retrieval server must adjust two parameters, $K_{\max}$ (the max number of communities) and $N_{\max}$ (the max number of peers in a community). By merging or partitioning the communities, the communities can be adjusted dynamically, so that the scale of communities is average and each peer is connected to the community with common interests. The RS can control the network scale and the communities' partition, and enhance the controllability and availability of P2P overlay network.

8.6 Evaluations

We use P2PSIM as our evaluation tool. P2PSIM is a P2P simulation on Linux which can simulate kademlia, chord for P2P, and can simulate more P2P protocols by extending protocols and verify their functionalities. In our experiments, we generated 128, 256, 521, 1024, 2048 ordinary peers separately. Then we assigned ten queries to each peer randomly. Based on these queries, the similarity of each peer can be calculated according to formula (3). We also assume that peers with common interests share similar files. The max number of peers in each community is 50. Then we set a value to each peer that represents the capacity of the peer. Superpeer and Backup Superpeer are elected based on the capacity. All the results are the average value of five separated experiments. Figure 8.3 shows the average search path length of unstructured P2P network using flooding algorithm, traditional semi-distribution P2P network and P2P network based on interest community. From the figure, we can find out that the average search path length of our approach was dramatically decreased.

8.7 Conclusion

High-load and high-concurrency ask for a very high capacity of server. In this paper, authors proposed an approach for large scale retrieval by using peer-to-peer network based on interest community. Social information is acquired from a social network service application. Members in a community share certain interests. Communities are under the control of superpeers and backup superpeers. The central server will send relevant recommendation information to communities so that the users can further clear what they want and reduce useless retrieval dramatically.

Acknowledgments This paper is supported by the National Science and Technology Support Project (2012BAH02F02-01) and the National High Technology Research and Development Program of China (2011AA01A107).

References

1. Xia, T., Wang, F., Liu P., Palanivelu S.: Managing and searching distributed multi-dimensional annotations with large scale image data. In: Proceedings of the International Workshop on Multimedia Content Analysis and Mining (MCAM 2007), vol. 4577, pp. 361–370. LNCS, Springer (2007)
2. Rasolofo, Y., Abbaci, F., Savoy, J.: Approaches to collection selection and results merging for distributed information retrieval. In: Proceedings of the 2001 ACM CIKM International Conference on Information and Knowledge Management, ACM, pp. 191–198. Atlanta, Georgia, 5-10 Nov (2001)
3. Gatani, L., Lo Re, G., Gaglio, S.: An adaptive routing protocol for ad hoc peer-to-peer networks. In: Proceedings of the Sixth IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks, pp. 44–50 (2005)
4. Montersor, A.: A robust protocol for building superpeer overlay topologies. University of Bologna, Bologna, Technical Report, UBLCS- 2004-8 (2004)
5. Ashraf Khan, S.K., Tokarchuk, L.N.: Interest-based self organization in group-structured P2P networks. In: Proceedings of the 6th IEEE Consumer Communications and Networking Conference, pp. 1–5 (2009)
6. Wang, Li, Hu, G.-X.: P2P semantic community model based on interest and trust evaluation. Comput. Eng. **35**(13), 11–13 (2009)
7. Garbacki, P., Epema, D.H.J., van Steen, M.: The design and evaluation of a self-organizing superpeer network. IEEE Trans. Comput. **59**(3), 317–331 (2010)
8. Liu, K., Bhaduri, K., Das, K., Nguyen, P., Kargupta, H.: Client-side web mining for community formation in peer-to-peer environments. SIGKDD Explorations **8**(2), 11–20 (2006)
9. Wang, Y.: Trust-based community formation in peer-to-peer file sharing networks. In: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2004), pp. 341–348. IEEE Computer Society, Beijing, 20–24 Sept 2004
10. Li, Y., Bandar, Z.A., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans. Knowl. Data Eng. **15**(4), 871–881 (2003)

Chapter 9
A RouterUpdate Method for Tor Anonymous Communication System

Tianbo Lu, Bing Xu, Shixian Du, Lingling Zhao and Xiaomeng Zhang

Abstract Among all the anonymous communication systems, Onion Routing is most widely used. Tor affords users with anonymous service in communication. It can be used in running anonymous web browsing and announcement, real-time communications, IRC, SSH and other TCP applications. After analyzing the source code of the Tor system, this paper introduced the network layout, working flow, the RouterUpdate method of establishing a virtual circuit and data sending or receiving in Tor system. The method helps Tor client using the relay nodes which have been stored to connect the internet of Tor without connecting the list server. Then the paper introduced a method to help the Tor client using all the applications whether or not using the SOCKS.

Keywords Anonymous communication · Onion routing · Tor · SOCKS · Directory server

9.1 Introduction

Information vulnerabilities gradually become a security hidden danger to people's life in net. Anonymous communication is an effective method to guard users' privacy, which can protect bilateral identities and communication relations from being obtained by attackers.

Based on MIX [1], an idea of multiple step objective route, transmitting data via many middle nodes is proposed. Onion routing is an efficient method in anonymous communication [2–4]. Tor, a second generation of onion routing, has been widely used [5–10]. Tor is used to keep watch on flow filter and sniff analysis,

T. Lu (✉) · B. Xu · S. Du · L. Zhao · X. Zhang
School of Software Engineering, Beijing University of Posts and Telecommunications, Beijing, China
e-mail: lutb@bupt.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_9,

making communication in overlay network composed of onion routers, and to realize anonymous external links, anonymous hide and so on. As Tor is a network of virtual passageways, it makes an anonymous foundation of a series of application. So people can share information in public networks without caring privacy may be violated. As Tor disperses user flow to many different places in internet, there is no single point that links a user and his destination. In this way, Tor helps decline risk of simple and high-level flow analysis.

The paper is organized as follows. Section 9.1 gives an overview of Tor. Section 9.2 presents the overall structure and the primary flow. Section 9.3 mainly shows the design and related algorithms of store routing nodes. Section 9.4 introduces the design and application about rewriting network function.

9.2 Overall Architecture

The procedure to build a virtual circuit is shown in [4]. Based on the introduction, let us make a summary of a primary Tor process.

In an anonymous system, the user who wants to hide his identity should start OP process, which is responsible to build communication links and encrypt and decipher data. It obtains node information from Directory Server and selects one from the set of Tor nodes to consult secret keys, and last it builds a safe information channel with the former node. The building process conforms to short-term Diffe-Hellman secret key exchange protocol, as well as the TLS which guards the privacy of information channels and the security of information retransmission further. Then all data should be transmitted in this channel. Next OP goes on expanding to other Tor nodes via the built channel and exchanging secret key to build a multilayer encrypted channel. Data should be encrypted according to the order of Tor nodes from the later layer to the former one. During the transmission process, every time encrypted data passes a Tor node, it will be deciphered. It won't stop until the data of the last node has been transmitted to the destination. In the process data goes back from destination, it will be encrypted one time when passing a Tor node, deciphered when arriving at OP and transmitted to the application program finally. As each Tor node only has its own encrypting and deciphering keys, external attackers and Tor cooperators won't obtain the plaintext of the communication data only if they could obtain secret key of all nodes in the route.

9.3 RouterUpdate Method Based on Voting Mechanism

As the available onion router list and the neighboring network information are obtained from the directory server, when a directory server cannot be connected for some reason, the Tor OP cannot get node information and it would be difficult

to connect to Tor network, thus losing the ability for anonymous communication. This has been the defects of Tor against certain network firewall.

9.3.1 RouterUpdate Method Design

The RouterUpdate (RU) method is divided into several steps.

- The routerlist initialization and load.

Loading part of this method is after the Tor initialization and before read the network information document on the directory server. After Tor OP sends request information to the directory server, if it is unable to obtain the document, the method will automatically select the network and routerlist in the loading document. After that, Tor OP can create virtual circuit through the routerlist.

- Analyze and collect the ORs in the current Tor network, dynamically selecting the optimized OR and adding into the pre-defined global list.

By checking the current storage router node information, the status of the Tor network relay nodes in the records will be inspected. The Tor relay node with better efficiency and performance will be elected by screening algorithm of RU method. Then, the nodes filtered out will be stored in the temporary routerlist.

- The third step is periodically writing a custom OR global list into the hard disk, and optimizing the store files.

In the main loop of the system, set a timer and counter. When the timer times out, RU will automatically write data to disk and counter will be plus one. When the number of writes in counter records is greater than 3, the counter will be cleared automatically, and store files in the disk will be optimized.

9.3.2 Custom Routerlist Loading

- Loading Point Processing

In the Tor system design, there is a pattern for the load of the routerlist. A typical loading process occurs in the initialization process, as follows:

(1) The system will try to load the network status recently used. (2) After an available network status document is obtained, the Tor client will load the OR indicated in the document into the client's routerlist after some necessary checks, and check the available nodes and download the RC of some OR if needed. (3) After the nodes loaded is the creation of the virtual circuit and the subsequent operation.

- Disk Data Reading

Client will access a file with the similar format of the default store file on disk. The default file name is "cached-best Routers". At this point, the client needs to define a metadata for store files operation to save the corresponding document, and to associate the router lists in files and programs. In this example, the client will define it as:

desc_store_t
fname_base: holds the name of the description file
fname_alt_base: holds the name of the backup file
mmap: point to the file data entry
description: This document textual description
type: the type of document
journal_len: the length of the document log
store_len: the length of the document.

- File Reading

Reading the file occurs during the router list initialization. RU method can read the file when the routing list is not assigned. The client will connect the global list and the predefined store file by setting desc_store_t during the initialization. Note that the initialization of the list and desc_store_t is carried out simultaneously.

9.3.3 Real-time Data Collection

- Store in the main loop

The Tor client will periodically maintain some everyday events. Every time unit, Tor will run some detection and maintenance work. Collection of router state is exactly suitable for periodic operation, so the client will add this feature to run_scheduled_events with the daily maintenance function. This program will be called once at intervals of 1 s.

- Collection standards

On the basis of the optimized node located, the client will maintain a smartlist with MAX_NUM_BESTNODES size to record the ORs which have the max bandwidth in all ORs, dynamically stores in the temporary router list. Bandwidth here is the average of known bandwidth and maximum acceptable bandwidth.

9.3.4 Writing into Disk

Each time after the collection and analysis of data, and every standard time passed, it will write standards-compliant nodes to disk.

9.3.5 Optimizing File

Because the client has been writing router information, the store file is bound to storage repeated even invalid router information. When loaded it will result in the router list bloated and inefficient. System optimizes the router information every three time to write, and every time system load the file, and automatically optimizes the routing information.

9.4 Expand Tor SCOKS Agency

Some application programs which do not support SOCKS can't use the anonymous service afforded by Tor and other services. The method of "Extend for Tor" (EFT) in this paper aims to allow those application programs to access the network by Tor as agency, but won't alter those programs.

9.4.1 The Design of EFT Method

EFT method can modify dynamically linked libraries by setting the LD_PRELOAD environment variable, so that make it point to custom library of EFT method. If resetting default dynamically linked libraries at exit, this can run Tor, at the same time, and can let it automatically load in each progress space of executable program.

In the library file, this module rewrite normal connect function and select function, as the Fig. 9.1 shows.

- About the process of establishing connection

EFT method make an application automatically call connect function which is defined by module when building TCP connection, so as to submit right of control to Tor OP, rather than rock-bottom protocol. Later, application can firstly establish connection with Tor OP, and then establish connection with destination address.

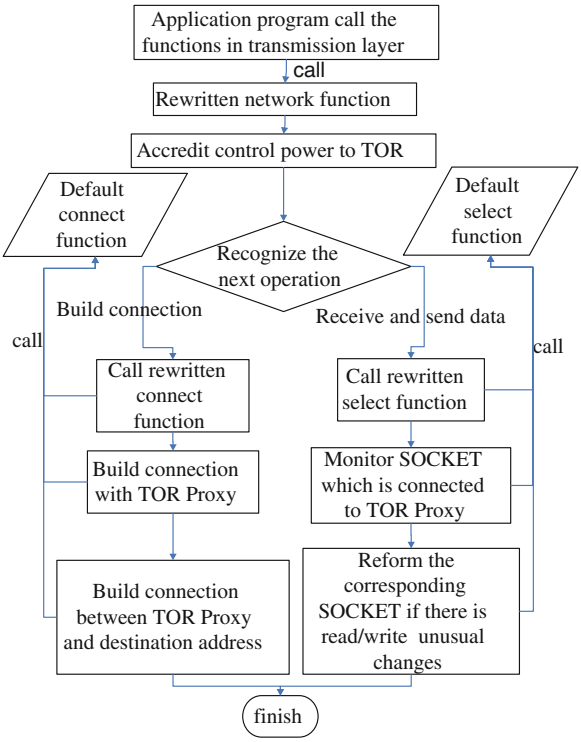


Fig. 9.1 Application call network function

- About the process of data transmission

This method can make rewritten select function inform Tor OP at first and submit the right of control to Tor when the state of data flow which is concerned with Tor system, instead of previous SOCKS. So the data flow is forced to go through Tor OP.

9.4.2 The Implementation Process of EFT Method

- Set range of application of Tor OP

Tor client need to be able to distinguish local network from external network. User can declare local address gateway and external address gateway by himself. By default, local address is localhost, others are external address. In addition, User can also declare corresponding address to use Tor OP by himself. And we can declare it in the file of PROXYconfig (Fig. 9.2).

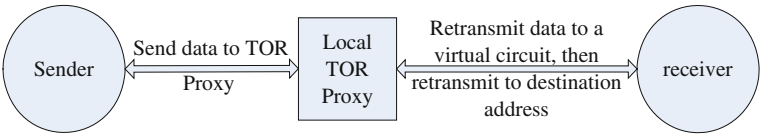


Fig. 9.2 The transmission process through Tor OP

- Data sending and receiving

The basic strategy of data sending is shown in the above figure. Tor OP plays a role in connecting the sender and receiver.

Under the request of receiving data, if client may use the disposal way of no blocking, in turn, the module should rewrite select function.

- About select function

It will achieve the monitoring of related SOCKET. When SOCKET need to read and write, module will know it by its defined select function and submit the message to the corresponding function to deal with. Notice that data firstly send to the Tor OP, then according to the destination address, Tor OP send it through virtual circuit.

9.5 Conclusion

Based on the study of source code of Tor system, the article analyzes anonymous communication system of Tor and its principle and introduces the design and application about rewriting network function. As it is difficult for Tor users to do further work in the condition that it could not be connected to Directory Server, this paper suggests collecting network data in User and sponsoring the connection project autonomously. However, there are still some points to be improved, such as information collection of User, collection algorithm and security of Tor system.

Acknowledgments This work is supported by the following programs: the National Natural Science Foundation of China under Grant No. 61170273; Research Innovation Program for Young People of China Beijing University of Posts and Telecommunications with title “Code Security Assurance”; 2010 Information Security Program of China National Development and Reform Commission with the title “Testing Usability and Security of Network Service Software”.

References

1. Chaum, D.: Untraceable electronic mail, return addresses, and digital pseudonyms. Commun. ACM **24**(2), 84–88 (1981)
2. Goldschlag, D.M., Reed, M.G.. Syverson, P.F.: Hiding routing information. In: Proceedings of Information Hiding, vol. 1174, pp. 137–150 (1996)

3. Goldschlag, D.M., Reed, M.G., Syverson, P.F.: Onion routing for anonymous and private internet connections. Commun. ACM. **42**(2), 39–41 (1999)
4. Reed, M.G., Syverson, P.F., Goldschlag, D.M.: Anonymous connections and onion routing. IEEE J. Sel. Areas Commun. **16**(4), 482–494 (1998)
5. Dingledine, R., Mathewson, N., Syverson, P.: Tor: the second-generation onion router. In: Proceedings of USENIX Security Symposium, pp. 21–21 (2004)
6. Edman, M., Syverson, P.F.: AS-awareness in Tor path selection. In: Proceedings of the 2009 ACM Conference on Computer and Communications Security, pp. 380–389 (2009)
7. Evans, N., Dingledine, R., Grothoff, C.: A practical congestion attack on tor using long paths. In: Proceedings of the 18th USENIX Security Symposium (2009)
8. McLachlan, J., Tran, A., Hopper, N., Kim, Y.: Scalable onion routing with Torsk. In: Proceedings of CCS, pp. 590–599 (2009)
9. Jansen, R., Hopper, N., Kim, Y.: Recruiting new tor relays with BRAIDS. In: Proceedings of the 2010 ACM Conference on Computer and Communications Security, pp. 319–328 (2010)
10. Tang, C., Goldberg, I.: An improved algorithm for Tor circuit scheduling. In: Proceedings of the 2010 ACM Conference on Computer and Communications Security, pp. 329–339 (2010)

Chapter 10
Management of Construction Schedules Based on Building Information Modeling Technology

Lianying Zhang, Xiang Zhang and Teng Ma

Abstract As construction projects becoming increasingly large and complex, the traditional methods of schedule management largely undermine the improvement of management level. However, the integration of Building Information Modeling (BIM) and scheduling information can help maintain control of scheduling goals and enhance project performance. This paper proposes a BIM-based construction schedule management framework and establishes a model to integrate scheduling information in life cycle. Moreover, information retrieval and integration and core supports for realizing the model are also examined. The study extends the existing research of construction schedule management, and can be used as guidance for BIM-based schedule management practice.

Keywords BIM · Construction management · Schedule management · Scheduling information management model

10.1 Introduction

The traditional methods of construction schedule management have several problems that hamper the collaboration between project participants and the maintenance of project schedules. These problems include dispersed and inaccurate project information, inefficiency of rearranging the schedules and low-level visualization of the schedule management system [1]. Moreover, as construction projects becoming increasingly large and complex, the industry demands for more efficient schedule management system that facilitates successful accomplishment of projects at a minimum waste of resources. Many previous studies recommend

L. Zhang (✉) · X. Zhang · T. Ma
College of Management and Economics, Tianjin University, Tianjin, China
e-mail: tjzly126@126.comzhanglianying@tju.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_10,

that Building Information Modeling (BIM) technology can be used as an efficient tool of providing visual information such as design parameters, project data and 3D model [2]. Yet, the studies regarding integration of project schedule to BIM model fail to view the system in the perspective of project life cycle. Therefore, this paper aims to propose a BIM-based schedule management system that considers life cycle of project. In addition, information retrieval and integration in BIM-based schedule planning and the core supports for BIM-based schedule management are also discussed in the paper.

10.2 BIM and BIM-Based Scheduling

According to Eastman et al. (2008), BIM is a process that define objects parametrically, and when related object changes the parameters also changes in accordance with the rules embedded in them [2]. And Kymmell (2008) defines BIM as a project and process simulation that allows making adaptations of the simulation parameters in a virtual environment that contains all the information required [3].

A number of studies indicate that BIM can considerably boost construction performance. Suermann (2009)'s research indicates that the implementation of BIM can substantially improve management of project schedule and quality [4]. Zuppa et al. (2009) also suggest the positive impact of BIM on project's schedule, quality, and cost [5]. Aslani and Chiarelli (2009) describe the advantages and beneficial of BIM, and they emphasize that for contractors, BIM facilitate tracking and managing changes, preparing for schedules and estimates [6]. Korman et al. (2008) conduct case studies to show that effectively using BIM requires integrating extra knowledge rather than resolving physical problems [7].

Regarding the integration of construction scheduling to BIM, many different approaches are found in previous academic studies and industry practices [8–12]. Ospina-Alvarado and Castro-Lacouture (2010) find that the existing literature with respect to the use BIM for scheduling purposes can be classified into "use of the model to generate the schedule as part of BIM" and "Link of the model to an external schedule for visualization" [13]. Other researches explore how to realize BIM-based 4D management model, especially integrate schedule information to BIM. Tse et al. (2005) propose a method to model objects and interfaces in BIM [14]. Fu et al. (2006) research on integrating information such schedule, cost, project, risk and energy saving, and provide a blueprint of the development of nD models in the future [15]. Zhang and Wang (2003) develop a 4D-MCPRU project management system, and realize the link between AutoCAD and Microsoft Project scheduling in both directions [16]. Nepal et al. (2009) propose several approaches for querying information from IFC-based BIM model [17]. Weise et al. (2009) develop a 4D simulation package called "scheduling assistant" which allows to import complete IFC models regarding 4D information, and apply it to Microsoft Project as a plug-in using IFC-interface [18].

10.3 BIM-Based Schedule Management System

10.3.1 The Establishment of BIM-Based Schedule Management System Framework

The management and control of schedule is one of the essential works of entire construction project management. The traditional schedule management system mainly relies on manual operation. The main problems exist in this system are: the overall system design concept is vague; the scheduling information is poor in visualization, availability, timeliness and accuracy; and unfavorable to system self-organization and self-running. This paper propose using Autodesk Revit to acquire available data to establish the BIM-based schedule management system in a pattern of 'Model-View-Controller' [19], as shown in Fig. 10.1.

Realization of Model layer: using Autodesk Revit to build 3D BIM, and export the graphics data in the model to DWG format graphics files. Then, Autodesk Revit and API are used to assist secondary development to export property data of the 3D model to a SQL Server database, and connect the graphic data and the property data one-on-one through producing distinctive building components ID. Meanwhile, independent construction schedule setting modules are developed to store construction scheduling data stored in the SQL Server database.

Realization of View layer: Displaying construction schedules based on Autodesk DWG Design Review, and display specific information of building

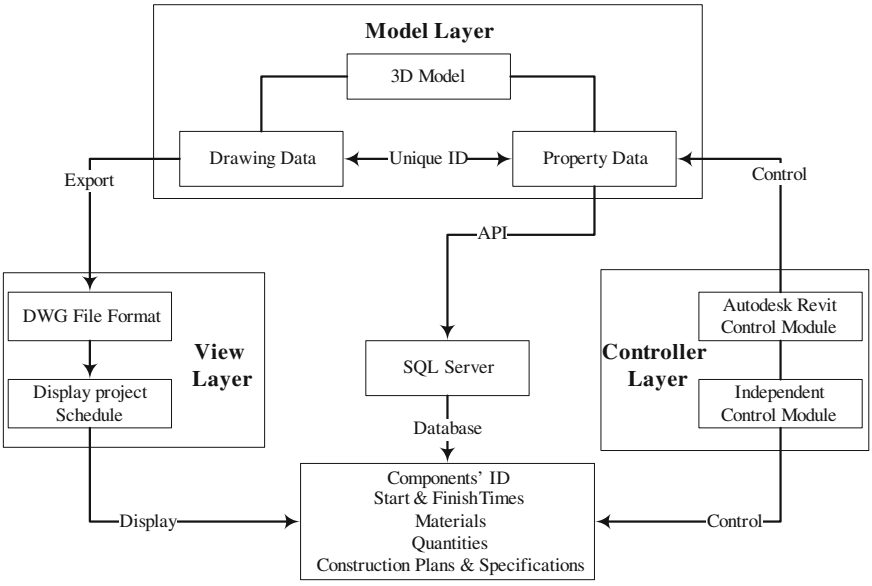


Fig. 10.1 BIM-based project schedule management system

component through property sheet control, for example, components' ID, start and finish times, materials, quantities and construction plans and specifications.

Realization of Controller layer: Development of independent construction schedule control modules and control current duration. Then compare start and finish times of every building component, and judge whether the setting component should be displayed. To control display or hidden state of every components and send messages to the View layer.

When construction scheduling information has been input into the 3D BIM, property information of a building component in the graphic module of Autodesk Revit will be obtained such as duration, labor, and resources. If one building component is selected in the View window, the graph of the component will be highlighted, and corresponding property bar in building component properties toolbar will also be highlighted, therefore construction procedure of the component are easy to be controlled. And if construction schedules are found to be different with the expectations, we can adjust the durations of related objects and current construction state, the system will automatically update the database, and refresh the 4D model, and it is convenient to check the current duration and schedule, and facilitate construction management.

10.3.2 BIM-Based Scheduling Information Management System

The establishment of basic framework provides architectural support for BIM-based schedule management system, yet in order to effectively implement it into construction schedule plan and daily activities, a powerful scheduling information management system is required.

The management of construction project information involves many participants, such as owner, designer, contractor, supplier, operator, government, and financial institution. There are huge amount of information, and the exchange of information are complicated, and the traditional way of information management is low efficient and arrange in disorder. To transfer construction scheduling information between all participants smoothly, and to be available for management departments, a project integrated control system must be built. This system is BIM-based platform, and it views project activities as basic objects based on computer network, and construct scheduling information processing platform that is under collaboratively management and control. Thus, the scheduling information can be shared and collaboratively managed across different departments, corporations and areas. So the key to construct BIM-based schedule management system model is to change the traditional way of information transformation and sharing, and integrate scheduling information of different project phases and participants effectively in the purpose of realizing the project information management in full life cycle.

To establish the BIM-based schedule management system, the difficulties are the creation, management, and sharing of BIM information. Currently, the main method to store BIM information to build data storage center based on IFC standards, and allow visit and revise of distributed, heterogeneous application systems to realize information integration.

This paper proposes creating BIM-based information management model in all project phrases by using information sub-models as kernels. The basic principle of the idea is to create information sub-models in all project phase, i.e. project planning phase, project design phase, construction phase, operation phase, in according to the need of the project management. Every information sub-model evolves automatically, and can retrieve, extend, and integrate data from the sub-model of previous phrase, and then create the sub-model of the present phrase. As the project continues, scheduling information model in full life cycle is created, as shown in Fig. 10.2.

From the project planning phase to the project design phase, then to the construction phase and operation phase, the project schedule is integrated step-by-step, and in the end, the complete project schedule is formed. In every project phase, the system will define information exchange model of the specific

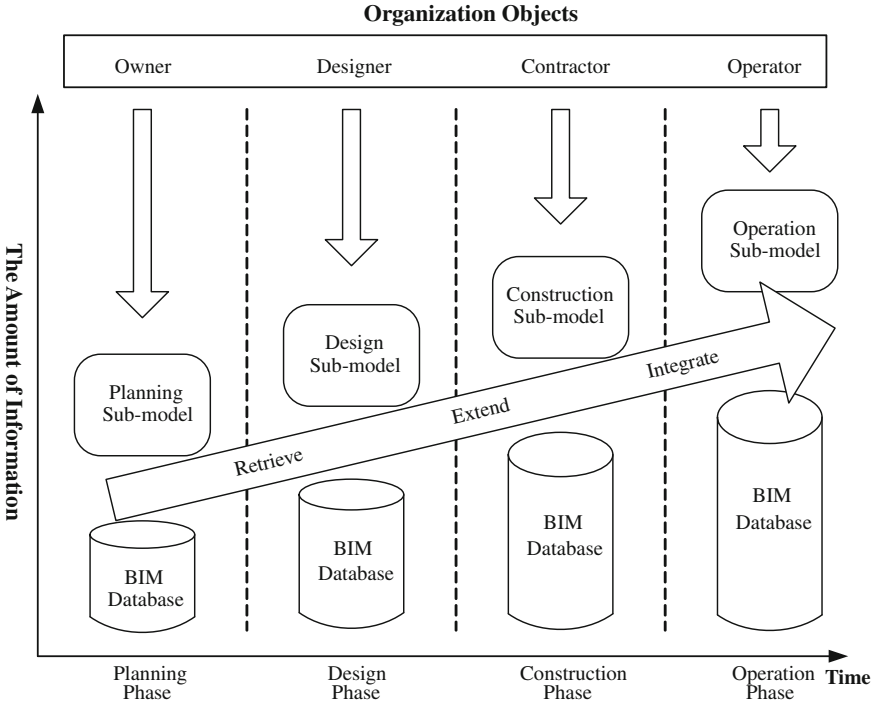


Fig. 10.2 The evolvement of BIM-based schedule management sub-models

application in that phase, and realize data integration and sharing by retrieving and integrating information sub-models.

10.3.3 Information Retrieval and Integration in BIM-Based Schedule Planning

The data in construction project schedule reflects durations of building objects and their overlapping relationships, and describe the state of the 3D model. For project schedule management, that data is the base of the 4D construction information model. Therefore, the organization and management of construction schedule data is one of the important steps of BIM information management system.

Construction site management is a complex and dynamic process with many unexpected changes, which contains huge amount of data of different kinds. And as the progress of the project, the amount of data is continuously accumulated, so high demand is required to manage those data in BIM information management system. Moreover, construction project is a complex system of high integration, and it involves different departments, types of work, resources, labor. Although the sources of data are complicated, every participant needs to obtain and share these data. After considering the above factors, we propose using SQL Server to store and retrieve schedule management information in database.

Data transfer and function calls between schedule and SQL Server database are programming by C language. In the process of project implementation, schedule data are collected, they are entered into Microsoft Project that automatically track the progress of project, and obtain usage information about schedule, resources and cost. The follow-up task duration, material, and labor are adjusted according to the situation at that time, and the updated schedule will feedback to the database of SQL Server.

10.3.4 Core Supports for BIM-Based Schedule Management

Software and basic technology that are used to load scheduling information in BIM model include: Autodesk Revit (3D Architecture, Structure and MEP modeling tool), Microsoft Project (project schedule planning tool), SQL Server (key development technologies for realizing schedule management system), IFC standard (realizing BIM data exchange and sharing).

In this paper, we propose using Microsoft Project to create project schedule management file. Project scheduling information are stored in database of SQL Server by using C language programming and API of Microsoft Project, and at the same time, scheduling information of standard file format in SQL Server are imported to Autodesk Revit by using API of Autodesk Revit. Moreover, BIM sub-

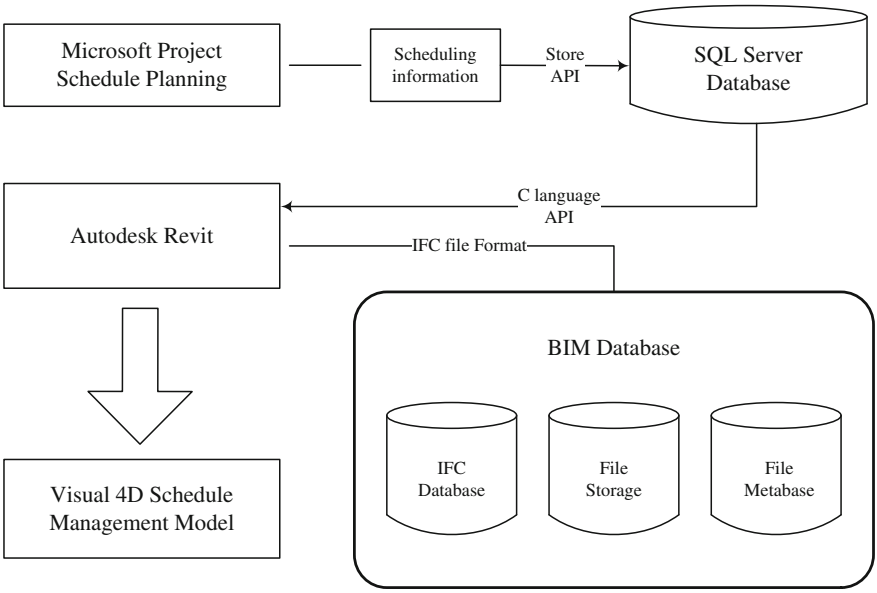


Fig. 10.3 Core supports and realization of 4D model

models data are imported to Autodesk Revit in IFC file format, and their property information are loaded to the schedule response nodes, so that BIM-based 4D project schedule model is established, as shown in Fig. 10.3.

10.4 Conclusion

Previous researches about project schedule and BIM model integration are unable to view the system in the perspective of project life cycle. This paper proposes using a BIM-based schedule management model that considers life cycle of project. Researchers try to establish BIM-based schedule management system framework by using a 'Model-View-Controller' pattern. It is the extension of the current research of using BIM for scheduling purposes. However, the real application of the model still needs to be further studied.

The application of BIM-based schedule management demonstrates its unique advantages and benefits, and it can substantially improve traditional schedule management methods. Although the application of BIM technologies in China has just started, its application is bound to have a profound impact on the construction industry.

Acknowledgments The authors wish to acknowledge the support and funding for this research provided by National Science Foundation of China (NSFC, Project No. 71272146).

References

1. Sullivan, C.C.: Best practices in integrated project delivery for overall improved service delivery management . McGraw-Hill Construction-Continuing Education Program (2009)
2. Eastman, C.M., Teicholz, P., Sacks, R., et al.: BIM handbook: A guide to building information modeling for owners, managers, architects, engineers, contractors, and fabricators. Wiley, Hoboken (2008)
3. Kymmell, W.: Building information modeling: planning and managing construction projects with 4D CAD and simulations. McGraw-Hill Professional (2008)
4. Suermann, P.C.: Evaluating the impact of building information modeling (BIM) on construction. University of Florida (2009)
5. Zuppa, D., Issa, R.R.A., Suermann, P.C.: BIM's impact on the success measures of construction projects. Technical Council on Computing and Information Technology of ASCE Reston, VA (2009)
6. Aslani P, Chiarelli L.: Building information model: the role and need of the constructors. ASCE (2009)
7. Korman, T.M., Simonian, L., Speidel, E.: Using building information modeling to improve the mechanical, electrical, and plumbing coordination process for buildings (2008)
8. Lu, N., Korman, T.: Implementation of building information modeling (BIM) in modular construction: Benefits and challenges. Construction Institute of ASCE Reston, VA (2010)
9. Sacks, R., Koskela, L., Dave, B.A., et al.: Interaction of lean and building information modeling in construction. J. Constr. Eng. Manage. **136**(9), 968–980 (2010)
10. Kuprenas, J.A., Mock, C.S.: Collaborative BIM modeling case study—process and results. ASCE (2009)
11. Goedert, J.D., Meadati, P.: Integrating construction process documentation into building information modeling. J. Constr. Eng. Manage. **134**(7), 509–516 (2008)
12. Tse, K.T.C., Wong, A.K.D., Wong, F.K.W:. Modeling objects and interfaces in building information modeling. ASCE (2009)
13. Ospina-Alvarado, A.M., Castro-Lacouture, D.: Interaction of processes and phases in project scheduling using BIM for A/E/C/FM integration (2010)
14. Tse, K.T.C., Wong, A.K.D., Wong, F.K.W.: Modeling objects and interfaces in building information modeling. ASCE (2005)
15. Fu, C., Aouad, G., Lee, A., et al.: IFC model viewer to support nD model application. Autom. Constr. **15**(2), 178–185 (2006)
16. Zhang, J.P., Wang, H.: A 4D++ site model and 4D management system for construction projects. China Civil Eng. J. **36**(3), 70–78 (2003) (In Chinese)
17. Nepal, M.P., Zhang, J., Webster, A., et al.: Querying IFC-based building information models to support construction management functions (2009)
18. Weise, M., Liebich, T., Tulke, J., et al.: IFC support for model-based scheduling. CiBW78 (2009)
19. Zhang, Y.: Research on BIM-based building information integration and management. Tsinghua University, Beijing (2009). (In Chinese)

Chapter 11
P-Hub Airline Network Design Incorporating Interaction Between Elastic Demand and Network Structure

Lie Han and Ning Zhang

Abstract This paper innovates the p-hub airline network median design method. Researchers present a new mathematical programming model, which incorporates the interaction between elastic demand in air passenger market and airline network structure. The model optimizes both the ticket prices and the profit of airline company, and subsequently determines the passenger volume influenced by different network structure. The effectiveness and practicability of the model are demonstrated by a realistic example of Chinese airline network which includes 15 major airports. Numerical analysis result indicates that hub locations tend to select the airports which have bigger passenger volume.

Keywords Traffic planning · Hub-and-spoke network · Elastic demand · Profit maximization

11.1 Introduction

The hub-and-spoke (HS) airline network, which includes hub and non-hub airports simultaneously, has already become major operational mode in mature aviation markets of developed countries. Hubs are special facilities that serve as switching, transshipment and sorting. When we design a HS airline network, we need to choose a fixed number P hub from all airports and allocate the remaining airports to these hubs. This design problem is known as p-hub median problem. The

L. Han (✉) · N. Zhang
School of Economics and Management, Beijing University of Aeronautics and Astronautics, Beijing, China
e-mail: han_lie@163.com

N. Zhang
e-mail: zhangng@vip.sina.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_11,

research of p-hub began with the pioneering work of O'Kelly, which gave a programming formulation of the single allocation p-hub median problem [1]. Campbell formulated the multiple allocation p-hub median problems firstly as a linear integer programming [2]. Skorin-Kapov et al. demonstrated that the LP relaxation of Campbell formulation leads to highly fractional solutions [3]. Mingguo Bai et al. developed an appropriate attribute index system to select spare hubs, and then applied the shortest path algorithm to design the HS network of fifteen Chinese airports [4]. Carello et al. investigated the cost of installing routes on the edge [5]. Yaman studied a problem which she named the uncapacitated hub location problem with modular arc capacities [6]. Yaman et al. investigated the capacitated version of problem, which the capacity of a hub is defined as the amount of traffic passing through the hubs [7].

These researches all attempted to find the best network design which has minimal total cost, based on the hypothesis that the volume of travelers were fixed. They have not considered that HS networks are absolutely different from traditional point-to-point (PP) networks in the transportation cost, flying routes, travel time and route distance. These differences will certainly affect the amount of travelers, which contradicts with the hypothesis of fixed number of travelers. For the reason to remedy this critical drawback, an original mathematic optimization model of p-hub median problem has been developed in this paper, which considers the interaction between elastic demand and network structure. The effectiveness and practicability of this model is proved by an example constructing a HS airline network containing 15 major airports in China.

11.2 Basic Assumptions of Airline Markets

In the airline market, every O-D pair is an independent submarket. The amount of travelers are dependent on the total travel cost including ticket price and travel time cost. Without loss of generality, we assume that the amount of travelers can be calculated by

$$q_{ij} = \theta_{ij} + \lambda_{ij}(p_{ij} + t_{ij}), \tag{11.1}$$

where q_{ij} is the amount of travelers in the submarket between airport i and j (ij submarket). λ_{ij} is the elastic demand coefficient, and $\lambda_{ij} < 0$. θ_{ij} is the number of possible customers in ij submarket. p_{ij} is the ticket price between airport i and j, t_{ij} is the travel time cost. Let π_{ij} denote the profit of airline company in ij submarket, which can be expressed by

$$\pi_{ij} = (p_{ij} - d_{\partial ij})q_{ij}, \tag{11.2}$$

where d_{ij} is the straight distance between airport i and j. Because the cost of transporting, a single traveler is correlative with the length of travel route. So d_{ij}

can measure the cost of transporting a single traveler. And p_{ij} is measured by unit of length, either. When $t_{ij} + c_{ij} \geq -\frac{\theta_{ij}}{\lambda_{ij}}$ in ij submarket, the airline company cannot obtain any profit from this submarket, and they will abandon this submarket. In this case, $\pi_{ij} = p_{ij} = q_{ij} = 0$.

11.3 Parameters of the Point-to-point Airline Network

We use superscript 0 to denote PP network. The time cost in PP network t_{ij}^0 can be expressed by

$$t_{ij}^0 = \gamma d_{ij}, \tag{11.3}$$

where γ is the translation coefficient between time cost and travel length. If there are some travelers in ij submarket, namely $q_{ij}^0 > 0$, then substituting Eqs. (11.1), (11.3) into Eq. (11.2) yields

$$\pi_{ij}^0 = \left(p_{ij}^0 - d_{ij}\right)\left(\theta_{ij} + \lambda_{ij}\left(\pi p_{ij}^0 + \gamma d_{ij}\right)\right), \tag{11.4}$$

Solving the maximization of Eq. (11.4) with variable p_{ij}^0, i.e., the first-order condition $\partial\pi_{ij}^0/\partial p_{ij}^0 = 0$, yields that when the profit reaches its peak value, we have

$$p_{ij}^0 = d_{ij} - \frac{q_{ij}^0}{\lambda_{ij}}, \tag{11.5}$$

And substituting Eq. (11.5) into Eq. (11.2), the maximal profit for the airline company is

$$\pi_{ij}^0 = -\frac{1}{\lambda_{ij}}(q_{ij}^0)^2, \tag{11.6}$$

In the case of travel demand in ij submarket $q_{ij}^0 = 0$, we can simply assume that there is not any potential customers in this submarket, i.e. $\theta_{ij} = 0$. Substituting Eqs. (11.5), (11.6) into Eq. (11.4) and rewriting the equation, parameter θ_{ij} can be expressed by

$$\theta_{ij} = \begin{cases} 2q_{ij}^0 - \lambda_{ij}(\gamma + 1)d_{ij}, & if \quad q_{ij}^0 > 0, \\ 0, & if \quad q_{ij}^0 = 0, \end{cases} \tag{11.7}$$

.

The total profit of whole PP airline network π^0 can be calculated by

$$\pi^0 = \sum_i \sum_j \pi_{ij}^0 = -\sum_i \sum_j \frac{1}{\lambda_{ij}} (q_{ij}^0)^2, \tag{11.8}$$

11.4 Parameters of the Hub-and-Spoke Airline Network

We use superscript 1 to denote HS networks and we adopt the strict uncapacitated multiple allocation HS network structure. There are n airports in which the set of origins, destinations and potential hub locations are identified.

Because the HS network concentrates the traveler flow through the hubs, which generates the economy of scale, there are the cost discounts in HS network. $\alpha(0 < \alpha \leq 1)$ is the discount factor between hubs, and $\beta(\alpha \leq \beta \leq 1)$ is the discount factor between non-hub and hub. As general setting of p-hub research, α and β are known as parameters. The transportation cost in HS networks is

$$c_{ijkm}^1 = \beta d_{ik} + \alpha d_{km} + \beta d_{mj} \tag{11.9}$$

The time cost of each traveler in HS networks is

$$t_{ijkm}^1 = \gamma \left(d_{ik} + d_{km} + d_{mj}\right) \tag{11.10}$$

In the HS networks, travelers of ij submarket transship at hub k and m. Combining Eqs. (11.1), (11.2), the airline company's profit in the ij submarket π_{ijkm}^1 is expressed by

$$\pi_{ijkm}^1 = \left(p_{ijkm}^1 - c_{ijkm}^1\right) q_{ijkm}^1 = \left(p_{ijkm}^1 - c_{ijkm}^1\right) \left(\theta_{ij} + \lambda_{ij} \left(p_{ijkm}^1 + t_{ijkm}^1\right)\right), \tag{11.11}$$

Solving the maximization of Eq. (11.11) with variable p_{ijkm}^1, i.e., the first-order condition $\partial \pi_{ijkm}^1 / \partial p_{ijkm}^1 = 0$, we obtains that when the profit reaches its peak value, the ticket prices are calculated by

$$p_{ijkm}^1 = \begin{cases} -\frac{\theta_{ij}}{2\lambda_{ij}} - \frac{1}{2} t_{ijkm}^1 + \frac{1}{2} c_{ijkm}^1, & if \quad c_{ijkm}^1 + t_{ijkm}^1 < -\frac{\theta_{ij}}{\lambda_{ij}} \\ 0, & if \quad c_{ijkm}^1 + t_{ijkm}^1 \geq -\frac{\theta_{ij}}{\lambda_{ij}} \end{cases} \tag{11.12}$$

Substituting Eqs. (11.9), (11.10), (11.12) into Eq. (11.1) yields that the amount of travelers is calculated by

$$q_{ijkm}^1 = \begin{cases} \frac{\theta_{ij}}{2} + \frac{\lambda_{ij}}{2} t_{ijkm}^1 + \frac{\lambda_{ij}}{2} c_{ijkm}^1, & if \quad c_{ijkm}^1 + t_{ijkm}^1 < -\frac{\theta_{ij}}{\lambda_{ij}} \\ 0, & if \quad c_{ijkm}^1 + t_{ijkm}^1 \geq -\frac{\theta_{ij}}{\lambda_{ij}} \end{cases} \tag{11.13}$$

11.5 Model and Numerical Example

On the basis of above analysis, we provide our mathematic optimization model for designing the HS network:

$$\max \quad \pi^1 = \sum_{i\in N}\sum_{j\in N}\sum_{k\in N}\sum_{m\in N} X_{ijkm}\left(p^1_{ijkm} - c^1_{ijkm}\right)q^1_{ijkm}, \tag{11.14}$$

$$s.t. \quad \sum_{k\in N} H_k = P, \tag{11.15}$$

$$\sum_{k\in N}\sum_{m\in N} X_{ijkm} = 1, \quad \forall i,j \in N, \tag{11.16}$$

$$X_{ijkm} \le H_k, \quad \forall i,j,k,m \in N, \tag{11.17}$$

$$X_{ijkm} \le H_m, \quad \forall i,j,k,m \in N, \tag{11.18}$$

$$H_k \in \{0,1\}, \quad \forall k \in N, \tag{11.19}$$

$$X_{ijkm} \in \{0,1\}, \quad \forall i,j,k,m \in N. \tag{11.20}$$

where $N = \{1,2,3,\ldots,n\}$, is the set of all n airports in the airline network. $p^1_{ijkm}, c^1_{ijkm}, q^1_{ijkm}$ in Eq. (11.14) are calculated by Eqs. (11.9), (11.10), (11.12), (11.13). H_k is a $0-1$ variable, when node k is a hub airport, $H_k = 1$, otherwise $H_k = 0$.X_{ijkm} is a $0-1$ variable, when the flying route from node i to node j need to transship at hub k and m, $X_{ijkm} = 1$, otherwise $X_{ijkm} = 0$. According to the features of the air passenger transportation, we limit that transshipment times are less than or equal to 2.

The objective function (11.14) is the total profit of the HS network. Constraint (15) ensures the number of hubs is P. Constraint (11.16) ensures that there is only one flying route between airport i and j. Constraint (11.17) and (11.18) ensure all flying routes transship at hubs only. Constraint (11.19), (11.20) are $0-1$ variable constraints.

To prove the effectiveness and practicability of this optimization model, we select 15 Chinese airports to design the HS airline network. The set of airports is quoted from the Ref. [4], which is (1) Beijing, (2) Shanghai, (3) Shenyang, (4) Zhengzhou, (5) Xi'an, (6) Wulumuqi, (7) Nanjing, (8) Hangzhou, (9) Changsha, (10) Wuhan, (11) Chengdu, (12) Guangzhou, (13) Haikou, (14) Kunming, (15) Xiamen. The model (11.14)–(11.20) is a NP-hard problem. We adopt the software GAMS to solve the model directly. The initial data of the traveler volume in point-to-point network q^0_{ij} and straight distances of each O–D pair d_{ij} is quoted from the Ref. [8]. We report the calculation results in Table 11.1.

We calculated several sets of results under a series of parameters to analyze the influence of different parameters. By observing the calculation results in Table 11.1, several principles are obviously revealed. Firstly, when parameter λ_{ij}

Table 11.1 The calculation results

P	α	β	γ	$\forall\lambda_{ij}$	Hub airports	π^0 (10^8 yuan)	π^1 (10^8 yuan)	π^1/π^0 (%)
1	0.8	0.9	0.1	−0.1	10	2611.87	2205.86	84.46
2	0.8	0.9	0.1	−0.1	1,9	2611.87	2518.97	96.44
3	0.8	0.9	0.1	−0.1	1,2,12	2611.87	2640.04	101.08
4	0.8	0.9	0.1	−0.1	1,8,11,12	2611.87	2781.12	106.48
5	0.8	0.9	0.1	−0.1	1,2,10,11,12	2611.87	2834.18	108.51
6	0.3	0.4	0.1	−0.1	10	2611.87	3598.85	137.79
7	0.3	0.4	0.1	−0.1	1,9	2611.87	3825.23	146.46
8	0.3	0.4	0.1	−0.1	1,4,12	2611.87	3876.71	148.43
9	0.3	0.4	0.1	−0.1	1,7,11,12	2611.87	4056.01	155.29
10	0.3	0.4	0.1	−0.1	1,2,10,11,12	2611.87	4112.73	157.46
11	0.8	0.9	0.1	−0.5	1	522.37	413.60	79.18
12	0.8	0.9	0.1	−0.5	1,12	522.37	580.54	111.14
13	0.8	0.9	0.1	−0.5	1,2,12	522.37	687.32	131.58
14	0.8	0.9	0.1	−0.5	1,2,11,12	522.37	759.59	145.41
15	0.8	0.9	0.1	−0.5	1,2,11,12,15	522.37	785.27	150.33

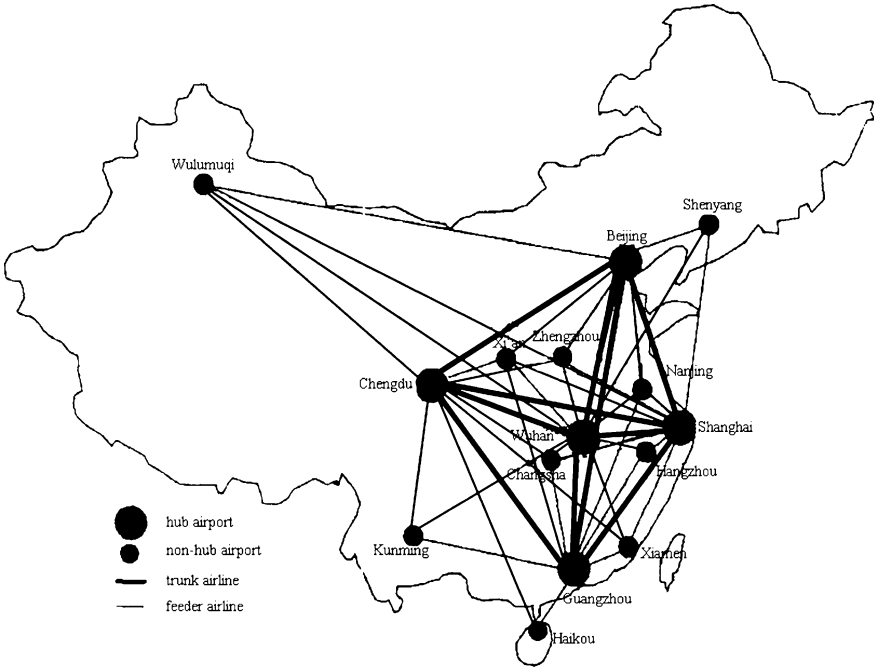


Fig. 11.1 The hub-and-spoke airline network of 15 Chinese cities

increases, namely, the elasticity of demand intensifies, the hubs tend to select the airports which have bigger amount of travelers. Because in this case, travelers are more sensitive to total cost. The hubs with more travelers may save more transportation costs. Secondly, when the elasticity of demand intensifies, the profit of HS network is remarkably bigger than PP network. Finally, when the discount factors become small, namely, the economy of scale becomes more obvious, the hubs tend to select the airports which positions are closer to geographic center of gravity. And reduced discount factor results in less total transport cost.

On the basis of calculation results, the best structure of airline network can be obtained. Figure 11.1 shows the design of Chinese airline network by above method with $P = 5$, $\alpha = 0.3$, $\beta = 0.4$, $\forall\lambda_{ij} = -0.1$, $\gamma = 0.1$. In this case, hubs are Beijing, Shanghai, Wuhan, Chengdu and Guangzhou.

11.6 Conclusion

Although HS airline networks take advantage of economies of scale and scope, foster hub airports, optimize resource of aviation industry, HS airline networks can also influence the transportation cost, flying routes and travel time, and further influence the airline company's profit and customer demand. When airline companies construct their hub-and-spoke networks, they must trade off various factors for maximizing the net profit.

This paper merges operational behavior of airline company and p-hub median problem together, presents a effective design method of HS network, which not only accords with the real environment, but also reflects the process of decision making. By applying 15 Chinese airports, researchers construct a HS airline network, which can provide reference for Chinese aeronautic transportation planning.

Acknowledgments This research is supported by the National Natural Science Foundation of China (70971003).

References

1. O'Kelly, M.E.: A quadratic integer program for the location of interacting hub facilities. Eur. J. Oper. Res. **32**(3), 393–404 (1987)
2. Campbell, J.F.: Hub location and the p-hub median problem. Oper. Res. **44**(6), 1–13 (1996)
3. Skorin-Kapov, D., Skorin-Kapov, J., O'Kelly, M.E.: Tight linear programming relaxations of uncapacitated p-hub median problems. Eur. J. Oper. Res. **94**(3), 582–593 (1996)
4. Bai, M.-G., Zhu, J,-F., Yao Y.: Design and application of hub-and-spoke network. Syst. Eng. **24**(5), 29–34 (2006) (In Chinese)
5. Carello, G., Carello, F., Ghirardi, M., Tadei, R.: Solving the hub location problem in telecommunication network design: A local search approach. Networks **44**(2), 94–105 (2004)
6. Yaman, H.: Polyhedral analysis for the uncapacitated hub location problem with modular arc capacities. SIAM J. Discret. Math. **19**(2), 501–522 (2005)

7. Yaman, H., Carello, G.: Solving the hub location problem with modular link capacities. Comput. Oper. Res. **32**(12), 3227–3245 (2005)
8. General Administration of Civil Aviation of China, Department of Planning and Development: Statistical data on civil aviation of china 201. Chinese civil aviation Press, Beijing (2010) (In Chinese)

Chapter 12
Aggregation Operators of Interval Grey Numbers and Their Use in Grey Multi-Attribute Decision-Making

Honghua Wu and Yong Mu

Abstract In this paper, authors propose a fast and efficient ranking method for interval grey numbers based on the idea of mean value and mean square deviation in statistics. If the degree of greyness of grey number is very small, the interval grey number is then big when the kernels of the interval grey numbers are equals. Authors extend data information weighted arithmetic averaging (WAA) operator, ordered weighted averaging (OWA) operator and hybrid weighted averaging operator (HWA) operator, meanwhile they propose interval grey numbers WAA operator, interval grey numbers OWA operator, and interval grey numbers HWA operator. According to these operators, authors develop an approach to solve grey multi-attribute multi-person decision-making problems, in which the attributive weights are completely known and the attributor values are interval grey numbers. Finally, an illustrative example is given.

Keywords Grey multi-attribute decision making · Interval grey number · Ranking method · Aggregation operators of interval grey numbers

12.1 Introduction

In 1982, Deng Julong proposed the theory of gray system [1], which has been widely used in modern society and various fields [2, 3]. In theory of gray system, the most basic element is interval grey number, and then the operation and ranking on interval grey number have long been touted [4–9]. At present, the research of interval grey number is mainly on the interval grey number operation axiom, algorithms and new grey algebraic system, while its application research is less

H. Wu (✉) · Y. Mu
School of Mathematical Sciences, University of Jinan, Jinan, China
e-mail: ss_muy@ujn.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_12,

involved. With the deepening of the research on the gray system theory and the expansion of its application scope, effective aggregation operators of interval grey numbers is becoming more and more important. It is necessary to discuss this problem because it cannot meet the actual need merely relying on the grey number basic algorithms.

Data information aggregation operator is an important research content in the modern information science and decision science, which has been widely used in decision analysis, fuzzy control, artificial intelligence, expert system, database system etc. [10–12]. WAA operator and OWA operator [13] are two kinds of common data information aggregation operators. The main difference between WAA operator and OWA operator is that the former weights on each data and then aggregates on the weighted data, while the later ranks the set of data first, and then weights and aggregates, whose weight is only based on the corresponding position. Because of the one-sidedness of WAA operator and OWA operator, Xu Zeshui proposed hybrid weighted averaging operator (HWA) [14] in 2003, which can not only take the importance degree of each data into consideration, but it can also reflect the data location importance degree. However, these operators are applied only to data and information with real expression conditions.

This paper is organized as follows: in Sect. 12.2, we propose a fast and efficient ranking method of interval grey numbers based on the ideal of mean value and mean square deviation in statistics. In Sect. 12.3, we extend data information WAA operator, OWA operator and HWA operator, meanwhile we propose interval grey numbers WAA operator, interval grey numbers OWA operator, interval grey numbers HWA operator. In Sect. 12.4, we develop an approach to solve grey multi-attribute multi-person decision–making problems. Finally, an illustrative example is given to demonstrate the feasibility and superiority for our ranking method and our approach in Sect. 12.5.

12.2 The Comparison of Interval Grey Number

The grey number is a certain interval or an uncertainty number under general count set in practical application. More formally, let "$\otimes$" be the grey number. Both lower bound and upper bound of grey number called interval grey number, written as $\otimes \in [a, b]$.

The ranking method for grey numbers is crucial for the uncertainty of grey decision-making. In the paper [7], Xie Naiming proposed a ranking method for grey numbers with their probability distributions. The ranking steps of three parameters interval grey numbers are set in paper [16]. However, there is a few studies of ranking method for interval grey numbers without knowing information of grey numbers at present.

In order to demonstrate our ranking method, we will introduce the concept related to grey number firstly.

Definition 1[3] Suppose interval grey number $\otimes \in [a,b], a<b$

(1) If $\otimes$ is a continuous grey number, so $\widehat{\otimes} = \frac{1}{2}(a+b)$ is kernel of the grey number.
(2) If $\otimes$ is a discrete grey number, $a_i \in [a,b](i=1,2,\cdots,n)$ are all possible values of grey number, so $\widehat{\otimes} = \frac{1}{n}\sum_{i=1}^{n} a_i$ is kernel of the grey number.

Definition 2[3] Suppose interval grey number $\otimes \in [a,b], a<b$ is random number having distribution information, so $\widehat{\otimes} = E(\otimes)$ is kernel of the grey number.

Definition 3[6] Suppose the background or field of interval grey number is $\Omega, \mu(\Omega)$ is the measurement for interval grey number, thus

$$g^0(\otimes) = \mu(\otimes)/\mu(\Omega)$$

is the degree of greyness of interval grey number, marked g^0.

By $\otimes \subset \Omega$ and the character of the measurement, the definition of 3 meet the standard, thus

$$0 \le g^0 \le 1$$

The degree of greyness of grey number reflects the degree of uncertainty. If g^0 is closer to 0, the uncertainty of degree of greyness of grey number is smaller; if g^0 is closer to 1, the uncertainty of degree of greyness of grey number is bigger. Obviously, if $\mu(\Omega) = 1$, then $g^0(\otimes) = \mu(\otimes)$.

Definition 4[6] Suppose the kernel of the gray number $\otimes$ is $\widehat{\otimes}$, the degree of greyness of grey number $\otimes$ is g^0, thus $\widehat{\otimes}_{(g^0)}$ is the simplified form for interval grey number.

The kernel $\widehat{\otimes}$ and the degree of greyness g^0 of grey number are similar to mean value and mean square deviation in statistics respectively. As we know, efficient estimator is to estimate sample dispersion of sampling distribution. Sample dispersion is small, the result is best. Based on this idea, if the degree of greyness of grey number is very small, the interval grey number is then big under the kernels of the interval grey numbers are equals.

Based on the above analysis, we propose a fast and efficient ranking method for interval grey numbers based on the kernel and the degree of greyness of interval grey numbers.

Definition 5 Suppose the interval grey numbers $\otimes_1 \in [a,b]$ $\otimes_2 \in [c,d]$, $\widehat{\otimes}_1$ and $\widehat{\otimes}_2$ are the kernel of $\otimes_1$ and $\otimes_2$, g_1^0 and g_2^0 are the degree of greyness of $\otimes_1$ and $\otimes_2$, so If $\widehat{\otimes}_1 < \widehat{\otimes}_2$, thus $\widehat{\otimes}_1 < \widehat{\otimes}_2$; If $\widehat{\otimes}_1 = \widehat{\otimes}_2$, thus (1) if $g_1^0 = g_2^0$, thus $\otimes_1 = \otimes_2$; (2) if $g_1^0 < g_2^0$, thus $\otimes_1 > \otimes_2$; (3) if $g_1^0 > g_2^0$, thus $\otimes_1 < \otimes_2$.

Example 1 The interval grey numbers $\otimes_1 \in [8, 18], \otimes_2 \in [-2, 18], \otimes_3 \in [2, 14]$ on the field $\Omega \in [-2, 20]$. If we take the interval length as the measure of grey numbers, please rank them.

Solution: According to the known conditions, we can calculate the measure of Ω, $\otimes_1, \otimes_2, \otimes_3$, $\mu(\Omega) = 20 - (-2) = 22$, $\mu(\otimes_1) = 10$, $\mu(\otimes_2) = 20$, $\mu(\otimes_3) = 12$; The kernels and the degree of greyness are $\widehat{\otimes}_1 = 13, \widehat{\otimes}_2 = 8, \widehat{\otimes}_3 = 8, g_1^0 = 0.45$, $g_2^0 = 0.95, g_3^0 = 0.54$. According to definition 5, thus $\otimes_1 > \otimes_3 > \otimes_2$.

12.3 Aggregation Operators of Interval Grey Numbers

For the sake of convenience, we give the interval grey number algorithm firstly.

Definition 6 Suppose $\otimes_1 \in [a, b], a < b, \otimes_2 \in [c, d], c < d$, and k is a arithmetic number, thus $\otimes_1 + \otimes_2, k\otimes$ are interval grey numbers also, and

$$\otimes_1 + \otimes_2 \in [a + c, b + d];\ k\otimes \in [ka, kb]$$

Axiom 1 [6] (the degree of greyness reduction axiom) The degree of greyness of sum, difference, product, quotient for two interval grey numbers that have different degree of greyness is not less than degree of greyness of the larger interval grey number.

We usually take degree of greyness of large interval grey number as degree of greyness of the result.

Based on the above algorithm, we give interval grey numbers WAA operator, interval grey numbers OWA operator, and interval grey numbers HWA operator.

Definition 7 Suppose $\otimes_i \in [a_i b_i](i = 1, 2, \cdots, n)$ is a set of interval grey numbers, and $GWAA : R(\otimes)^n \to R(\otimes)$

$$GWAA_\lambda(\otimes_1, \otimes_2, \cdots \otimes_n) = \lambda_1 \otimes_1 + \lambda_2 \otimes_2 + \cdots + \lambda_n \otimes_n \tag{12.1}$$

where $R(\otimes)$ is the set of all interval grey numbers, $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_n)^T$ is the weight vector of interval grey numbers $\otimes_i (i = 1, 2, \cdots, n), \lambda_i \in [0, 1], \sum_{i=1}^{n} \lambda_i = 1$, the function GWAA is called interval grey numbers WAA operator, marked GWAA.

The characteristic of GWAA operator: to weight each interval grey number for a set of interval grey number (i.e. there are weighted on proper weights according to the importance of each interval grey number), and then to aggregate the weighted interval grey number. Specially, if $\lambda = (\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n})^T$, GWAA operator degenerate interval grey numbers arithmetic average operator GWA:

$$GWA(\otimes_1, \otimes_2, \cdots \otimes_n) = \frac{1}{n} \sum_{i=1}^{n} \otimes_i \tag{12.2}$$

Theorem 1 Suppose $\otimes_i \in [a_i b_i](i = 1, 2, \cdots, n)$ is a set of interval grey numbers, and $GWAA_\lambda(\otimes_1, \otimes_2, \cdots \otimes_n)$ is also an interval grey number, and

$$GWAA_\lambda(\otimes_1, \otimes_2, \cdots \otimes_n) \in [\sum_{i=1}^{n} \lambda_i a_i, \sum_{i=1}^{n} \lambda_i b_i] \tag{12.3}$$

$$g^0[GWAA_\lambda(\otimes_1, \otimes_2, \cdots \otimes_n)] = \underset{i=1}{\overset{n}{\vee}} g^0(\otimes_i) \tag{12.4}$$

Definition 8 Suppose $\otimes_i \in [a_i b_i](i = 1, 2, \cdots, n)$ and $GOWA : R(\otimes)^n \rightarrow R(\otimes)$

$$GOWA_\omega(\otimes_1, \otimes_2, \cdots \otimes_n) = \omega_1 \otimes_{\sigma(1)} + \omega_2 \otimes_{\sigma(2)} + \cdots + \omega_n \otimes_{\sigma(n)} \tag{12.5}$$

where $(\sigma(1), \sigma(2), \cdots, \sigma(n))$ is permutation of $(1, 2, \cdots, n)$, $\otimes_{\sigma(i-1)} \geq \otimes_{\sigma(i)},$, $i = 2, 3, \cdots n$, $\omega = (\omega_1, \omega_2, \cdots, \omega_n)^T$ is the weight vector associated on function GOWA, $\omega_i \in [0, 1]$, $\sum_{i=1}^{n} \omega_i = 1$. The function GOWA is called interval grey numbers OWA operator, marked GOWA. Specially, if $\omega = (\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n})^T$, GOWA operator degenerates interval grey number arithmetic average operator GWA.

Theorem 2 Suppose $\otimes_i \in [a_i b_i](i = 1, 2, \cdots, n)$ is a set of interval grey numbers, and $GOWA_\omega(\otimes_1, \otimes_2, \cdots \otimes_n)$ is an interval grey number also, and

$$GOWA_\omega(\otimes_1, \otimes_2, \cdots \otimes_n) \in \left[\sum_{i=1}^{n} \omega_i a_{\sigma(i)}, \sum_{i=1}^{n} \omega_i b_{\sigma(i)}\right] \tag{12.6}$$

$$g^0[GOWA_\omega(\otimes_1, \otimes_2, \cdots \otimes_n)] = \underset{i=1}{\overset{n}{\vee}} g^0(\otimes_i) \tag{12.7}$$

where $\otimes_{\sigma(i)} \in [a_{\sigma(i)}, b_{\sigma(i)}](i = 1, 2, \cdots n)$ is ranking i in $(\otimes_1, \otimes_2, \cdots, \otimes_n)$.

Definition 9 Suppose $\otimes_i \in [a_i b_i](i = 1, 2, \cdots, n)$ is a set of interval grey numbers, and $GHWA : R(\otimes)^n \rightarrow R(\otimes)$

$$GHWA_{\omega,\lambda}(\otimes_1, \otimes_2, \cdots \otimes_n) = \omega_1 \otimes'_1 + \omega_2 \otimes'_2 + \cdots + \omega_n \otimes'_n \tag{12.8}$$

where $\omega = (\omega_1, \omega_2, \cdots, \omega_n)^T$ is the weight vector associated on function GHWA, $\omega_i \in [0, 1]$, $\sum_{i=1}^{n} \omega_i = 1, \otimes'_i (i = 1, 2, \cdots, n)$ is ranking i in a set of weighted interval grey numbers $(n\lambda_1 \otimes_1, n\lambda_2 \otimes_2, \cdots, n\lambda_n \otimes_n)$, $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_n)^T$ is the weight vector for $\otimes_i (i = 1, 2, \cdots, n)$, $\lambda_i \in [0, 1]$, $\sum_{i=1}^{n} \lambda_i = 1$, n is a balanced factor. The function GHWA is called interval grey numbers HWA operator, marked GHWA.

The characteristic of GHWA operator: not only we can consider the importance of each interval grey number, but we can also embody the importance of the interval grey number location.

Theorem 3 Suppose $\otimes_i \in [a_i b_i](i = 1, 2, \cdots, n)$ is a set of interval grey numbers, and $GHWA_{\omega,\lambda}(\otimes_1, \otimes_2, \cdots \otimes_n)$ is also an interval grey number, and

$$GHWA_{\omega,\lambda}(\otimes_1, \otimes_2, \cdots \otimes_n) \in [\sum_{i=1}^{n} \omega_i a'_i, \sum_{i=1}^{n} \omega_i b'_i] \tag{12.9}$$

$$g^0[GHWA_{\omega,\lambda}(\otimes_1, \otimes_2, \cdots \otimes_n)] = \underset{i=1}{\overset{n}{\vee}} g^0(\otimes_i) \tag{12.10}$$

where $\otimes'_i \in [a'_i, b'_i](i = 1, 2, \cdots, n)$ is ranking i in $(n\lambda_1\otimes_1, n\lambda_2\otimes_2, \cdots, n\lambda_n\otimes_n)$.

12.4 Grey Multi-attribute Multi-person Decision-making Problem

Because any gray number $\otimes$ can be expressed $\otimes \in [a, b]$, the decision information can be expressed the interval grey number [17]. According to GWAA operator, GOWA operator and GHWA operator, we develop an approach to solve grey multi-attribute multi-person decision-making problems, in which the attributive weights are completely known and the attributor values are interval grey numbers.

We denote $D = \{d_1, d_2, \cdots, d_t\}$ as the set of decision maker, $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_t)^T$ is weight vector of decision maker, $\lambda_k \in [0, 1], \sum_{k=1}^{t} \lambda_k = 1$. The project $A_i \in A$ attribute value for $S_j \in S$ are interval grey numbers $\otimes_{ij}^{(k)} \in [a_{ij}^{(k)}, b_{ij}^{(k)}](a_{ij}^{(k)} \le b_{ij}^{(k)}, (i = 1, 2, \cdots, n; j = 1, 2, \cdots, m; k = 1, 2, \cdots, t)$, which are given by decision maker $d_k \in D$. We get decision matrix $R_k = (\otimes_{ij}^{(k)})_{n\times m}$ $(k = 1, 2, \cdots, t)$.

Step 1 Normalize the decision matrix $R_k = (\otimes_{ij}^{(k)})_{n\times m}$ $(k = 1, 2, \cdots, t)$. In order to eliminate the differences of attribute values of alternatives in the dimension and increase comparability of different projects, we normalized $\otimes_{ij}^{(k)} \in [a_{ij}^{(k)}, b_{ij}^{(k)}]$ using the gray range transform [17]' as following.

On the cost index value

$$x_{ij} = \frac{b_j^* - b_{ij}}{b_j^* - a_j^*}, \; y_{ij} = \frac{b_j^* - a_{ij}}{b_j^* - a_j^*} \tag{12.11}$$

On the benefit index value

$$x_{ij} = \frac{a_{ij} - a_j^*}{b_j^* - a_j^*}, \; y_{ij} = \frac{b_{ij} - a_j^*}{b_j^* - a_j^*} \tag{12.12}$$

where $a_j^* = \min\limits_{1 \le i \le n}\{a_{ij}\}$, $b_j^* = \max\limits_{1 \le i \le n}\{b_{ij}\}\,(j = 1, 2, \cdots, m)$. The normalized decision matrix is $R_k' = (\otimes_{ij}^{\prime(k)})_{n\times m}$. where $\otimes_{ij}^{\prime(k)} \in [x_{ij}, y_{ij}](i = 1, 2, \cdots, n;$ $j = 1, 2, \cdots, m)$.

Step 2 Aggregate attribute value $\otimes_{ij}^{\prime(k)} a(k = 1, 2, \cdots, t)$ for project A_i using GHWA operator. We get aggregated attribute value $\otimes_{ij}'$,

$$\otimes_{ij}' = GHWA_{\omega,\lambda}(\otimes_{ij}^{\prime(1)}, \otimes_{ij}^{\prime(2)}, \cdots \otimes_{ij}^{\prime(t)}) = \sum_{k=1}^{t} \omega_k \otimes_{ij}^{\prime(\sigma(k))} \tag{12.13}$$

where $\omega = (\omega_1, \omega_2, \cdots, \omega_t)^T$ is the weight vector associated with function GHWA, $\omega_k \in [0, 1]$, $\sum\limits_{k=1}^{t} \omega_k = 1$, $\otimes_{ij}^{\prime(\sigma(k))}$ rank i in the set of weight vector $(t\lambda_1 \otimes_{ij}^{\prime(1)}, t\lambda_2 \otimes_{ij}^{\prime(2)}, \cdots, t\lambda_t \otimes_{ij}^{\prime(t)})$, n is a balanced factor. We get multi-person decision matrix $R' = (\otimes_{ij}')_{n\times m}$.

Step 3 Aggregate attribute value for the rank of multi-person decision matrix $R' = (\otimes_{ij}')_{n\times m}$ using GWAA operator. We get aggregated attribute value $\otimes_i'$,

$$\otimes_i' = GWAA(\otimes_{i1}', \otimes_{i2}', \cdots, \otimes_{in}') = \sum_{j=1}^{m} w_j \otimes_{ij}' \tag{12.14}$$

where $w = (w_1, w_2, \cdots, w_m)^T$ is attribute weight vector, $w_i \in [0, 1]$, $\sum\limits_{i=1}^{m} w_i = 1$.

Step 4 Compute the kernels $\widehat{\otimes}_i'$ and the degree of greyness $g^0(\otimes_i')$ of aggregated attribute value $\otimes_i'$.

Step 5 Use definitions 5 ranking $\otimes_i'$ $(i = 1, 2, \cdots, n)$, so we can get the result.

12.5 Example Analyses

Example 2 [15] If a company's board of directors consist of d_1, d_2, d_3, weight vector for d_1, d_2, d_3 is $\lambda = (0.5, 0.3, 0.2)^T$. They want to choose a project form A_1, A_2, A_3, A_4, so they score A_1, A_2, A_3, A_4 (range from 0 to 100) on investment cycle (S_1), investment risk and loss value (S_2), fixed assets investment (S_3), and management cost (S_4). In order to determine the weight of attributes, the board of directors issued a survey. Judgment matrixes are given by 12 experts respectively. They obtained attribute weight vector $w = (0.2, 0.3, 0.3, 0.2)^T$ using the Delphi method and analytic hierarchy process [7]. We get three decision matrixes that are given in the following table, trying to determine the best investment project. Tables 1, 2, 3.

Step 1 Normalize the decision matrix $R_k = (\otimes_{ij}^{(k)})_{n\times m}$ $(k = 1, 2, 3)$ using formula (12.11). We get normalized decision matrix R_1', R_2', R_3'. We can get interval

Table 1 The decision matrix given by decision maker d_1

	S_1	S_2	S_3	S_4
A_1	[82, 87]	[85, 95]	[92, 96]	[60, 70]
A_2	[90, 100]	[75, 85]	[60, 70]	[70, 75]
A_3	[60, 70]	[75, 80]	[95, 100]	[65, 70]
A_4	[75, 82]	[67, 73]	[50, 60]	[85, 89]

Table 2 The decision matrix given by decision maker d_2

	S_1	S_2	S_3	S_4
A_1	[60, 65]	[75, 80]	[90, 95]	[65, 68]
A_2	[84, 86]	[60, 65]	[70, 75]	[82, 85]
A_3	[60, 64]	[70, 77]	[66, 69]	[90, 95]
A_4	[65, 73]	[80, 86]	[89, 93]	[64, 70]

Table 3 The decision matrix given by decision maker d_3

	S_1	S_2	S_3	S_4
A_1	[68, 72]	[78, 83]	[80, 92]	[66, 73]
A_2	[82, 86]	[68, 70]	[72, 75]	[78, 84]
A_3	[88, 93]	[84, 90]	[73, 77]	[92, 96]
A_4	[85, 88]	[70, 75]	[72, 76]	[86, 90]

grey number simplified form by the kernels and the degree of greyness of the interval grey number. So we get simplified representation matrix form normalized decision matrix.

Step 2 Aggregate attribute values $\otimes_{ij}^{\prime(k)}$ $(k = 1, 2, 3)$ for project A_i using GHWA operator. We get aggregated attribute value $\otimes_{ij}^{\prime}(i = 1, 2, 3, 4; j = 1, 2, 3, 4)$. $\omega = (0.4, 0.3, 0.3)^T$ is a set of weight vector associated with function GHWA. We get multi-person decision matrix

$$R' = \begin{pmatrix} [0.59, 0.74] & [0.14, 0.41] & [0.05, 0.25] & [0.77, 1.04] \\ [0.05, 0.24] & [0.62, 0.86] & [0.7, 0.89] & [0.45, 0.61] \\ [0.68, 0.77] & [0.41, 0.64] & [0.46, 0.58] & [0.39, 0.56] \\ [0.44, 0.65] & [0.59, 0.83] & [0.64, 0.84] & [0.33, 0.48] \end{pmatrix}$$

Step 3 Aggregate attribute values for the rank of multi-person decision matrix $R' = (\otimes_{ij}')_{n\times m}$ using GWAA operator. We get aggregated attribute value $\otimes_i'(i = 1, 2, 3, 4)$,

$$\otimes_1' = [0.33, 0.56]\ \otimes_2' = [0.49, 0.70]\ \otimes_3' = [0.48, 0.63]\ \otimes_4' = [0.52, 0.73]$$

Step 4 Compute the kernels $\widehat{\otimes}_i'(i = 1, 2, 3, 4)$ of aggregated attribute value $\otimes_i'$:

$$\widehat{\otimes}_1' = 0.445\ \widehat{\otimes}_2' = 0.595\ \widehat{\otimes}_3' = 0.555\ \widehat{\otimes}_4' = 0.625$$

Compute the degree of greyness $g^0(\otimes_i')(i = 1, 2, 3, 4)$ of aggregated attribute value $\otimes_i'$ using formula (12.10) and formula (12.4).

Step 5 Rank $\otimes_i'(i = 1, 2, \cdots, n)$ using the kernels $\widehat{\otimes}_i'$ and the degree of greyness $g^0(\otimes_i')(i = 1, 2, 3, 4)$, so we can get the result, $A_4 \succ A_2 \succ A_3 \succ A_1$. Therefore, A_4 is the best investment project.

12.6 Conclusion

The operation and ranking method for interval grey numbers is the starting point of the research of the grey system theory, which plays an important role in the development of the grey system theory. Based on the kernel and the degree of greyness of interval grey number, authors propose a fast and efficient ranking method, which can perfectly solve the ranking problem for interval grey numbers.

In this paper, authors propose GOWA operator, GWAA operator, and GHWA operator. Those operators extend the research field of information aggregation, and provide an efficiency approach for grey multi-attribute decision-making problems.

References

1. Julong, Deng: Control problems of grey systems. Syst. Control Lett. **1**(5), 288–294 (1982)
2. Sifeng, Liu, Yi, Lin: Grey information: theory and practical application. Springer, London (2005)
3. Sifeng, Liu, Yaoguo, Dang: Grey System Theory and Its Application. Science Press, Beijing (2010)
4. Yi, Y.: Extended grey numbers and their operations. In Proceedings of the 2007 IEEE International Conference on Systems, Man and Cybernetics, pp. 2181–2186, (2007)
5. Yi, Y., Liu, S.: Kernels of grey numbers and their operations. Fuzzy Systems.. FUZZ-IEEE 2008. (IEEE World Congress on Computational Intelligence), pp. 826–831, (2008)
6. Sifeng, Liu, Zhigeng, Fang, Naiming, Xie: Algorithm rules of interval grey numbers based on the "Kernel" and the degree of greyness of grey numbers. Syst. Eng. Elect. **32**(2), 313–316 (2010)
7. Naiming, Xie, Sifeng, Liu: On comparing grey numbers with their probability distributions. Syst. Eng. Theory Pract. **29**(4), 169–175 (2009)
8. Zhigeng, Fang, Sifeng, Liu, et al.: Study on improvement of token and arithmetic of interval grey numbers and its GM (1, 1) model. Eng Sci **7**(2), 57–61 (2005)
9. Haiying, Guo, Yanjun, Pang: A kind of order relations of interval grey number via kenning degree. Hebei Inst. Archit. Sci. Technol. **20**(1), 85–86 (2003)
10. Yager, R.R., Kacprzyk, J.: The Ordered Weighted Averaging Operators: Theory and Applications. Kluwer, Norwell (1997)
11. Zeshui, Xu: Uncertain Multiple Attribute Decision Making: Methods and Application. Tsinghua University Press, Beijing (2004)
12. Torra, V.: Information Fusion in Data Mining. Springer, New York (2003)

13. Harsanyi, J.C., Cardinal, W.: Individualistic ethics and interpersonal comparisons of utility. J. Political Econ. **63**, 309–321 (1955)
14. Zeshui, Xu, Qingli, Da: An overview of operators for aggregating information. Int. J. Intell. Syst. **18**, 953–969 (2003)
15. Huayou, Chen, Jinpei, Liu, Hui, Wang: A class of continuous ordered weighted harmonic (C-OWH) averaging operators for interval argument and its applications. Syst. Eng. Theory Pract. **28**(7), 86–92 (2008)
16. Jiefang, Wang, Sifeng, Liu: Method of ranking three parameters interval grey numbers and its application in interval DEA model. Syst. Eng. Elect. **33**(1), 106–109 (2011)
17. Dang, Luo: Analysis Method for Grey Decision-making Problems. Yellow River Water Conservation press, Zhengzhou (2005)

Chapter 13
The Information Discovery Service of Electronic Product Code Network

Siwei Zheng and Ge Yu

Abstract Aiming at the problem of low query efficiency and delay of Object Naming Service (ONS), an improved Chord algorithm is used in this paper. By introducing an improved Chord algorithm for information discovery service of EPC network, this paper classifies neighboring nodes into a group, stratifies all the nodes according to the different performance, and selects more powerful nodes as the super nodes to manage ordinary nodes in the same region. The results of simulation experiments show that the query hops and network delay are both reduced while query efficiency is improved to a certain extent. Improved Chord algorithm can effectively solve the problem of ONS query bottlenecks, thus the query efficiency in information discovery service of EPC network is improved.

Keywords EPC network · Object naming service · Chord algorithm · Super node · Discovery service

13.1 Introduction

In 1999, in support of the Uniform Code Council (UCC), the concept of the Electronic Product Code (EPC) was proposed in the automatic identification Laboratory (Auto-ID Labs) [1] established at the Massachusetts Institute of Technology (MIT), whose purpose is to form an open network where worldwide items can achieve real-time information sharing, that is EPC network [2].

EPC network consists of five parts: Electronic Product Code (EPC), Information Identification System (ID System), EPC Middleware, Information Discovery Service, EPC Information Services (EPCIS). In EPC network, the item information

S. Zheng (✉) · G. Yu
Hangzhou Institute of Service Engineering, Hangzhou Normal University, Hangzhou, China
e-mail: sijia67@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_13,

is randomly stored in any information server in scattered, so the key issues which EPC Network Information Discovery Service needs to solve is to how to retrieve the relevant information of specific items in the unpredictable information servers efficiently and accurately. At this time it will use Object Naming Service (ONS), which can provide a query service to EPC Information Services according to EPC code, and ultimately realize the users' access to the relevant information. But the existing ONS has some defects: ONS is a resource query service depending on the root ONS server, so when more and more query requests come, the root ONS server will be overloaded. Once the root ONS server collapses, the entire query system will face paralysis, which will severely restrict the query efficiency. But EPC network that will be built in the future will host information sharing of global items, the sheer numbers of query requests that it will face can be imagined. It can be seen, in EPC Network, an efficient information discovery service is particularly important. This paper will propose an information discovery service based on DHT network, and try to solve the disadvantage of low query efficiency and delay in the existing ONS and improve query efficiency of EPC network [3].

13.2 P2P Network Search Technology

13.2.1 DHT

Distributed Hash Table (DHT) [4] is a data structure that can effectively achieve storage, management, and query of data in the distributed system or P2P network. Whose principle is that all the nodes in the network is constructed into a hash table, and splited into one by one fragments, then these fragments are stored in physical connected nodes according to certain rules (These rules include Chord, Pastry, CAN, etc.), and each node will maintain a fragment of the entire hash table, which is called a node routing table. In this way, the resource query of entire network is realized by one by one node routing table. This query idea using hash table has become increasingly widespread in network query.

13.2.2 Chord Algorithm

Chord is a resource search algorithm proposed by the Massachusetts Institute of Technology based on DHT. Given a keyword, then the keyword is mapped to a node, usually it is denoted as (k, v), k refers to the keyword identifier, which is obtained through the hash operation of resource keyword; v is called node identifier, which is the hashed value of IP address of the node. In the compatible hash, the keywords are all stored in the position called a successor node—successor (k), which represents the first node of node identifier that is equal to k or immediately

after k. All the identifiers will be arranged in a circle, and the value follows from 0 to 2^m-1. In order to make two nodes will not hash to the same identifier, usually m must be long enough [5, 6].

Each node has its own routing table (Finger Table) with m entries at most. The i-th item of the node n in the routing table is successor node s with a distance of 2^{i-1} at least to node n in the circle, that is, $s = \text{successor}(n + 2^{i-1}) = (n + 2^{i-1}) \bmod 2^m$, $1 \leq i \leq m$ (count modulo 2^m). Through maintaining the routing table, each node can query any other nodes via information of some nodes in the routing table. And more close from the node in the Chord circle, more information the node will know.

13.3 Application of Chord Algorithm in Information Discovery Service of EPC Network

Chord algorithm has the advantage of achieving quick query resource location, but there are still some questions: (1) It does not use all the nodes discriminatively according to the differences of node performance, and makes the high-performance nodes do not be fully utilized, while the low-performance nodes are used overload, which seriously affects the efficiency of the query and the stability of the network. (2) It does not consider the actual physical topology between nodes. Chord algorithm makes each node mapped to a logical space, after the nodes are hashed, the actual physical topology will lose. It causes that the nodes which are close in the physical network may be very far apart after mapped function, and ultimately it will lead to a repeat path and take up network bandwidth.

Based on Chord algorithm, information discovery service of EPC Network takes into account the different performance and actual physical topology of the nodes. According to the difference of node performance, all the nodes are divided to super node layer (Super Node, SN) and ordinary node layer (Ordinary Node, ON). At the same time, the node's IP address is grouped by CIDR (Classless Inter-Domain Routing), and the nodes that are close in the physical address are divided into a group.

13.3.1 The Stratification of Nodes

Through using the idea of Chord algorithm and regarding information servers storing related information of items as independent distributed data nodes. According to the difference of node performance, all the nodes are divided to super node layer (Super Node, SN) and ordinary node layer (Ordinary Node, ON). At the same time because there are a large number of hierarchies in the real world, the division can be match with the actual business model. Super nodes bear the task of

query and transmission, and have three tables: index record table, routing forwarding table and query cache table. Index record table stores a section of keyword index managed by super nodes and the position of the object of the key value in the area centered on super nodes. The requester can quickly locate in the related ordinary node through index record table and obtain the related address information service of data node. Routing forwarding table stores routing information between the super nodes. Query cache table stores node information of recent multiple queries [7].

13.3.2 The Grouping and Join in of Nodes

Through CIDR (Classless Inter-Domain Routing), the node's IP address is grouped. With respect of traditional network protocol address, CIDR can improve the utilization of the IP address better, and can shorten the routing table. CIDR allocates network address based on variable length block, and uses the various length prefix to instead of the network number and subnet number of classification address. In the improved model, the nodes which have the same upper 16 bits of the IP address are divided into a group, and the nodes that are close in the physical address are divided into a group.

When a new node wants to add to the hierarchical network, first, make the comparison on the upper 16 bits of IP address of nodes, if the upper 16 bits already exists, the new node will join the group, and register its own resource information to the super node, and otherwise the new node will create a new group.

13.3.3 Resources Release

(1) Each node in the network uses its own IP address to make hash operation and obtains a node identifier ID.
(2) First, the items to be networked use theirs own EPC codings to make hash operation, and the results are recorded as Key. This Key value maps detailed items information through Chord routing algorithm to the ordinary node that is nearest the node identifier ID.
(3) At the same time, the ordinary node will release resource information to the super node that is responsible for the management of the area. Through Chord routing algorithm, super node build the index record of Key value in the corresponding super node, which is the mapping relationship between the Key value and node Identifier ID.

13.3.4 Resources Query

Once the query request is issued, nodes in the network will raise a query based on Chord routing algorithm:

(1) Once nodes receive the user's query of Key = Hash (EPC encoding), firstly they will research the node that is closest to the Key value in the ordinary nodes in their own area. Then they send the received query request to that request node.
(2) If the resource needed is not found in the ordinary node layer, the node will continue to send query request to the super node in its own area. When this super node receives the query request, it will continue to forward query request to other super nodes.
(3) When a message is sent to a super node, it will initiate a query in the area where the super node is. The super node will first check whether there is an index matched with the Key value in its own query cache table and index record table, if there is, then it will return the resource address of the source query node. Otherwise, the super node will send query request to another super node which is closest to the Key value.
(4) After the super node receives the resource query request, if the matched information with the Key value can be found in the index record table, the query request node will receive the query request issued by the super node. Otherwise, it will return a query failed information.
(5) According to returned query information, the user will directly query related resources of items from nodes storing item information, and obtain detailed item information.
(6) After be successfully queried, the super node in the area where the query is issued will record the Key value of this time and the node information. The benefit of the query is that when other nodes query the same Key value again, the system does not need to find information in the network of other areas, and can directly obtain the Key value and corresponding node's IP address through query cache table. It reduces query hops, and improves the query efficiency. Figure 13.1 shows the model of information discovery service based on Chord algorithm.

13.4 Experiments and Analysis

The paper generates the simulation by NS2, and makes a performance compare within ONS query algorithm, Chord routing algorithm and improved Chord routing algorithm used in EPC network from two aspects of average query hops and average query delay respectively.

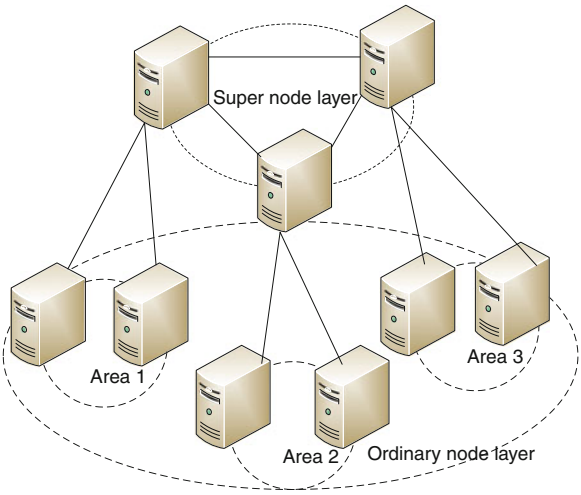


Fig. 13.1 Information discovery service based on Chord algorithmChord algorithm

The experiment adopts 5 different node numbers (2000, 4000, 6000, 8000, 10000), and each node stores 20 documents. Experiment will be carried out 10 times for each set of values.

As can be seen from Fig. 13.2, with the growth of the number of nodes, the average query hops of Chord algorithm and improved Chord algorithm are also growing, but they do not dramatically increase, and remain in a relatively stable range. The average query hops of Chord algorithm increase in the proportion of $O(\ln N)$ with the increase of the number of nodes N. Improved Chord algorithm curve is more flat than Chord algorithm curve, and its average query hops is few. This is because that the node division according to the physical position makes nodes in the same network segment are divided in the same group, and the hierarchical Chord model makes that each super node is responsible for ordinary nodes

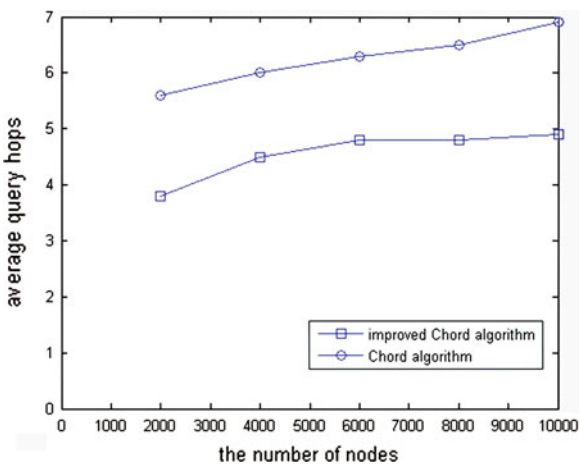


Fig. 13.2 Comparison of average hops on two models

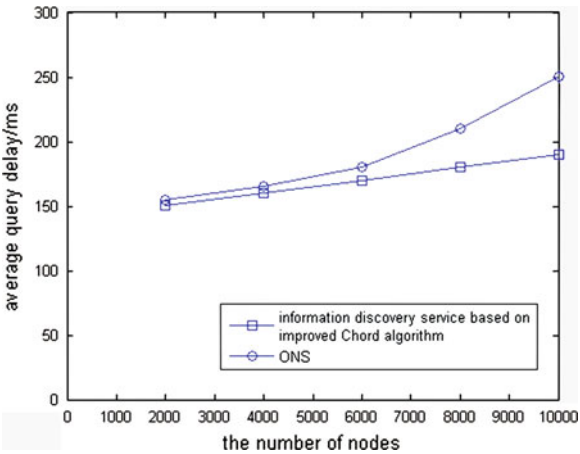


Fig. 13.3 Comparison of average delay on two models

of the area, so access speed between nodes in an area becomes fast, and it reduces query hops, finally query efficiency has been significantly improved to a certain extent.

As can be seen from Fig. 13.3, the average query delay of information discovery service based on improved Chord algorithm is around 150 ms. With the increase of the number of nodes, the value is in a flat range. This is because when the nodes issue the query request, they are no longer dependent on the central node, but exchange data among the nodes, and grouped node routing reduces forwarding times of the entire network, so the number of nodes have few influence on the query time. The average query delay of ONS gradually increases as the number of nodes increase, this is because the query mechanism of ONS depends on the root servers, once the query request is too large, the request will wait in line, at the same time, ONS will experience two address resolution processes of ONS returning the URI address and DNS returning the IP address, which results in a delay of the query.

13.5 Conclusion

Aiming at the problem of overload in ONS root server, this paper presents an improved Chord algorithm of information discovery service. Improved information discovery service has the following advantages: (1) facing the information resolution of a large number of items, it has high performance in load balancing, avoiding the bottleneck problem of root ONS. And average query delay has reduced by 25 % than ONS. It demonstrates that it can meet the massive query needs of EPC network. (2) The nodes are grouped according to the physical topology, discovery algorithm only forwards information among the super nodes, and super nodes and neighboring ordinary nodes constitute an autonomous area,

which all improve query efficiency to some extent. (3) Query cache table in the super node is also conducive to quickly locate to the previous query node information. The items information in the EPC network are related to the commercial interests and privacy, so security protection mechanisms play an important role in information discovery service of EPC network. The next step is to study the security problem of information discovery service.

References

1. AutoID Labs homepage [EB/OL]. http://www.autoidlabs.org/ (2010)
2. Xiaoli, D, Baoping, Y: ONS service in EPC network. Microelectron. Comput. **22**(2), 17–21 (2005)
3. EPCglobal Inc.V.1.0.1. Object Naming Service (ONS) Standard [S] (2008)
4. Stoica, I., et al.: Chord: a scalable peer-to-peer lookup protocol for Internet applications. IEEE/ACM Trans Netw. **11**(1), 17–32 (2003)
5. Liu H, Cheng L: Research on naming service architecture of internet of things based on DHT. Appl. Res. Comput. **6**(28), 2327–2329 (2011)
6. Yong, H: The summarizing of Chord and improvement algorithms in P2P systems. Comput Knowl. Technol. **8**(1), 38–71 (2006)
7. Liu, D: Study on the information discovery service of Internet of Things based on P2P [D]. Zhengzhou Univercity, Zhengzhou (2011)
8. Minghui, J.: Hierarchical bidirectional Chord. International conference on educational and information technology (ICEIT2010), pp. 486–489 (2010)

Chapter 14
Modeling Goals for Information System by a Heuristic Way

Bin Chen, Qiang Dong and Zhixue Wang

Abstract In order to construct goals model completely, a heuristic way is proposed in this paper based on multiple views. A meta-model is proposed for guidance of modeling of organizational view, business view and goal view. In the meta-model, the relationship among actor, task and goal in meta-model shapes an iron triangle and establishes basis of the heuristic way for goal modeling. In this way, goals can not only be constructed directly by decomposition of goals, but also be elicited indirectly by analysis of the purpose of every task appearing in business view and by asking actors appearing in organizational view why they do the work they undertake. For illustration, a case about a hospital management information system is offered to illustrate idea of this method. This heuristic way is very practical for goal modeling, which can be seen as a helpful complement for directly goal modeling.

Keywords Information system · Goal modeling · Heuristic way

14.1 Introduction

In goal-oriented requirements engineering practice, goals provide rationales for requirements of information system. Analysts often elicit functional and non-functional requirements from analysis of stakeholders' goals [1]. But there are still many vacancies about a systematical and effective way on how to acquire goal model. For goals of information system, it is often directly constructed as a hierarchy from high level to low level [2, 3]. But in practice, it is very idealistic because few stakeholders can describe all goals undertaken by the stuff in an

B. Chen (✉) · Q. Dong · Z. Wang
College of Command Information System, PLA University of Science and Technology, Nanjing, China
e-mail: chenbinmsn@msn.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_14,

enterprise. Furthermore, stakeholders don't prefer to directly describe abstract goals but to describe concrete operational scenario. To solve this problem, a heuristic way is presented in this paper for goal modeling based on multiple views guided by a meta-model. A case about a hospital management information system is offered to illustrate idea of this method. UML notations are directly adopted to illustrate the construction of meta-model and case study.

14.2 An Enterprise Meta-Model for Guiding Goal Modeling

An enterprise represents circumstance in which information system will run. For information system requirements modeling, multiple views can be used for describe the characteristics of the enterprise, which commonly include goal view, organizational view and business view. For guiding multiple views modeling, an enterprise meta-model is offered firstly. A view can be seen as a projection of meta-model. Concepts and relations in a view are instances of meta-concepts and meta-relations in meta-model. Only minimal and necessary concepts, relations and constrains for requirements modeling are included in the enterprise meta-model. Contrast to complicated jargons in a specific domain, the enterprise meta-model affords a good communication bridge for stakeholders and analysts to understand each other. Concepts space of the enterprise meta-model is illustrated in Fig. 14.1 adopting UML notations. For restriction of paper length, not all details but meta-concepts and meta-relations of the meta-model are presented below.

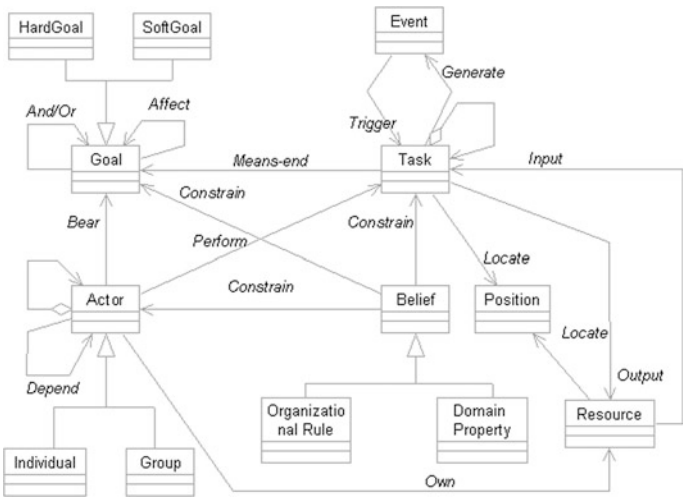


Fig. 14.1 Concepts space of enterprise meta-model

The enterprise meta-model include eight abstract meta-concepts. (1) ⟨Actor⟩. An actor represents an active entity in an organization, which can perform some task to achieve some goal. (2) ⟨Belief⟩. A belief represents a fact which actors should obey in some domain. A belief constrains performance of tasks and achievement of goals. (3) ⟨Task⟩. A task represents a serial of activities adopted for achievement of some goals. (4) ⟨Goal⟩. A goal represents a situation or a state expected to be achieved for satisfying users' needs or solving problems in the enterprise. There are two kinds of goals including soft goal and hard goal [4]. (5) ⟨Resource⟩. A resource represents a physical or information entity needed in performance of some task or achievement of some goal. (6) ⟨Event⟩. An event represents change of some state in the enterprise. An event expresses dynamical dependency between tasks, which happens when some state is reached or some task performs or some situation is satisfied. (7) ⟨Position⟩. A position represents a place where a task is performed or a resource is acquired.

There are thirteen kinds of meta-relations among meta-concepts. (1) Specialization relation describes abstract concept is specialized as concrete concept. (2) "Part-of" relation describes whole is partitioned as parts. (3) "Means-end" relation between task and goal describes some tasks are means to achieve some goal. (4) "Perform" relation between actor and task describes some actor can perform some task. (5) "Constrain" relation between belief and goal or task describes a belief constrains a task, a goal or an actor. (6) "Own" relation between actor and resource describes an actor owns a kind of resource. (7) "Bear" relation between goal and actor describes an actor has a responsibility to achieve a goal. (8) "Trigger" relation between task and event describes an event may trigger performance of a task. (9) "Generate" relation between task and event describes an event may be generated by performance of a task. (10) "Locate" relation between position and task or resource describes performance of a task locates somewhere or a kind of resource locates somewhere. (11) "Dependency" represents dependency relation between actors to accomplish their responsibilities because every actor only has limit capabilities. (12) "And/Or" [5] represents decomposition of goal, which describes a father goal may be decomposed into some sub-goals in an "and" way or an "or" way. (13) "Affect" represents achievement of one goal may affect achievement of the other goal. There are two "affect" relation. "Support" relation includes sufficient support relation and necessary support relation. "Denial" relation includes sufficient denial relation and necessary denial relation [6].

14.3 Mechanism and Process of a Heuristic Way for Goal Modeling

In meta-model presented above, goals can be borne by actors because "bear" relation exits between actor and goal, and goals can be achieved by tasks because "means-end" relation exits between task and goal. So goals can not only be

constructed directly by decomposition of some high level goals, but also be elicited indirectly by analysis of the purpose of every task appearing in business view and by asking actors appearing in organizational view the reason why they do their work. This heuristic way is indirect but very practical for goal modeling, which can be seen as a helpful complement for directly goal modeling. A program of hospital management information system (HIMS) is used to illustrate the modeling process. In order to achieve comprehensive electronic management and improve service capability and efficiency, HIMS utilizes advanced medical facilities. For restriction of paper length, not all details are given in the case.

(1). Organizational View Modeling

Organizational view describes static organizational characteristics of an enterprise, which includes two kinds of models: organizational structure model and actor dependency relationships model.

1. Organizational structure modeling

From social viewpoint, an enterprise can be regarded as an organizational network. According to partition of responsibilities, larger group may be decomposed into smaller groups until individuals appear. Finally, a hierarchy shapes in the organizational decomposition. For example, a hospital is made up of many groups, such as administrators department, common services department, policlinic, residency department, etc. These groups are made up of smaller groups. Figure 14.2 describes the decomposition of policlinic department. Arrow lines with blue diamond represent "part-of" relationships; rectangles represent groups;

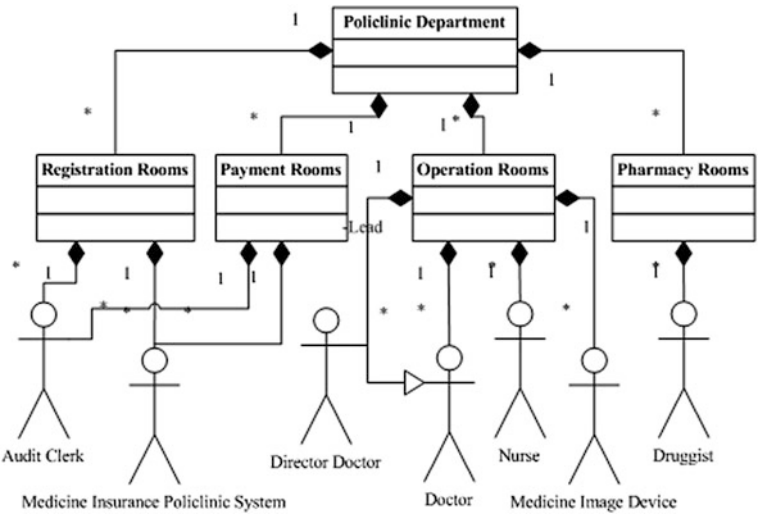


Fig. 14.2 Organizational structure modeling

and little people heads represent individuals. Hereinto, director doctor is also a kind of doctor, but he or she can lead other doctors.

2. Actor dependency relationships modeling

For fulfilling some tasks, actors may cooperate with each other. For getting some resources, actors may need the help of others who own the resources. So dependency relationships shape. Of course, organizational rules actors abide and goals actors bear should be analyzed in the course of modeling. The dependency relationships between members in policlinic group are illustrated in Fig. 14.3. Policlinic administrator can perform tasks such as print registration or payment bill. Patient depends on administrator to perform tasks such as print registration and payment, and depends on doctor or photo device to diagnose diseases. These are all called task dependency. Doctor affords a prescription for patient, which is called resource dependency.

(2). Business View Modeling

Business view describes dynamic behavior characteristics of an enterprise, which also includes two kinds of models: tasks decomposition model and activity flow model.

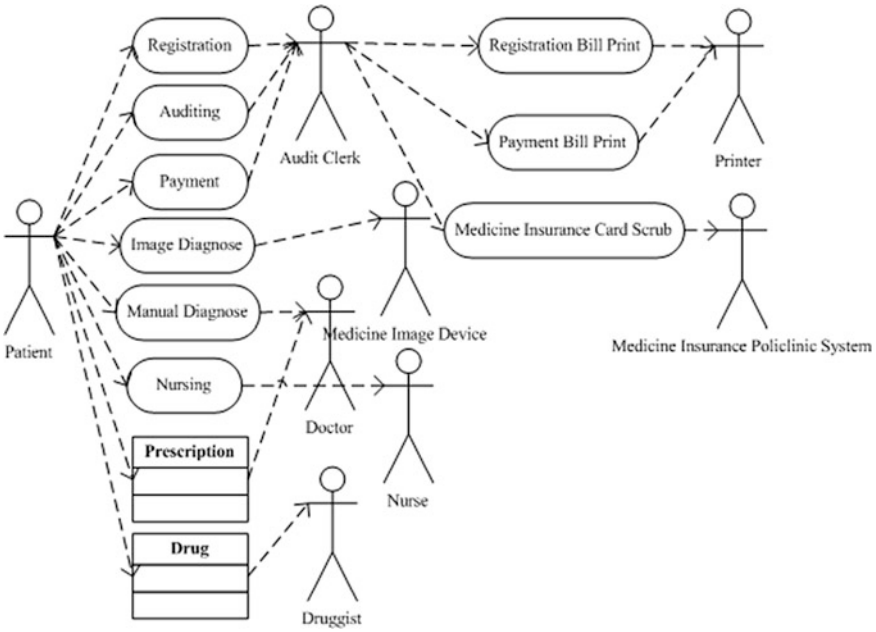


Fig. 14.3 Actor dependency relationships modeling

1. Tasks decomposition modeling

For an enterprise, daily business is comparatively stable. Top tasks are often not less than five. Tasks decomposition can start from top tasks and doesn't finish until basic actions which are indecomposable or don't need to decompose. In the example of HIMS, daily business in a hospital includes some basic tasks such as "policlinic", "hospitalization", "pharmacy management" as well as support tasks such as "administration", "common service". These tasks can be decomposed into more concrete sub-tasks. For example, father task of "policlinic" can be decomposed into sub-tasks such as "registration", "diagnose", "treating", "auditing", "payment" and "drug taken". There are "Part-of" relations between father task and sub-tasks denoted with solid diamond arrows. Furthermore, "diagnose" can be classified as "image diagnose" and "manual diagnose". There are specialization relations between father task and sub-tasks denoted with a hollow triangle arrows. Figure 13.4 shows that.

2. Activity flow modeling

In order to describe the performance details of of a task, activity flow is used to show the performance order of sub-tasks and actors who participate in the sub-tasks, and resource the sub-tasks utilize, and events triggering the sub-tasks or generated by the sub-tasks, position where the sub-tasks perform. Of course, goals means-ended by tasks and domain properties or organizational rules constraining tasks should be analyzed in the course of modeling. Figure 14.5 shows activity flow modeling for policlinic. Activities of "policlinic" include "registration", "diagnose", "auditing", "payment" and "drug taken". Input of "registration" activity is a medicine insurance card; output of "registration" activity is a registration bill. The medicine insurance card and registration bill are all resources. If a patient wants to perform "policlinic" activity, he or she should make a registration with his or her medicine insurance card in a registration room firstly. The registration room is the position of performance of "registration" activity. Then an auditing clerk accomplishes the "registration" activity throughout medicine insurance policlinic system according to the patient's medicine insurance card. The patient, auditing clerk and medicine insurance policlinic system are all actors who perform the "registration" activity cooperatively.

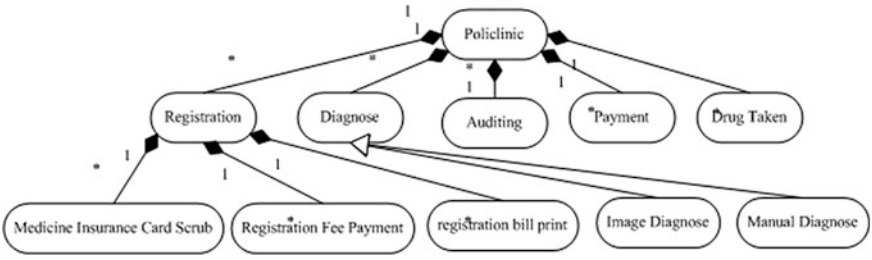


Fig. 14.4 Tasks decomposition for policlinic

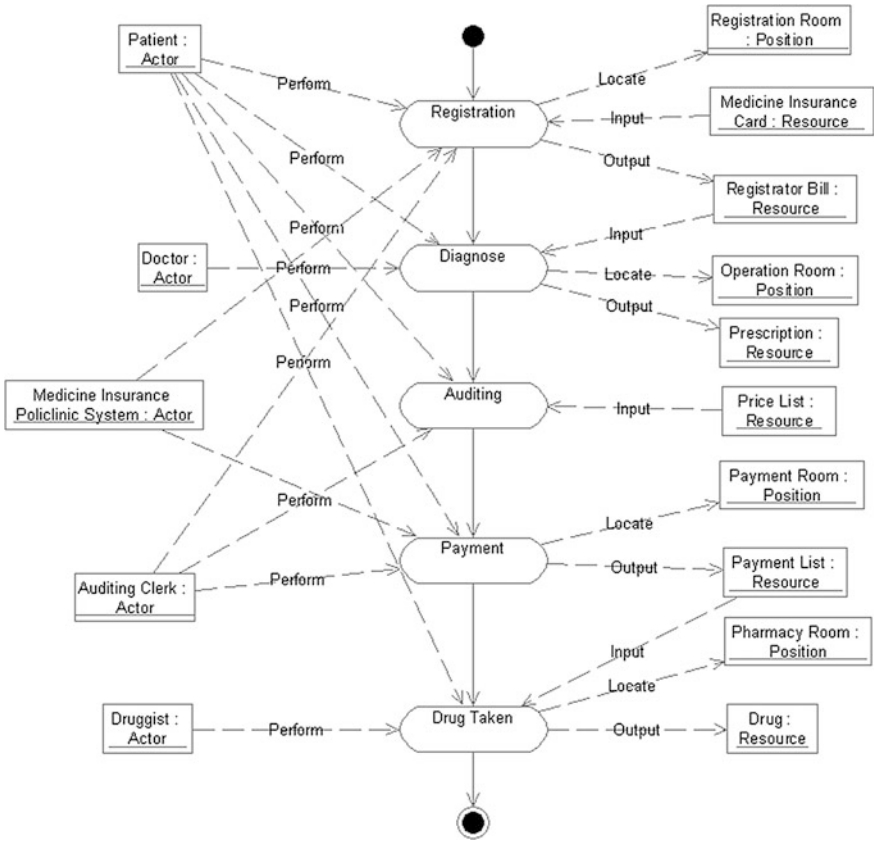


Fig. 14.5 Activity flow modeling for policlinic

(3). Goal View Modeling

Goal view describes intentional characteristics of an enterprise, which includes one kind of models. That is the goal model. Goal model is often hierarchical which is constructed directly from some high level goals to low level goals. Of course, domain properties or organizational rules constraining achievement of goals should be analyzed.

1. Construct elementary goal model in goal view

Simultaneously with constructions of business view and organizational view, goal view is constructed by analysis of business documents and talking with stakeholders. In goal view, some high level business goals about system-to-be can be found firstly from stakeholders, especially from enterprise administrators. An elementary goal model can be gradually constructed in a top-down way by decomposing higher level goal into lower level goals. This goal model may be incomplete and new goals may be found indirectly in the other two views

meanwhile. For example, a high level goal "Policlinic Treatment" can be refined as sub-goals "Registration", "Work-up" and "Drug Taken" by a "And" decomposition way. The sub-goal "Registration" can be refined farther as goal "identity authentication" and goal "Registration Bill Gotten". The sub-goal "Work-up" can be refined farther as goal "Diagnose" and goal "Prescription Gotten". The sub-goal "Drug Taken" can be refined farther as goal "Payment identification" and goal "Drug gotten". In Fig. 14.6, rectangles and lines decorated with black color show the elementary goal model. Hereinto, dotted lines with arrows labeled with "≪ And ≫" represent "and" decomposition between father goal and sub-goals.

2. Elicit goals from organizational view

Because "bear" relation between actor and goal exists in meta-model presented above, some goals can be elicited heuristically by asking actors why their dependency relations are shaped after organizational structure model and dependency model are constructed basically in organizational view. In this way, some new goals may be found and supplemented into goal view. By analysis of actor dependency relationships just described in Fig. 14.3, some goals can be elicited such as "Registration", "Auditing Prescription", "Payment Identification", "Diagnose By Image", "Diagnose By Manual", "Nursing Treatment", "Prescription Gotten", "Drug Gotten", "Registration Bill Presentation", "Payment Bill Presentation", "Identity Authentication". Five new goals appear except the goals appear in the elementary goal model such as "auditing prescription", "registration bill presentation", "payment bill presentation", "diagnose by image" and "diagnose by manual", which should be correspondingly supplemented into the elementary goal model. In Fig. 14.6, rectangles and lines decorated with red color show that. Hereinto, Real lines with triangle arrows labeled with "≪ Or ≫" represent "or" decomposition.

3. Elicit goals from business view

Because "means-end" relation between task and goal strictly exists in meta-model presented above and there is no reason for existence of task which has no any purpose, goals can be elicited heuristically by asking what purpose every task wants to achieve. In business view, business model can be constructed by continual decomposition of tasks from high level to low level. Accordingly, high level goal can be elicited by high level task, and low level goal can be elicited by low level task. In this process, some new goals may be found and supplemented into goal view. By analysis of tasks decomposition described in Fig. 14.4 and activity flow in Fig. 14.5 for policlinic, some goals can be elicited such as "Policlinic Treatment", "Registration", "Diagnose", "Auditing Prescription", "Payment Identification", "Drug Gotten", "Identity Authentication", "Registration Bill Presentation", "Registration Payment Identification", "Diagnose By Image", "Diagnose By Manual". Only the goal "Registration Payment Identification" is new in current goal model, which must be supplemented into the goal model. In Fig. 14.6, rectangles and lines decorated with green color show that.

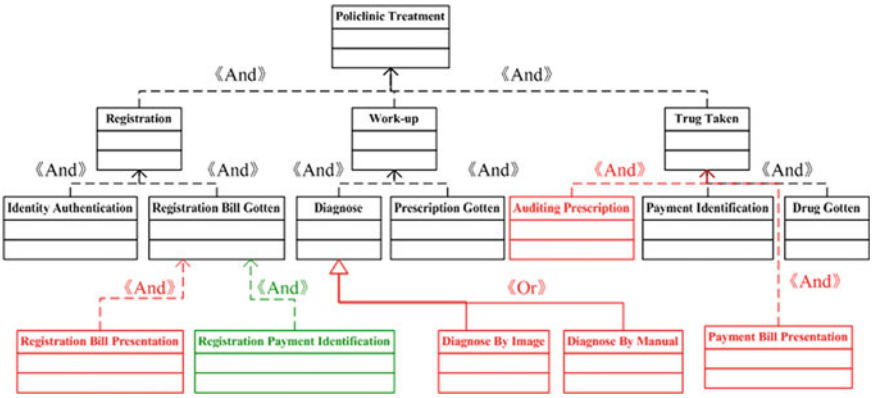


Fig. 14.6 Goal model for policlinic treatment

14.4 Conclusion

For information system requirements modeling, multiple views can be used to describe the characteristics of the enterprise, which commonly include organizational view, business view and goal view. With guidance of meta-model proposed in this paper, analysts can not only directly construct goals in a top-down way, but also heuristically elicit goals from organizational view and business view. With this method, analysts can construct goal model completely.

References

1. Dardenne, A., van Lamsweerde, A., Fickas, S.: Goal-directed requirements acquisition. Sci. Comput. Progr. (0167–6423) **20**(3), 3–50 (1993)
2. Kavakli, E.: Modeling organizational goals: Analysis of current methods. Symposium on Applied Computing, Proc. 2004 ACM symposium on applied computing, pp. 1339–1343 (2004)
3. Castro, J., Kolp, M., Mylopoulos, J.: Towards requirements-driven information systems engineering: the Tropos project. Inf. Syst. **27**(6), 365–389 (2002)
4. Anton, A.I.: Goal identification and refinement in the specification of software-based information systems. Ph.D. Dissertation, Georgia Institute of Technology, Atlanta (1997)
5. Chen, B., Wang, Z.-X., et al.: Requirement analysis method based on three-dimensional classification approach of goal modeling. J. Syst. Simul. **20**(15), 3986–4005 (2008) (In Chinese)
6. Chen, B., Wang, Z.-X., An algorithm modeling goals precedence relations based on transferability closure. Syst. Eng. Elect. **31**(2), 463–467 (2009) (In Chinese)

Chapter 15
Environment Monitoring System Based on Internet of Things

E. Tang, Fu Chen and Quanyin Zhu

Abstract In order to obtain the remote environmental parameters, the sensor module and the mobile software is designed for a new environment monitoring system based on Internet of Things (IOT) in this paper. SHT10 digital humidity, temperature sensor and TSL2561 light-to-digital converter are used to get the environmental parameters such as the temperature, the humidity and the luminance. IOT technology is used to transform the parameters data to remote server through GPRS network, and SIM900 chip is opted to accomplish GPRS function. Ajax and mobile platform are combined to make sure that the environment data on the server can be easily accessed. Experiments demonstrated that by using Ajax in the web application, the time to reach the information reduced a lot, and the users can monitor the environmental parameters easiest via their mobile cell.

Keywords Environment monitoring · Internet of things (IOT) · Ajax · Mobile platform

15.1 Introduction

Production environment monitoring and control is very important in realizing industrial automatization and high efficiency. With the development of the Internet of Things, currently, most environment monitoring systems are using a distributed

E. Tang · F. Chen · Q. Zhu (✉)
Faculty of Computer Engineering, Huaiyin Institute of Technology, Huaian, China
e-mail: topchenfu@qq.com

E. Tang
e-mail: hyitzqy@126.com

F. Chen
e-mail: 34353659@qq.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_15,

framework [1]. Wireless Sensor Network (WSN) technology is used in some other monitoring systems [2]. But WSN network's transmit distance isn't very long, the problem doesn't exist in General Packet Radio Service (GPRS) network because GPRS network based on GSM, so the transmit distance is almost unlimited. GPRS is the shorted form of GPRS, it is a breakthrough of GSM net-works only circuits witched provided thinking mode, and realize packet switching only by increasing the functionality entities and transforming part of the existing base- station system [3].

Recent years, with the fast development of the mobile platform, the mobile platform has become the most popular way to obtain information people need.

Ajax technology is the most important way to design a Web 2.0 application. Web 2.0 enables the design of highly interactive User Interfaces (UIs) for web applications [4].

15.2 Principle of Ajax

Ajax is an acronym for asynchronous JavaScript and XML. With Ajax, Web applications can send data to, and retrieve data from, a server asynchronously (in the background) without interfering with the display and behaviour of the existing page.

The comparison between the traditional mode and Ajax mode for Web applications is shows in Fig. 15.1.

In the traditional web application, user's actions in the interface trigger an HTTP request back to a web server. The server does some processing and then returns an HTML page to the client with a HTML page format. While user has a

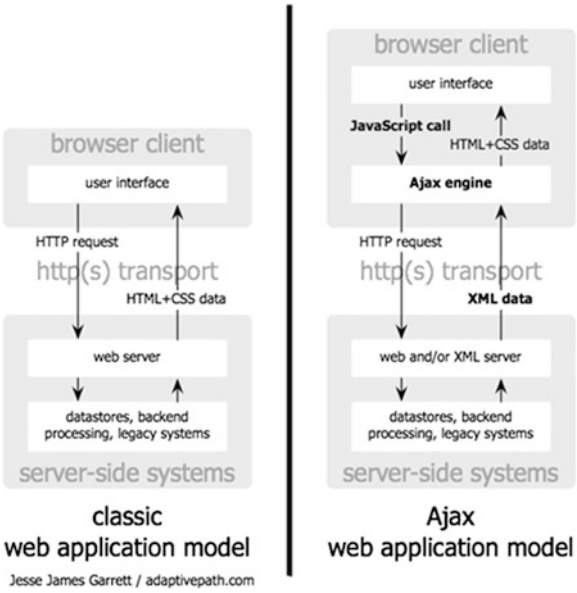


Fig. 15. 1 Comparison between the traditional mode and Ajax mode

request, the server will return a whole page to handle the request, then the user has to wait for a long time even if there is only a few simple data interaction [5].

In the application based on Ajax, when the user client needs some little data, Ajax engine will trigger request and the server return the client end response with XML format or JSON format. Ajax engine only refresh those part of the page which are changed by using DOM object model, thereby saving a great deal of time and enhancing the work efficiency [2].

15.3 The Android Mobile Platform

Android is a Linux-based operating system primarily designed for mobile devices such as smart phones and tablet computers utilizing ARM processors. It is developed by the open handset alliance, led by Google. Android became the world's leading smart phone platform at the end of 2010 [6]. For the first quarter of 2012, Android had a 59 % smart phone market share worldwide [7].

Because of the android devices and other smart phones, mobile platform has become the largest way to obtain information; the original PC lost its position because of the big size and weight. People now can get the information they want more efficiently by using smart phones like android devices.

15.4 IOT Technologies

IOT is the sensor equipment to the power grid, railways, bridges, tunnels, roads, buildings, water supply systems, dams, oil and gas pipelines as well as household appliances and other real objects, to link up via the Internet, and then run a specific program to reach remote control or direct communication between things. IOT is connected through the interface with the wireless network through the device in the various types of objects on the Radio Frequency Identification (RFID), sensor, two-dimensional code, "smart" objects, people and objects communicate and dialogue, communication and dialogue between objects one to another can also be achieved, this object linking network known as the "Internet of Things (IOT)".

IOT extends Web 2.0 but clearly raises the question of our ability to develop ever more powerful tools. Either object becomes "actors/partners" acting under our control with their associated software intelligence: that is to say not only assistants but especially counsellors, policy makers, organizers or economic agents.

15.5 System Design

The proposed environment monitoring system is separated by two parts, one is the hard-ware which contains the sensors, the microprocessor and GPRS module, the other is the software, which contains a B/S based platform designed by ASP.NET, and a mobile application for Android platform.

The general architecture of the system is shown in Fig. 15.2.

The environmental parameters are gained by the sensor module, than the data would sent to the remote server thought the Internet using GPRS technology, the user could get the update environmental parameters by using PC or a smart phone, and the administrator could add or delete the user.

15.5.1 System Flow Chart

The proposed system working produce is shown in Fig. 15.3.

When the system running, the environmental parameters was gained by the sensor module, the original parameters data was sent to the remote server through the GPRS network, the data handled by the server and stored in the database, the web services also provided by the server, so the environmental parameters is easily obtained by the users through the web or the mobile platform.

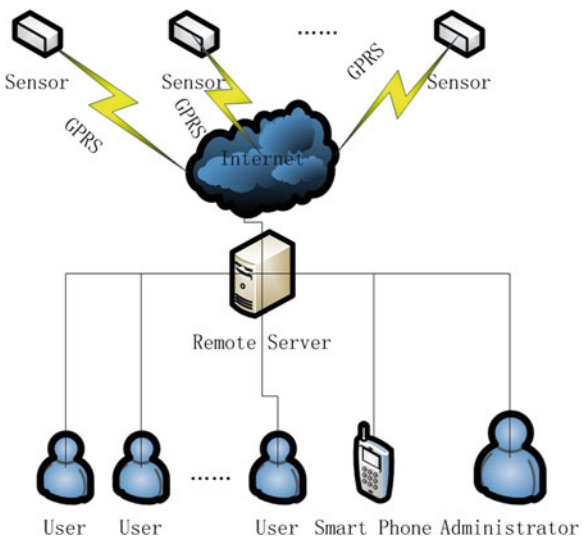


Fig. 15.2 The general architecture of the proposed system

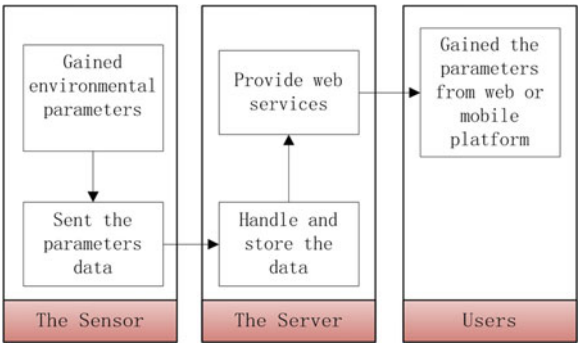


Fig. 15.3 The flow chart of the proposed system

15.5.2 Detail Design

15.5.2.1 The Sensor Module

The sensor module of the system contains a MCS-51 microprocessor, a SHT10 digital humidity and temperature sensor, a TSL2561 light-to-digital converter and a SIM900 GPRS module, the architecture of the module is shown in Fig. 15.4.

SHT10 digital humidity and temperature sensor is the low cost version of the reflow solderable humidity sensor series [8]. It provides a high-accuracy RH and temperature measurement. It has a 0–100 % RH of RH operating range and a −40 to +125 °C (−40 to +257 °F) of temperature operating range.

TSL2561 light-to-digital converter provides a 1-40000Lux luminance measure range [9].

SIM900 is a quad-band GSM/GPRS module that works on frequencies GSM 850 MHz, EGSM 900 MHz, DCS 1800 MHz and PCS 1900 MHz. SIM900

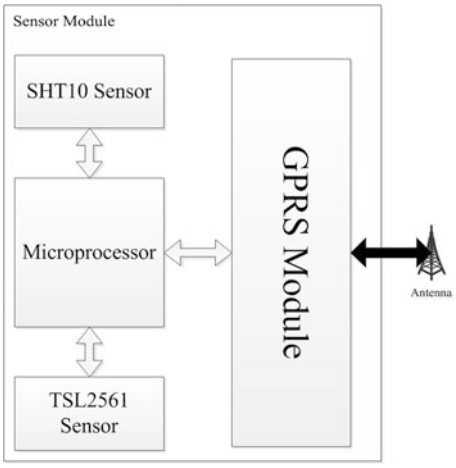


Fig. 15.4 The architecture of the sensor module

features GPRS multi-slot class 10/class 8 (optional) and supports GPRS coding schemes CS-1, CS-2, CS-3 and CS-4 [10].

The charge of controlling the two sensors to gain the environmental parameters is taken by MCS-51 microprocessor, the information will sent to the remote server though GPRS module over a RS-232 interface.

The environmental parameters was transmitted by GPRS network, so the power consumption is an important problem to the system. In the sensor module, there was a trickle charge timekeeping chip called DS1302, the chip is used to provide the real time, and the users can set the sampling frequency of the sensor module. And the whole module was in sleep mode between the two sampling processes. Besides these the sensor module also has a lithium battery and a photovoltaic cell to ensure the power supply of the whole system.

15.5.2.2 The Web Application

The web application is built by ASP.NET and SQL Server 2008 on Windows. It's separated into two parts: one of them is a Windows service which runs background to handle the data sent by the sensor module: the other one is the front web application which provides several functions to the user.

The functions provided by the web application are shown in Fig. 15.5.

At the sensor overview page, all the active sensors will be displayed in the content. When the mouse moved on some sensor's area, the sensor tracking dialog will be displayed, and the display effect is provided by a jQuery plug-in called power float. Ajax engine gets the latest sensor data from the remote server, then display the dialog to show the most recently environmental parameters from the sensor module.

The display style of the floating dialog is shown in Fig. 15.6.

There are two user groups in the web application, one is the administrator, the other is the common user, the administrator has all the permissions of the system, like add a sensor or a user, and only the administrators have access to change the

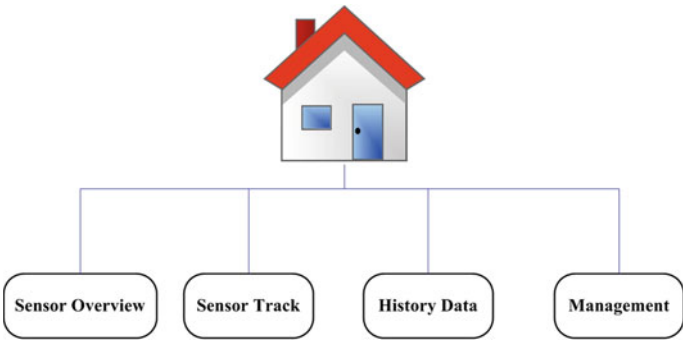


Fig. 15.5 The functions of the web application

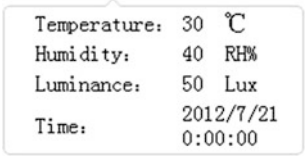


Fig. 15.6 The display style of the floating dialog

system configuration like sampling frequency and the system local time, etc. The common users only have the permission to access the environmental parameters. The username and the password are stored in the database, and the password is MD5-encrypted.

15.5.2.3 The Android Application

The newest version now is Android 4.1, considering the version compatibility, Android 4.1 may not be a good choice to develop application, and Android 2.3 was used at a wider range, so Android application was developed for Android platform 2.3, which means the application can run on devices running Android 2.3 or later versions.

Several functions are provided by the application, the main function is to check the sensor's data when the user logged in the application and choose the sensor. And the sensor's status and the update environmental parameters will be shown at the screen. The running status of the application is shown in Fig. 15.7.

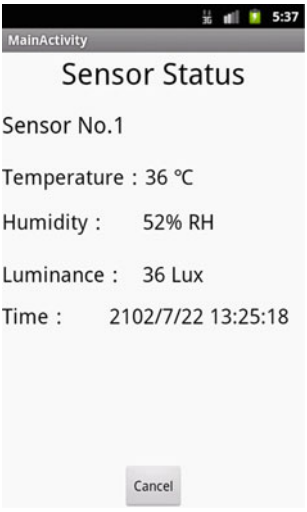


Fig. 15.7 The android application

15.6 Conclusion

Depend on hardware and software design, mobile platform subverts the way people obtaining information. It provides an easier way to get needed information than PC. With development of sensors, MCU, Web 2.0, Ajax, and IOT technologies, researchers proposed an environmental parameters monitor system which can support users. Users only need to take more care of the experience instead of the information itself. This paper makes an easier and practical way to obtain the environmental parameters by mobile cell.

References

1. Wu, L., Hu, J.: Design and Implementation of Production Environment Monitoring System Based on GPRS-Internet. 2010 Fourth International Conference on Genetic and Evolutionary Computing, pp. 818–821 (2010)
2. Han, W., Fang, K.L., Li, X.H., et al.: Ajax Applied in Environment Monitoring System based on WSN., International Symposium on Computer Science and Computational Technology, pp. 608–611 (2008)
3. He, H.J., Yue, Z.Q., Wang, X.J.: Design and Realization of Wireless Sensor Network Gateway Based on ZigBee and GPRS. Second International Conference on Information and Computing Science, pp. 196–199 (2009)
4. McIntosh, S., Adams, B., Hassan, A.E., et al.: Using indexed sequence diagrams to recover the behaviour of AJAX applications. 13th IEEE International Symposium on Web Systems Evolution, pp. 1–10 (2011)
5. Jesse, J., Garrett, A.: A new approach to web applications. http://www.adaptivepath.com/ideas/ajax-new-approach-web-applications. Available at February 18, 2005 (2008)
6. Palo, A.: Google's Android becomes the world's leading smart phone platform http://www.canalys.com/newsroom/google%E2%80%99s-android-becomes-world%E2%80%99s-leading-smart-phone-platform Post at Monday (2011)
7. New York (PRWEB): Android Smartphone Activations Reached 331 Million in Q1'2012 Reveals New Device Tracking Database from Signals and Systems Telecom http://www.prweb.com/releases/2012/5/prweb9514037.htm, Post at May 16, 2012 (2012)
8. Sensirion: Datasheet SHT1x. http://www.sensirion.com/ pp. 1–2 (2012)
9. TAOS Inc: TSL2560, Tsl2561 Light-To-Digital Converter. http://www.taosinc.com/. pp. 1–2 (2009)
10. SIMCom Wireless Solutions: SIM900 Hardware Design V_2.02 http://wm.sim.com/upfile/20126416756f.pdf. 7 May 2012 (2012)

Chapter 16
Utility Theory Based Approach for Converged Telecommunications Applications

Muhammad Athar Saeed, Li Jian and Sadia Murawwat

Abstract Convergence of heterogeneous products and services is now a leading phenomenon of telecommunications industry. These new and innovative converged applications are able to meet the diversified but pooled demands of the subscribers, so their demand is rapidly increasing. In this paper researchers have done both qualitative and quantitative analysis to find out subscriber satisfaction levels for these services. This analysis is based on utility theory, probability theory and Bayes theorem. The aim of this analysis is to find out subscribers' satisfaction levels and their willingness to pay for converged applications for decision making by manufacturers and service providers. The research found that cost effective converged application with maximum satisfaction is preferred by consumers and will play leading role in expansion of this market in future.

Keywords Utility theory · Convergence · Subscriber's satisfaction · Telecommunication applications

16.1 Introduction

Evolution for refinement driven by the motive of to bring more ease and comfort in human lives is as old as existence of human beings. Inventions by man are good example of it. In telecommunications sector, span of thousands of years from

M. A. Saeed (✉) · L. Jian
School of Management and Economics, Beijing Institute of Technology, Beijing, China
e-mail: athsaeed@yahoo.com

L. Jian
e-mail: lijianbit@bit.edu.cn

S. Murawwat
Department of Electrical Engineering, Lahore College for Women University, Lahore, Pakistan
e-mail: sadiamurawwat@yahoo.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_16,

ancient times' conveying messages through fire lighting and drum beating to telegraphy in the recent past and modern mobile telephony and other state of the art telecom products in today's world contains the quest for evolution for refinement for ease and comfort. Since last two decades, huge developments have been made in this sector. Analog devices evolved to digital, wired to wireless, then wireless extended to mobile and nomadic access, narrowband transmission to broadband [1]. Many technologies play significant role in this mobile evolution. xG continues its developing process to meet new requirements from 1G, 2G, 2.5G, 2.75G, 3G and 3G+. WiFi, WiMAX, LTE all are strong competitors in this struggle. These all advancements have now changed dimension to the phenomena of convergence. However for telecommunications operators, it is continuously evolving subscriber's demand to communication, customization, convenience and personalization [2].

Our research is on the application layer and in the next two sections we discuss convergence and the vital players of this layer. In Sects. 16.4 and 16.5, we have used utility theory approach to quantize the satisfaction level of our subscribers and results are mentioned in Sect. 16.6 while Sect. 16.7 concludes this paper.

16.2 Convergence

According to International Telecommunication Union (ITU) recommendation Q.1761, convergence is a mechanism by which an IMT-2000 user can have his basic voice as well as other services through a fixed network as per his subscription option [3]. In opinion of European Telecommunication Standards Institute (ETSI), Fixed Mobile Convergence (FMC) is concerned with the provision of network capabilities which are independent of the access techniques.

An important annex of this rule is related to inter-network roaming; subscriber should be able to roam between diverse networks and to be able to use the same steady set of services through those visited networks [4]. Some examples of mobile/WLAN convergence are:

- WLAN/GPRS 'Handover' declared by Nokia
- WLAN/GSM VoIP terminal publicized by Motorola
- NTT DoCoMo: FOMA–WiFi
- BT "Fusion": GSM–WiFi
- France Telecom "Business Anywhere": GPRS–WiFi
- O2 Germany "surf@home": UMTS–WiFi
- Korean KT & KTF "OnePhone": CDMA–Bluetooth
- Dual Phone, by Deutche Telekom's T-Com.

Convergence is depicted in Fig. 16.1.

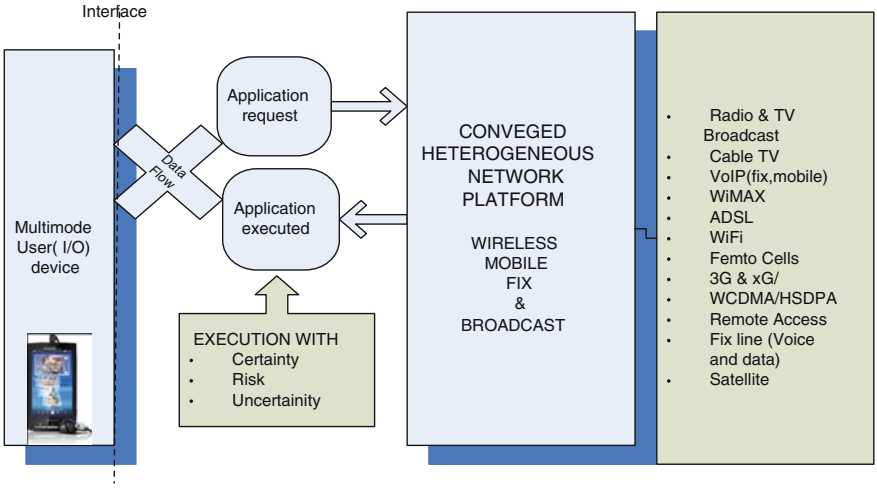


Fig. 16.1 Converged network platform

16.3 Key Players of Application Layer

In this section, we discuss three key players at application layer of any telecom environment namely subscribers, applications and devices measurement described.

Subscriber is any person availing the services of telecom network on payment. Their classification for research purpose is available in literature based on Age (Teenagers, Young, Adults, Old), Gender (Male, Female), Network access type (TDMA, CDMA, OFDMA), Network scale (PAN, LAN, MAN, WAN, GAN), Application type (Voice, Video, Combination of both), Technology (Fixed, Wireless, Mobile, Nomadic) and Market (Home, Enterprises, Corporate) but not limited to these.

Applications of telecommunication networks are being provided since 1970's starting from analog and then 1990's voice and in late 1990's with text and voice under 2.5G technology and progressing continuously 3G and 3G+. The drive behind is more and more data rate technically named as broadband. International Telecommunications Union (ITU) recommendation I.113 has defined: 'A broadband as a transmission capacity that is faster than primary rate ISDN at 1.5–2 Mbits/s' [5]. Federal Communications Commission (FCC) defined in 1999: 'Broadband is 200 kbits/s (0.2 Mbit/s) in one direction, the lower end of the broadband spectrum to be 200 kbits/s and advanced broadband is at least 200 kbit/s in both directions' [6] however in 2010, it defined as "high-speed, switched, broadband telecommunications capability that enables users to originate and receive high-quality voice, data, graphics, and video telecommunications using any technology" [7].

Future network requires multimode device that can be connected and switched between networks based on different technologies, including cellular, WiFi, DSL/Cable, Bluetooth, UWB and more. Qualcomm is behind the multimode dimension

by developing power efficient chipsets that integrate a growing array of functionality and supporting many access technologies while LG and Samsung are launching their own chipset. Multimode devices have a key position in the commercialization of LTE. Commercial launch of LTE with multimode devices is taken off.

16.4 Utility Theory Based Approach

It is human instinct that possession and consumption of goods give pleasure or satisfaction. In economics, this pleasure or satisfaction is named utility. Jeremy Benthan (1784–1832) gave the idea to measure pleasure and pain in his famous book 'Introduction to the Principals of Morals and Legislation' in 1780s which laid the foundation stone of the utility theory and this theory was further developed by William Stanley Jevons and others in mid 19th century. The theory assumes/postulates that consumer derives satisfaction from goods, wants to maximize the satisfaction, income is limited, and the satisfaction increases as per law of diminishing marginal utility and utility is measureable [8]. It is human behavior that he wants to maximize the satisfaction but his income is constrained. To analyze this behavior, the concepts of indifference curve and budget constraint in economics are helpful.

Indifference curve shows different combinations of two commodities that gives same level of satisfaction or in other words set of various quantities of two goods which are equally desirable. The indifference curves have the properties: i. farther from origin preferred to closer ones, ii. There is an indifference curve through every possible bundle, iii. Indifference curves cannot cross and iv. Indifference curves slope downward [9].

Budget constraint is important concept in utility theory and shows various combinations of goods that a consumer can buy given his income and price of goods. In [10] budget line is defined as 'a line showing the possible combinations of two goods that a consumer can purchase'.

After mapping the indifference curves and budget line, we can find out the point where consumer satisfaction is maximized. Graphically it will be the intersection of budget line and the farther most indifference curve from origin.

16.5 Applying Utility Theory Approach to Convergence of Telecom Net-work Application

In view of human behavior of maximizing the utility given the limited resources, a telecom subscriber would naturally want to get more and more benefit and satisfaction from telecom application, device or network by spending certain amount of

Table 16.1 Wireless mobile application and sample data rate

Application	Data rate
VoIP	4–64 kbps
Interactive gaming	50–85 kbps
Music, Speech, Video Clips	5–384 kbps
Web browsing, email, instant messaging, telnet, file download	0.01–100 Mbps
IPTV, movie download, P2P video sharing	>1 Mbps

money and utility theory help here to find the solution. For our purpose, some important factors that contribute to the approach and analysis are following.

The term 'subscribers' is used instead of consumers and they are of three types i.e., home or individuals, enterprise and corporate. Each kind spends a range of their income for the purpose of execution of applications. A general framework for their respective budget has been taken for analysis (Table 16.1).

Voice and video are taken as two goods for our utility analysis and their different combination encompasses various applications' demands. Satisfaction level means the quantified utility level. Our range of satisfaction is depicted in following table.

Satisfaction levels are classified as 'E', 'G','F' and 'P'. In first type, the bandwidth and data rate required by the application request is equal or greater than needed. There is no delay and all parameters for Quality of Service (QoS) are maintained by network. Hence leading to conclusion that subscriber satisfaction level is 'E'. In second and third alternative, satisfaction level is not as much of 'E' but still satisfaction and is denoted as 'F' and 'G' in Table 16.2. 'P' type subscriber is unsatisfied.

Indifference Curves are named as satisfaction curves in this analysis and shows different combinations of two commodities that gives same level of satisfaction. When a subscriber's satisfaction level changes, the curve also changes otherwise not. These curves for satisfaction levels 'E', 'G' and 'F' are shown in Fig. 16.2 and labeled as 'SE', 'SG' and 'SF' respectively. As long as the subscriber has same satisfaction level, his curve is same. These curves never touch zero levels. 'SE', 'SG' & 'SF' hold all combination of video and voice units at respective satisfaction levels of 'E', 'G' and 'F'. At level 'P', there is dissatisfaction, so, satisfaction curve is not drawn. For plotting these curves, the different combinations of voice and videos are used and units for both goods i.e., video and voice are kbits/sec.

Table 16.2 Wireless mobile application and sample data rate

Level		Satisfaction status
00	Poor (P)	Unsatisfied
01	Fair (F)	Partially satisfied
10	Good (G)	Partially satisfied
11	Excellent (E)	Satisfied

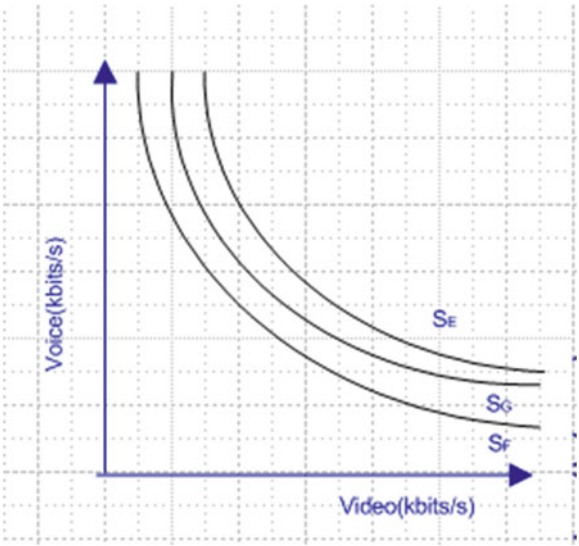


Fig. 16.2 Satisfaction curves

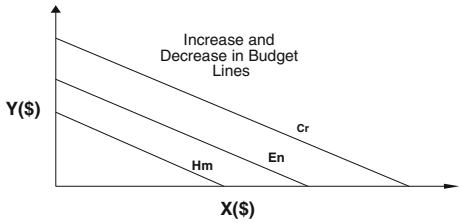


Fig. 16.3 Budget lines

Budget constraint is very important in telecommunications. If the applications are executed with good quality of service and cheap rates then it is attracted by common subscribers so it is a significant factor in implementing and penetration of any network. If the converged network is providing all its services at very high rates that are not affordable to its subscribers then convergence phenomenon for wireless mobile applications will not be a success story. Figure 16.3 shows budget line for purchase of two commodities namely voice (V_c) and video (V_d):

Let 'N' is the amount in dollars for purchase and the cost/kbps of Vc is PVc and cost/kbps of Vd is as PVd. The straight line connecting N/PVc and N/PVd will be the budget constraint of a subscriber with defined 'N' dollar for purchase of units of Vd and Vc. This line shows the combination of Vd and Vc that a subscriber can get in the given budget. Different subscribers such as individuals, enterprises and corporations have different budget constraints for access of broadband applications and their budget lines are straight with certain ratios of voice and video units.

Optimal solution to choose between different combinations of voice and video given the budget constraint of the subscriber lies at intersection point between budget line and satisfaction curve. In Fig. 16.4, points 'x', 'y' and 'z' contains different quantities or units of voice and video but they give same level of

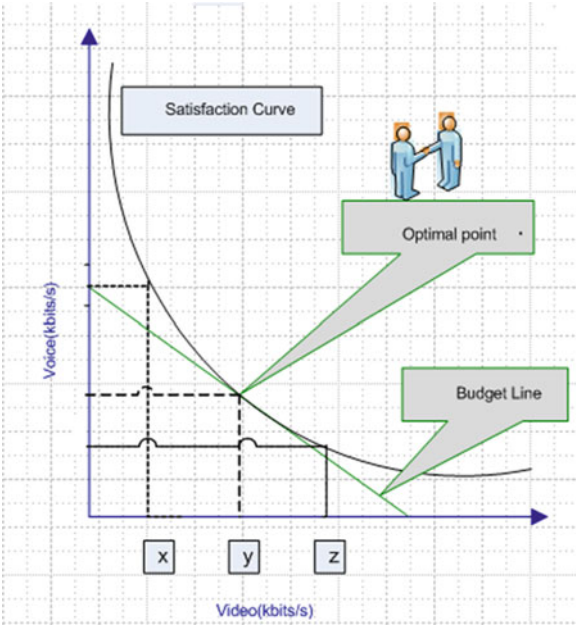


Fig. 16.4 Satisfaction maximization point

satisfaction to the subscriber as all three points lie on the same satisfaction curve. However, Point 'y' is intersection point with budget line while points 'x' and 'z' lies above the budget line and hence not attainable at this level. Therefore subscriber will opt for point 'y'.

16.6 Analysis and Results

Analysis is done in this section to aid our idea of quantizing satisfaction through utility theory. Analytical approach is scientific aspect that leads to decision making. Previous section shows the qualitative analysis, whether subscriber is satisfied, unsatisfied or partially satisfied depending upon the execution level of his/her request. Once we know the preference level of the particular group of subscribers, it is helpful for the telecom service provider to design and offer products keeping in view of the groups' specific requirements and price level.

Now we quantify and analyze the relation of satisfaction and change in budget and first we define some probabilities in this regard.

P(S) = Probability of satisfaction, P(PS) = Probability of partial satisfaction, P(US) = Probability of un satisfaction, P(S|BI) = probability of satisfaction given the budget increase or Probability that increase in budget will be favourable for satisfaction, P(PS|BI) = probability of partial satisfaction given the budget increase or Probability that increase in budget will be favourable for partial

satisfaction, P(US|BD) = probability of un satisfaction given the budget decrease or Probability that decrease in budget will be favourable for un satisfaction.

Based on results obtained from survey, we have calculated the following probabilities:

i. Probability of satisfaction P(S) = 0.3
ii. Probability of un satisfaction P(US) = 0.2
iii. Probability of partial satisfaction P(PS) = 0.5
iv. Probability of satisfaction given the budget increase P(S|BI)P(S|BI) = 0.7
v. Probability of partial satisfaction given the budget increase P(PS|BI) = 0.2
vi. Probability of un satisfaction given the budget decrease vi. P(US|BD) = 0.6
vii. Joint probability of Satisfaction and budget increase
P(S&BI) = P(S|BI). P(S) = (0.7) (0.3) = 0.21
viii. Joint probability of partial Satisfaction and budget increase
P(PS&BI) = P(PS|BI). P(PS) = (0.2) (0.5) = 0.15
ix. Joint probability of un-satisfaction and budget decrease
P(US&BD) = P(US|BD). P(US) = (0.6) (0.2) = 0.12

16.7 Conclusion

The research concludes that the global adoption of converged telecom applications depends on many factors. Service operators, vendors and manufacturers aim to promote fixed and mobile converged applications. This utility theory based analysis highlights the fact that service provision without subscriber satisfaction and adoptability is nothing. The approach emphasizes on application layer platform, shedding light on one of the important issues of convergence-subscriber satisfaction. Less cost converged solution with maximum subscriber satisfaction would be deployed. In conclusion, subscriber preference, contentment and inclusion of budget factor are important aspects to be considered in making telecommunication converged application thriving.

References

1. Mohr, W., Konhauser, W.: Access network evolution beyond third generation mobile communications. IEEE Commun. Mag. **38**(12), 122–133 (2000)
2. Pelt, M.: Convergence of Media and Telecom, NEM & Integration of Broadband and Multimedia. Alcatel Research and Innovation NEM Workshop (2004)
3. International Telecommunications Union: Principles and requirements for convergence of fixed and existing IMT 2000 systems. Recommendation Q.1761-ITU-T (2004)
4. Agung, W.: The APT wireless forum interim meeting. Document No. AWF-IM2/32, Asia-Pacific Telecommunity (2006)

5. International Telecommunications Union: Principles and requirements for convergence of fixed and existing IMT 2000 systems. Recommendation I.113-ITU-T (1997)
6. Federal Communications Commision: Inquiry concerning the deployment of advanced telecommunications capability to all Americans in a reasonable and timely fashion, and possible steps to accelerate such deployment pursuant to Section 706 of the telecommunications Act of 1996 [Report].CC Docket No. 98–146, para. 20 (1999)
7. Federal Communications Commission: Inquiry concerning the deployment of advanced telecommunications capability to all Americans in a reasonable and timely fashion, and possible steps to accelerate such deployment pursuant to Section 706 of the telecommunications Act of 1996 [Report]. GN Docket No. 10–159, para 4 (2010)
8. Garb, G.: Micro-economics theory, Applications, Innovations. Macmillan publishing Co., Inc., New York (1981)
9. Perloff, J.M.: Microeconomics, 4th edn. China Machine Press, China (2008)
10. Goodwin, N., Nelson, J.A., Ackerman, F., Weisskopf, T.: Microeconomics in context, 2nd edn. ME Sharpe, New York (2008)

Chapter 17
Open the Black Box of Information Technology Artifact: Underlying Technological Characteristics Dimension and its Measurement

Yuan Sun, Zhigang Fan, Jinguo Xin, Yiming Xiang and Hsin-chuan Chou

Abstract Previous research suggests the great importance of scrutinizing the information technology (IT) artifact itself. Yet, there is still limited attention to meet that goal. In this research, authors employed multiple research methods to investigate the technological characteristics dimensions, mainly including grounded theory method, focus group in-depth interview, content analysis, panel expert judgments and survey method. The results indicate that the complexity, utilitarian, hedonic, communication, flexibility, reliability, integration, accessibility, timeliness, accuracy, completeness, currency, format and stability are the critical and universal technological characteristics dimensions. Corresponding measurement items were also developed and illustrated. The findings have the potential to be applied in future investigations on the technological characteristics of IT artifact.

Keywords Information technology artifact · Technological characteristics · Users' IT acceptance and usage · Dimension and measurement

Y. Sun (✉)
School of Business Administration, Zhejiang Gongshang University, Hangzhou, China
e-mail: d05sunyuan@zju.edu.cn

Z. Fan
Alibaba Business School, Hangzhou Normal University, Hangzhou, China
e-mail: learvan@126.com

J. Xin
Department of Economics, Hangzhou Dianzi University, Hangzhou, China
e-mail: profxjg@163.com

Y. Xiang
School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou, China
e-mail: china8848@sohu.com

H. Chou
Inventec Appliances (Pudong) Corporation, Shanghai, China
e-mail: Jou.Galaxo@iac.com.tw

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_17,

17.1 Introduction

IT usage research has examined a wide range of technologies, such as word processing software, electronic mail, spreadsheets, microcomputers, world wide web, expert system, debugging software, telemedicine, group support system, computerized physician order entry systems, web-based transactional system, blog system, content management system, enterprise resource planning system and etc. However, these studies have not differentiated different kinds of information technologies (IT) artifacts. Hence, the underlying technological characteristics remain taken for granted, unspecified, unexamined, and under-theorized.

Orlikowski and Iacono [1] lamented the lack of theorization of the IT artifact in the information system (IS) field, and then identified five categories of information technology conceptualizations based on the coding of the 188 articles: the tool view, the proxy view, the ensemble view, the computational view, and the nominal view. The ensemble view was the most comprehensive one, focusing on the dynamic interactions between people and technology—construction, implementation, organization and the deployment of technology in society. In this paper, we adopt the ensemble view to investigate the underlying technological characteristics dimension of IT artifact and develop its corresponding measurement.

17.2 Literature Review

Although the IS researchers have realized the importance of theorizing and incorporating IT artifacts with current theories in IS research, only a few explicit attempts have been made. Leifer [2] pointed out that computer-based information systems could be categorized into decentralized systems, centralized systems, distributed systems and stand-alone systems. Ngwenya and Keim [3] distinguished group decision support system as synchronous and asynchronous one. King and He [4] categorized types of information system usage into job-related, office, general (such as email and telecom) and internet and e-commerce. Osrael et al. [5] classified the service-oriented systems into stateful and stateless categories. Lau et al. [6] considered that the negotiation agents system can be divided into adaptive and non-adaptive classes.

Based on the literature review, we have found that most of the current studies just employ a simple category to explore the black box of information technology artifact. Hence, the underlying technological characteristics dimensions of IT artifact and their measurement items are needed to investigate. As we know, different information system serves different purposes in our personal, social, and work lives. For example, enterprise resource planning systems help to improve our work performance, online games provide us with entertainment, videoconferencing improves us with communication or collaboration function, and etc. Depending on those purposes, we can conclude that technologies have three kinds

of technological characteristics: utilitarian, hedonic, and communication. Other potential technological characteristics can be considered are complexity [7], flexibility [8] and etc. Some information technologies serve multiple purposes, and have multiple technological characteristics. For example, group support systems can represent utilitarian by improving work performance, as well as communication by using as a tool for group members to discuss and collaborate with each other. Above identified technological characteristics from the literature review are the footstones for further investigations.

17.3 Research Methods

Based on initial potential technological characteristics identified from the literature review, we employed Ontology Theory (OT) [9] and General System Theory (GST) [10] to generate a series of important technological characteristics of IT artifact and establish their underlying dimension structure. We got twenty five important technological characteristics after above step. Then we used Grounded Theory (GT) [11] method to refine the dimension structure and the corresponding technological characteristics. After that, we got fourteen technological characteristics, and then conducted the focus group in-depth interview and content analysis to extract and purify the critical items to measure each technological characteristics dimension of IT artifact. And then we further refined these critical measurement items by integrating the measurement items from the related literatures and by panel expert judgments. Finally, we employed the pilot and large-scale survey to verify and refine the measurement items. We employed factor analysis to examine the validity of the measurement, and used Cronbach's alpha to examine the reliability.

17.4 Results and Discussions

The final technological characteristics dimension of IT artifact, their definition and measurement items are listed in the Table 17.1. The final technological characteristics are the ones that can be assessed by users. Other technological characteristics which cannot be assessed by users, such as scalability (refers to the extent to which provisioned computing resources can dynamically adjust to variable loads such as changes in the number of users, required storage capacity, and processing power [12]) are not included in this study.

To examine the technological characteristics dimension of IT artifact and their measurement items, we needed to reach out as many and diversified survey participants using different kinds of information technology as possible in the final large scale survey. Hence, we employed an online survey method through a popular local online survey web site. As an incentive, respondents who fully

Table 17.1 Technological characteristics dimension of IT artifact and their measurement

Technological characteristics	Definition	Measurement items from users' perspective	Sources
Utilitarian	Refer to the degree to which user believes that using a specific technology will assist in accomplishing his/her work goals	Using the information technology would enable me to accomplish tasks more quickly Using the information technology would improve my job performance Using the information technology in my job would increase my productivity Using the information technology would enhance my effectiveness on the job	[14, 15]
Hedonic	Refer to the value derived from enjoying the process of work	I would have fun interacting with the information technology Using the information technology would provide me with a lot of enjoyment It is interesting for me to use the information technology I would feel excited to use the information technology	[14, 15]
Communication	Refer to the value of facilitating communication, cooperation, and collaboration among a group of user	The information technology can facilitate me to communicate with other users The information technology can facilitate me to cooperate with other users The information technology can facilitate me to collaborate with other users	[16, 17]
Complexity	Refer to the degree to which a certain innovation is difficult to understand and use	The users' interaction with the information technology is clear and understandable It is easy for users to get the information technology to do what they want it to do Learning to use the information technology has been easy for users Overall, the information technology is easy to use	[7, 18]

(continued)

Table 17.1 (continued)

Technological characteristics	Definition	Measurement items from users' perspective	Sources
Flexibility/ Adaptability	Refer to the way the information technology adapts to changing demands of the user	The information technology can be easily adapted or extended to fulfill application requirements The information technology can be adapted to meet a variety of needs The information technology can flexibly adjust to new demands or conditions The information technology is versatile in addressing needs as they arise	[8, 19, 20]
Reliability	Refer to the dependability of information technology operation	I can depend on the information technology to provide me with the information I need I do not find information technology errors very often when I use the information technology The information technology has consistent information The information technology delivers what it promises	[21–23]
Stability	Refer to the robustness of the information technology	The information technology ensures data robustness (no loss of data) between the business processes Overall, the information technology is very stable I'm confident that the information technology is a stable system	[21]
Integration	Refer to the way the system allows data to be integrated from various sources	The information technology effectively integrates data from different areas of the company The information technology pulls together information that used to come from different places in the company The information technology effectively combines data from different areas of the company The information technology allows for integration with other systems	[8, 20]

(continued)

Table 17.1 (continued)

Technological characteristics	Definition	Measurement items from users' perspective	Sources
Accessibility	Refer to the ease with which information can be accessed or extracted from the system	The information technology allows information to be readily accessible to me The information technology makes information very accessible The information technology makes information easy to access I would find it easy to get access to the information technology	[8, 20]
Timeliness/ Response time	Refer to the degree to which the system offers timely responses to requests for information or action	It takes small elapsed time for the information technology to respond to my requests The information technology provides information in a timely fashion The information technology returns answers to my requests quickly	[8, 20, 21]
Accuracy	Refer to correctness of the output information provided by information technology	The information technology produces correct information There are few errors in the information I obtain from the information technology The information provided by the information technology is accurate	[8, 20]
Completeness	Refer to the degree to which the system provides all necessary information	The information technology provides me with a complete set of information The information technology produces comprehensive information The information technology provides me with all the information I need	[8, 20, 21]

(continued)

Table 17.1 (continued)

Technological characteristics	Definition	Measurement items from users' perspective	Sources
Currency	Refer to the degree to which the information is up to date	The information technology provides me with the most recent information The information technology produces the most current information The information from the information technology is always up to date	[8, 20]
Format	Refer to how well the information is presented	The information provided by the information technology is well formatted The information provided by the information technology is well laid out The information provided by the information technology is clearly presented on the screen Layout provided by the information technology is in good structure	[8, 20]

completed the questionnaire had a chance to win various prizes by lucky draw. Finally, we got a useful sample of 103 university students and 121 enterprise employees. Respondents were 55.36 % male and their average age was 33.9 years. To assess the reliability of the measurement instrument, we computed the Cronbach's alpha for all variables [13]. The values of all Cronbach's alpha in our study were between 0.89 and 0.94, thus passing the reliability test. The items for all variables were submitted to factor analysis to assess the validity of the measurement instrument. The extraction procedure was Principal Components using the varimax method for factor rotation. The analysis produced 14 components accounting for 62.23 % of the variance. There were no significant cross-loads, and the minimum load exceeded 0.71 for all items, indicating high construct validity. Each technological characteristic was composed of a single variable. Therefore, the reliability and validity of the measurement instrument conceivably indicated that our technological characteristics dimensions of IT artifact and their corresponding measurement items were of high quality, and could be used in the future study investigating the technological characteristics.

Based on all technological characteristics dimension of IT artifact identified above, authors classify them into two groups, i.e., triggers group and contextual constraints group. For triggers group (including flexibility, reliability, integration, accessibility, timeliness, accuracy, completeness, currency, format and stability), technological characteristics play the role of mediation in the interactions between users and technologies. The technological characteristics in this group emphasize that IT-supported task can be considered as an important factor influencing the user IT usage. And the technological characteristics can indirectly influence the users' acceptance and usage through the users' cognitive believes. For contextual constraints group (including technological complexity, utilitarian, hedonic and communication), technological characteristics play the role of environmental background. The technological characteristics in this group emphasize that users' IT usage can be influenced by information technology event and change, and the technological characteristics moderate the effect of users' cognitive believe on IT usage.

17.5 Conclusion

Through the grounded theory method, focus group in-depth interview, content analysis, panel expert judgments and survey method, this study systematically investigates the underlying technological characteristics dimension of IT artifact and develops its corresponding measurement. Final large scale survey result indicates that complexity, utilitarian, hedonic, communication, flexibility, reliability, integration, accessibility, timeliness, accuracy, completeness, currency, format and stability are the critical underlying technological characteristics dimensions of IT artifact. High reliability and validity indicates that the developed items are qualified to measure these critical technological characteristics. Future

studies could integrate these technological characteristics into the users' acceptance and usage models, and use corresponding measurement items to test these theoretical models. Furthermore, the technological characteristics identified in this study are critical and universal ones, which can be used both in individual information technology and enterprise information technology. Therefore, more technological characteristics could be identified for the specific information technology through examining the specific information technology in future study.

Acknowledgments This research was supported by Humanities and Social Sciences Research Project of the Ministry of Education of China (11YJC630189) and Zhejiang Provincial Natural Science Foundation of China under Grant (LQ12G02009). And this material is based upon work funded by China Postdoctoral Science Foundation Funded Project (2011M500105, 2012T50560) and National Natural Science Foundation of China (70972119). Besides, this research was supported in part by Zhejiang Provincial Philosophy and Social Sciences Project (12JCGL11YB) and Zhejiang Provincial Key Research Base–Decision Science and Innovation Management (RWSKZD02-201206) and the Contemporary Business and Trade Research Center of Zhejiang Gongshang University, which is the Key Research Institute of Social Sciences and Humanities Ministry of Education.

References

1. Orlikowski, W.J., Iacono, C.S.: Research commentary: desperately seeking the "IT" In it research-a call to theorizing the IT artifact. Inf. Syst. Res. **12**, 121–134 (2001)
2. Leifer, R.: Matching computer-based information systems with organizational structures. MIS Q. **12**, 63–73 (1988)
3. Ngwenya, J., Keim, R.: The effects of augmenting face-to-face meetings with a web-based asynchronous group support system. J. Inf. Technol. Theor. Appl. **5**, 47–62 (2003)
4. King, W.R., He, J.: A meta-analysis of the technology acceptance model. Inf. Manag. **43**, 740–755 (2006)
5. Osrael, J., Froihofer, L., Goeschka, K.: Replication in service-oriented systems. J. Softw. Eng. Fault Toler. Syst. **19**, 91–117 (2007)
6. Lau, R.Y.K., Li, Y., Song, D., et al.: Knowledge discovery for adaptive negotiation agents in e-marketplaces. Decis. Support Syst. **45**, 310–323 (2008)
7. Nadkarni, S.: A task-based model of perceived website complexity. MIS Q. **31**, 501–524 (2007)
8. Wixom, B.H., Todd, P.A.: A theoretical integration of user satisfaction and technology acceptance. Inf. Syst. Res. **16**, 85–102 (2005)
9. Bailey, K.D.: Typologies and Taxonomies: An Introduction to Classification Techniques. Sage Publications, Inc, Thousand Oaks (1994)
10. Wand, Y., Storey, V.C., Weber, R.: An ontological analysis of the relationship construct in conceptual modeling. ACM Trans. Database Syst. **24**, 494–528 (1999)
11. Strauss, A.L.: Qualitative Analysis for Social Scientists. Cambridge University Press, New York (1987)
12. Saya, S., Pee, L., Kankanhalli, A.: The impact of institutional influences on perceived technological characteristics and real options in cloud computing adoption. In: Thirty First International Conference on Information Systems, St. Louis, pp. 1–11 (2010)
13. Nunnally, J.C.: Psychometric Theory. McGraw-Hill, New York (1967)
14. Wakefield, R.L., Whitten, D.: Mobile computing: a user study on hedonic/utilitarian mobile device usage. Eur. J. Inf. Syst. **15**, 292 (2006)

15. van der Heijden, H.: Hedonic information systems. MIS Q. **28**, 695–704 (2004)
16. Dennis, A.R., Fuller, R.M., Valacich, J.S.: Media, tasks, and communication processes: a theory of media synchronicity. MIS Q. **32**, 575–600 (2008)
17. Zigurs, I.: Buckland BK A theory of task/technology fit and group support systems effectiveness. MIS Q. **22**, 313–334 (1998)
18. Sharma, R., Yetton, P.: The contingent effects of training, technical complexity, and task interdependence on successful information systems implementation. MIS Q. **31**, 219 (2007)
19. Gorla, N., Somers, T.M., Wong, B.: Organizational impact of system quality, information quality, and service quality. J. Strateg. Inf. Syst. **19**, 207–228 (2010)
20. Nelson, R.R., Todd, P.A., Wixom, B.H.: Antecedents of information and system quality: an empirical examination within the context of data warehousing. J. Manag. Inf. Syst. **21**, 199–235 (2005)
21. Wu, J.H., Wang, Y.M.: Measuring ERP success: the key-users' viewpoint of the ERP to produce a viable is in the organization. Comput. Hum. Behav. **23**, 1582–1596 (2007)
22. Abugabah, A., Sanzogni, L.: Re-conceptualizing information systems models: an experience from ERP systems environment. Int. J. Infono. **3**, 414–421 (2010)
23. Chung, B.Y., Skibniewski, M.J., Lucas Jr, H.C., et al.: Analyzing enterprise resource planning system implementation success factors in the engineering—construction industry. J. Comput. Civil Eng. **22**, 373–382 (2008)

Part II
Algorithms and Applications

Chapter 18
Joint Optimization About Pattern Synthesis of Circular Arrays Based on the Niche Genetic Algorithm

Yuan Fei, Zhao Ming, Huang Zhong Rui and Zhang Zhi

Abstract The high side lobe level is a serious problem for the circular array pattern. In order to solve this problem, a new method for the pattern synthesis of circular array is proposed in this paper. It makes the location of the array element and the coefficient as joint variables for its optimization model based on the niche genetic algorithm, which can overcome the shortcomings of premature and bad local searching capability existing in simple genetic algorithm. This approach can not only enhance the variables freedom degree but also accord with the academic global optimization. A measure is proposed to alleviate the dependence of convergence on the initial population, in which two reproduction operations of different types are alternatively used in generating chromosomes. The simulation results show the efficiency of this method.

Keywords Circular arrays · Niche · Genetic algorithm · Unite optimization · Side lobe level

Y. Fei (✉) · Z. Ming · Z. Zhi
Electronic Engineering Institute, Hefei, Anhui, China
e-mail: 18756073857@163.com

Z. Ming
e-mail: zhongrui66@hotmail.com

Z. Zhi
e-mail: lujiangliuhan@163.com

H. Z. Rui
Department of Information Engineering, Electronic Engineering Institute, Hefei, Anhui, China
e-mail: 564728486@qq.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_18,

18.1 Introduction

Array signal processing has been widely used in diverse applications including radar, communications, sonar, speech, intelligence analysis and medical image [1–3]. Circular arrays have more superior performance as opposed to the linear arrays. Circular arrays not only can provide two-dimensional azimuth and elevation angle information but also can move through the cycle incentives to control the beam direction flexibility. And also it can achieve beam scanning in all directions with the same aperture. Moreover, circular arrays' beam pattern is not sensitive to the frequency change [4, 5], which makes it more competitive in the broadband signal processing. But the higher side lobe level of circular arrays' pattern is their great shorting. In recent years, the use of sparse arrays to reduce the side lobe level is becoming a hotspot [6], but there is less literature for the circular array. [7–10] respectively, applied on the genetic algorithm, differential algorithms and their improved algorithm for designing a circular sparse array. [11, 12] is to make improvements of weighted coefficients to the formation of the beam pattern. The similarity of the above methods is that all of them are under certain circumstances and only make one of the sensor position and element weighting coefficient as optimization variables to optimize the maximum side lobe level. Array antenna pattern is jointly determined by the sensor position and element weighting coefficients, so the above methods are equivalent to a slice of the path from the domain of the function of the broken-line optimization, apparently lacking of a true sense of the global optimal solution.

This paper presents a new method for pattern synthesis of the circular array based on the niche genetic algorithm. The idea is that the joint optimization of sensor position and element weighting coefficient can improve the freedom of independent variables so as to achieve the defined region global optimization. It can get lower peak side lobe level. In addition, the basic genetic algorithm has been improved in order to avoid prematurity. It alternates the two genetic reproduction operations and extends the crossover-way, so it can effectively reduce the dependence of the algorithm's convergence on selection of the initial group [13]. In the end, it takes the angle difference between the chromosomal genes for the sensor position, which can not only reduce the optimization space but also improve the efficiency of the algorithm.

18.2 Pattern Function and Design Equations of the Circular Array Antenna

Consider a circular array [14], the radius $R = n \cdot \lambda$, n is a constant, λ is the wavelength. Let us consider N antenna array elements spaced on a circle according to Fig. 18.1, selecting the center of the circle as a reference point, d_i is the azimuth of i unit, and the desired beam direction is (ϕ_0, θ_0).

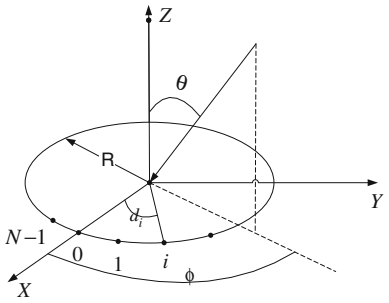


Fig. 18.1 Circular array antenna

Its pattern function can be expressed as:

$$F(\phi,\theta)=\sum_{i=0}^{N-1} w_i * \exp\left(j\frac{2\pi}{\lambda}R(\sin\theta\cos(\phi-d_i)-\sin\theta_0\cos(\phi_0-d_i))\right) \quad (18.1)$$

where: w_i is the i^{th} array element weights amplitude; ϕ starting at the X axis positive direction of the azimuth; θ starting at Z axis in the positive direction of the pitch angle. Here, the focus is to consider the changes in characteristics of the pattern with azimuth, that is to say, only consider the circular array where the plane direction of figure, and $\theta=\theta_0=90^\circ$ at this time, the array pattern function for the circular array can be written as:

$$F(\phi)=\sum_{i=0}^{N-1} w_i * \exp\left(j\frac{2\pi}{\lambda}R(\cos(\phi-d_i)-\cos(\phi_0-d_i))\right) \quad (18.2)$$

The peak side lobe level of the circular array pattern MSLL [15] is calculated as follows:

$$MSLL=\max_{\phi\in\Theta}\left\{\left|\frac{F(\phi)}{\max(F(\phi))}\right|\right\} \quad (18.3)$$

Let Θ denotes the side lobe beam pattern areas, so that the main lobe of the null-power beamwidth is $2\phi_0$, and

$\Theta=\{\phi|0^\circ\le\phi\le\phi-\phi_0\}\cup\{\phi|\phi-\phi_0\le\phi\le360^\circ\}$, $\max|F(\phi)|$ is for the entire airspace.

Now let us constrain that the interval of neighboring elements can not be smaller than a certain fixed length (or angle) d_c, namely:

$$\min\{d_i-d_j\}\ge d_c, 1\le j<i\le N \quad (18.4)$$

The azimuth of N elements known as:

$$d_N \leq 360^\circ - d_c \tag{18.5}$$

Let us split d_i into $x_i + (i-1)d_c$, then

$$\mathbf{d} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ \cdots \\ d_N \end{bmatrix} = \begin{bmatrix} 0 \\ x_2 + d_c \\ x_3 + 2d_c \\ \cdots \\ x_N + (N-1)d_c \end{bmatrix} = \begin{bmatrix} 0 \\ x_2 \\ x_3 \\ \cdots \\ x_N \end{bmatrix} + \begin{bmatrix} 0 \\ d_c \\ 2d_c \\ \cdots \\ (N-1)d_c \end{bmatrix} = \mathbf{X}_1 + \begin{bmatrix} 0 \\ d_c \\ 2d_c \\ \cdots \\ (N-1)d_c \end{bmatrix} \tag{18.6}$$

In order to satisfy the Eq. (18.4), the elements of the vector $\mathbf{X}_1$ must be ordered in descending, and $x_N \leq 360^\circ - Nd_c$, so its search space of elements will be reduced from $[0, 360^\circ]$ to $[0, 360^\circ - Nd_c]$, and for this it can improve the efficiency of the algorithm [16].

Because the array pattern of the circular array is normalized, so here we limit the array element weighting coefficients range is $[-1, 1]$. In order to improve the randomness of the initial chromosome and enhance the convenience of selection, crossover and mutation operations, the real and imaginary parts of the complex weighting coefficients is respectively expressed as $\mathbf{X}_2, \mathbf{X}_3$, as a vector of dimension equal to the number of array elements.

Now making the sensor position vector $\mathbf{X}_1$ and element weighting coefficient vector $\mathbf{X}_2, \mathbf{X}_3$ as the joint optimal variable of the real-coded chromosome $\mathbf{X}$, where $\mathbf{X} = \left[\mathbf{X}_1^T, \mathbf{X}_2^T, \mathbf{X}_3^T\right]^T$. From this we are available to achieve the following optimization model:

$$\begin{gathered} \min_{\mathbf{X}}(MSLL) \\ s.t.\ \mathbf{X} = \{(x_1, \cdots, x_{3N}) \mid 0^\circ \leq x_1 \leq \cdots \leq x_N \leq 360^\circ - Nd_c, -1 \leq x_{N+1}, \cdots, x_{3N} \leq 1\} \end{gathered} \tag{18.7}$$

18.3 Pattern Synthesis on Niche Genetic Algorithm

18.3.1 The Overview of the Niche Genetic Algorithm

The basic idea of the niche genetic algorithm [17] is: firstly, we should give the definition of the distance between groups of each chromosome. Then the real distance of each individual can be calculated. If this distance is less than the given

value of L, the individual which has the smaller fitness value will be punished, greatly reducing its fitness. After this treatment, the close proximity of the poor individual in subsequent competition will be at a disadvantage, which will gradually be eliminated. Only the superior one can be exist. That is to say the distance of L functions meets the target of the function to lower the chromosome of a larger penalty and will be gradually phased out in the subsequent competition.

Chromosome genes variable in this article is divided into two categories: Firstly, the position variable, and the other incentive variables. Between the two, their unit is not the same and the corresponding range is not in an order either, if direct them into the Euclidean distance formula:

$$\begin{aligned} F_{u,v} &= \alpha F_{1u,v} + \beta F_{2u,v} \\ &= \frac{\alpha}{N}\sqrt{\sum_{i=1}^{N}(\mathbf{X}_u^i - \mathbf{X}_v^i)^2} + \frac{\beta}{2N}\sqrt{\sum_{j=1}^{2N}(\mathbf{X}_u^j - \mathbf{X}_v^j)^2} \end{aligned} \tag{18.8}$$

Among them, $\mathbf{X}_u^i$ and $\mathbf{X}_u^j$, respectively, denotes the gene i and j in the chromosome u, $\mathbf{X}_v^i$ and $\mathbf{X}_v^j$, respectively, denotes the gene i and j in the chromosome v.

The values of $F_{1u,v}$ and $F_{2u,v}$ may vary greatly in the genetic manipulation. Compromising the case of a fitness function will no longer play a role, it is necessary to deal with the two fitness functions. The fitness functions after treatment as follows:

$$\begin{aligned} \widetilde{F}_{u,v} &= \alpha\widetilde{F}_{1u,v} + \beta\widetilde{F}_{2u,v} \\ &= \frac{\alpha}{N}\sqrt{\sum_{i=1}^{N}(\widetilde{\mathbf{X}}_u^i - \widetilde{\mathbf{X}}_v^i)^2} + \frac{\beta}{2N}\sqrt{\sum_{j=1}^{2N}(\widetilde{\mathbf{X}}_u^j - \widetilde{\mathbf{X}}_v^j)^2} \end{aligned} \tag{18.9}$$

where,

$$\begin{aligned} \widetilde{\mathbf{X}}_u^i &= \mathbf{X}_u^i / \max(\mathbf{X}_u^1, \ldots, \mathbf{X}_u^N, \mathbf{X}_v^1, \ldots, \mathbf{X}_v^N), i = 1, \ldots, N \\ \widetilde{\mathbf{X}}_v^i &= \mathbf{X}_v^i / \max(\mathbf{X}_u^1, \ldots, \mathbf{X}_u^N, \mathbf{X}_v^1, \ldots, \mathbf{X}_v^N), i = 1, \ldots, N \\ \widetilde{\mathbf{X}}_u^j &= \mathbf{X}_u^j / \max(\mathbf{X}_u^{N+1}, \ldots, \mathbf{X}_u^{3N}, \mathbf{X}_v^{N+1}, \ldots, \mathbf{X}_v^{3N}), j = N+1, \ldots, 3N \\ \widetilde{\mathbf{X}}_v^j &= \mathbf{X}_v^j / \max(\mathbf{X}_u^{N+1}, \ldots, \mathbf{X}_u^{3N}, \mathbf{X}_v^{N+1}, \ldots, \mathbf{X}_v^{3N}), j = N+1, \ldots, 3N \end{aligned}$$

After normalization, we can guarantee that $\widetilde{F}_{1u,v}$ and $\widetilde{F}_{2u,v}$ on an order of magnitude and can be comparable. α and β is the weight of the type (18.8) and (18.9) respectively. By adjusting α and β, we can choose the main way to optimize the array, that is, to optimize the sensor position-based or array element excitation optimization-based or equal emphasis.

18.3.2 Pattern Synthesis Based on Niche Genetic Algorithm

Equations (18.2), (18.3) show that the objective function in (18.7) is a highly nonlinear NP-hard problem, which can be optimized via niche genetic algorithm. Improvements have been dealed with the basic niche genetic algorithm in order to avoid the algorithm slow convergence and bad stability. Their specific implementation steps are as follows (Fig. 18.2).

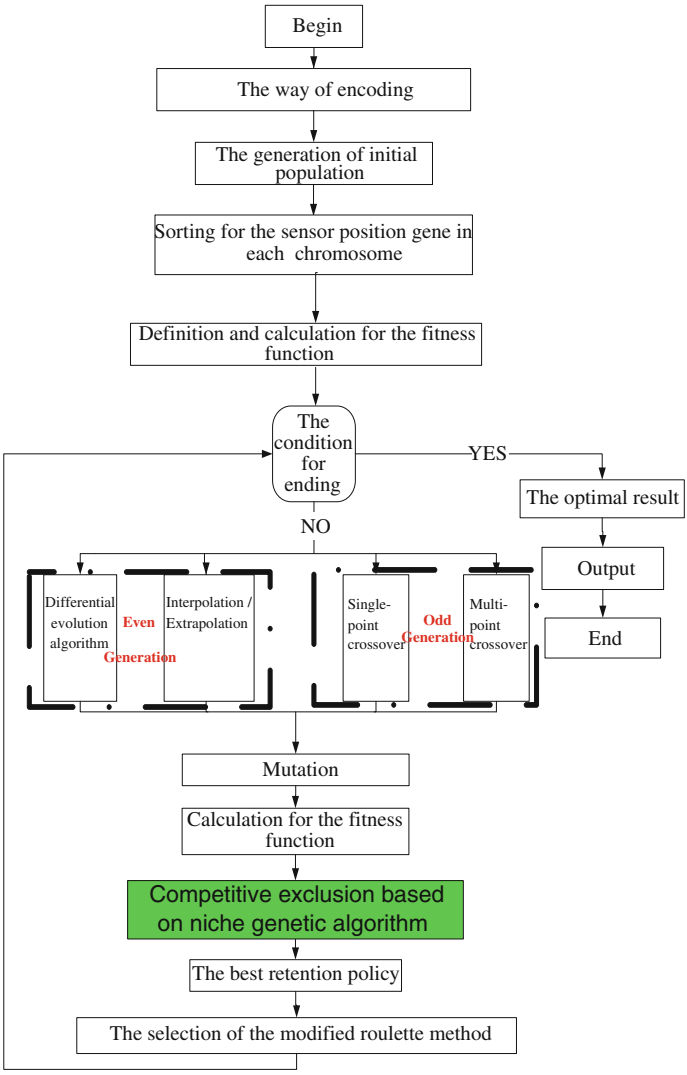


Fig. 18.2 The flow chart of the niche genetic algorithm

18.4 Simulation Results

To demonstrate the effectiveness of our method (niche genetic algorithm) in pattern synthesis of the circular array antenna, we choose two other methods the differential evolution algorithm (DE) and the genetic algorithm (GA), as comparison, and evaluate the performance of each method by investigating an important parameter in pattern synthesis, i.e., the peak side lobe level outside the main lobe while maintaining the zero power width is constant.

In the example, there is a circular array of radius $R = 2\lambda$, and its desired beam direction is $(200°, 90°)$, $\theta = \theta_0 = 90°$. The number of array elements is $N = 16$ and neighboring elements of the minimum two-array azimuth difference is not less than $d_c = 14.3641°$ (the distance of the half-wavelength). The number of individuals in the arrangement is $M = 100$, and the maximum evolution generation $G = 200$, the variable dimension is $3N = 48$.The differential amplitude control coefficient $F = 1$ and interpolation/extrapolation coefficient $C = 0.25$, $\alpha = 0.05$, $v\beta = 0.95$, the mutation probability is taken as 0.015, the width of the main lobe of the null power point $2\phi_0 = 24°$, the discrete points of side lobe is 138, the discrimination interval is about 2.5°.

Figure 18.3 shows the pattern of the uniform circular array whose number of elements $N = 16$, the maximum side lobe level is -4.4336 dB.

Figure 18.4 shows the diagram for the optimized circular array direction, the figure shows that the peak side lobe level is -13.1 dB. In comparison with [9] in which the peak side lobe level is -11.5922 dB and [10] in which the peak side lobe level is -11.3468 dB, it is lower 1.5078 and 1.7532 dB, respectively. When compared with uniform circular array it has been obtained nearly 9 dB reduction. So this experiment indicates the validity of the presented optimization algorithm.

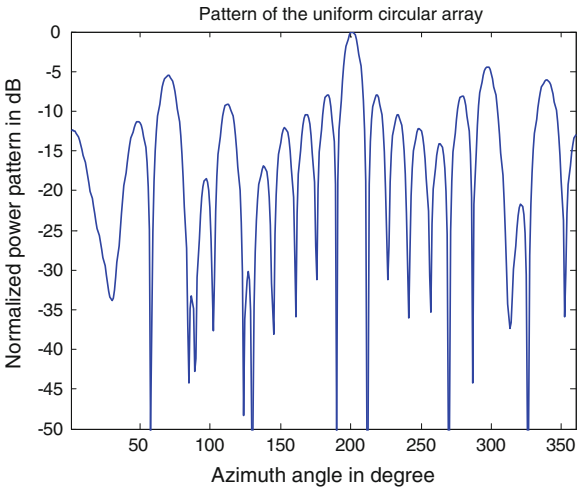


Fig. 18.3 The pattern of the uniform circular array

Fig. 18.4 The pattern of the sparse circular array

Figure 18.5 shows the position distribution of the array element in the uniform circular array and the sparse circular array. The dots represent element positions of uniform circular, while the triangles represent that of the sparse array.

Figure 18.6 shows the convergence curve of the fitness function based on niche genetic algorithm proposed in this paper. And the simulation times is 10, where the solid line represents the average of the fitness function among the 10 simulations, the dotted line represents each simulation's curve. From the chart it can be seen that, the application of this algorithm can not only enhance optimization ability of the fitness function but also improve the convergence speed of the objective function. So that it can obtain a lower peak side lobe level. The superiority of the proposed algorithm and the correctness of the selected fitness function can be demonstrated through the simulations.

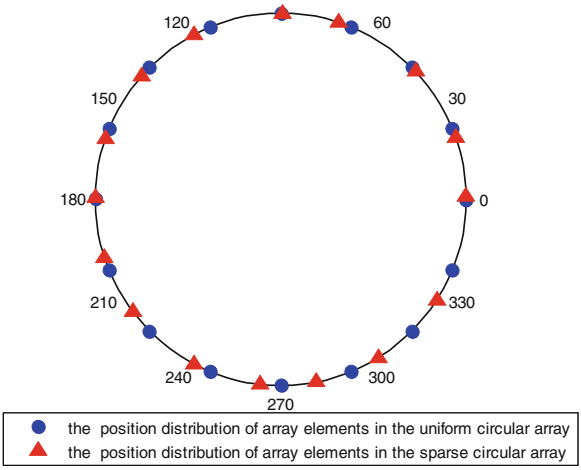


Fig. 18.5 The position distribution of the array elements

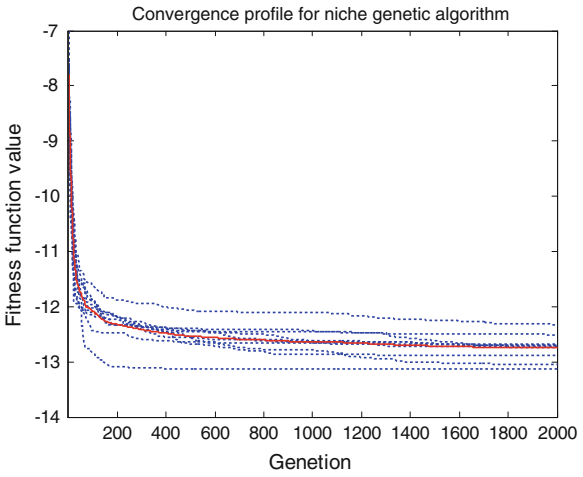


Fig. 18.6 Convergence profile for niche genetic algorithm

18.5 Conclusion

As to the peak side lobe level of the circular array pattern, researchers have made the sensor position and the weighted coefficients as joint optimization variables to minimize the peak side lobe level for the fitness function based on niche genetic algorithm optimization. It can overcome the lack of convergence to local optimal of the traditional algorithm. Researchers take two kinds of crossover-way to overcome the convergence dependent on the selected initial population in the presented method. In addition, taking the angle difference between the chromosomal genes for the sensor position, which can remove redundant optimization space, and improve the efficiency of algorithm optimization? Finally, the simulation experiment illustrates the effectiveness of the method. Much reduction of the side lobe can be obtained in comparison with the reference methods.

References

1. Mohamad, G.: Wideband smart antenna theory using rectangular array structures. IEEE Trans. Signal Process. **50**(9), 2143–2151 (2002)
2. Wu, M.Q.: The development of digital array radar. China Acad. Electron. Inf. Technol. **1**(1), 11–16 (2006)
3. Zheng, H.S., Xu, Z.D., Jin, M.X.: A novel receiver architecture for DBF antenna array. J. Electron. Sci. Technol. China **5**(1), 33–37 (2007)
4. Steyskal, H.: Digital beamforming aspects of wideband circular arrays[C]. In: IEEE Aerospace Conference. IEEE, Piscataway, pp. 1–6 (2008)
5. Miao, P.C., Yin, Q.Y., Zhang, J.G.: A wideband beamforming method based on directional uniform circular arrays. Sci. Sinica (Informationis) **41**(2), 246–256 (2011)
6. Chen, K.S., Yun, X.H., He, Z.S.: Synthesis of sparse planar arrays using modified real genetic algorithm. IEEE Trans. Antennas Propag. **55**(4), 1067–1073 (2007)

7. Wang, L.Y., Bao, Z.Y., Chen, K.S.: Study on the sparse of the circular array. China Radar **2**, 1–4 (2008)
8. Dhanesh, G.K., Mohamed, H., Anders, R.: Synthesis of uniform amplitude unequally spaced antenna arrays using the differential evolution algorithm. IEEE Trans. Antennas Propag. **51**(9), 2210–2217 (2003)
9. Bao, Z.Y., Chen, K.S., He, Z.S., Han, C.L.: Sparse circular arrays method based on modified DE algorithm. Syst. Eng. Electron. **31**(3), 497–499 (2009)
10. Bao, Z.Y., Chen, K.S., He, Z.S., Han, C.L.: A sparse circular arrays method based on modified genetic algorithm [J]. Appl. Electron. Tech. **10**, 110–112 (2008)
11. Zhou, Y.P., Zhang, Z.Q.: An approach of beamform ing based on improved genetic algorithm. Comput Simul **27**(8), 208–211 (2010)
12. Jin, J., Wangi, H.L., Liu, M.: Partially adaptive beam-forming methods based on genetic algorithms [J]. J. Commun. **27**(12), 92–97 (2006)
13. Li, D.F., Gong, Z.L.: Application of genetic algorithms in the pattern synthesis of ultra-low sidelobe linear array antenna. Acta Electronica Sinica **31**(1), 82–84 (2003)
14. Liu, X.X., Zhang, L.T., Wu, S.L., Mao, R.K.: Pattern synthesis of circular arrays based on directional elements. Acta Electronica Sinica **32**(4), 701–704 (2004)
15. Ma, Y.H.: Null steering using genetic algorithms by controlling only the current phases. J. Microwares **17**(2), 41–46 (2001)
16. He, X.H., Wu, Z.P., Wu, S.J.: Pattern synthesis with desired magnitude response for arbitrary arrays. Acta Electronica Sinica **38**(10), 2292–2296 (2010)
17. Yin, Y.T., Yang, S.Y.: Application of the niche genetic algorithm in antenna array pattern synthesis [J]. Appl. Sci. Technol. **34**(12), 12–16 (2007)

Chapter 19
The Application of Wavelet Analysis and BP Neural Network for the Early Diagnosis of Coronary Heart Disease

Shengping Liu, Guanlan Chen and Guoming Chen

Abstract Based on the relationship between coronary artery blockages and heart sound signals, a novel processing method on heart sound signal of early diagnosis of coronary heart disease was proposed. With the wavelet analysis, the heart sound signal was decomposed and reconstructed, and the coefficient of each layer was extracted. The information content of the first and the second heart sound signal (S1, S2) was calculated from Shannon entropy. The time threshold was applied to obtain the interval between S1 and S2. All the characteristic values were combined into a matrix containing nine elements, which was regarded as the input of a BP neural network for the identification of heart sound signal. The results show that the proposed algorithm is highly accurate for the early diagnosis of coronary heart disease. The recognition rate of the simple aortic regurgitation, the aortic regurgitation, the mitral valve stenosis and mitral valve insufficiency were 73.33, 80.00, 86.67 and 93.33 % respectively. It provides a non-invasive early diagnosis method of coronary heart disease.

Keywords Coronary heart disease · Heart sound signals · Wavelet analysis · BP neural network

19.1 Introduction

Coronary heart disease (CHD) is most commonly equated to the atherosclerotic of coronary artery, but it could arise from other factors such as the coronary vasospasm, which is caused by spasm of the blood vessels of heart. Heart sounds can be heard by ears or a stethoscope on the chest wall during cardiac systolic and

S. Liu (✉) · G. Chen · G. Chen
Department of Biomedical Engineering, Chongqing University of Technology, Chongqing, China
e-mail: lsp0717@sohu.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_19,

diastolic period. Also, it can be conveniently recorded with electronic instruments. Normally, the heart spurting speed and other factors can produce physiological murmurs. The variation of heart sounds and noise appearance are the early signs of organic heart disease, which can be detected through the heart auscultation before the other signs and symptoms appearance. Analysis of heart sound signal (HSS) has a great significance for the early diagnosis of cardiovascular diseases [1]. Its accuracy and reliability will directly affect the evaluation of clinical diagnosis and prognosis effect.

In order to understand the characteristics of heart sounds comprehensively, some researchers utilized the artificial neural network to analyze heart sounds. The phonocardiogram was directly input into the BP neural network (BPNN) for analyzing. Since the influence of randomness of signal variation, it is hard to accurately classify the HSSs [2]. For overcoming the disadvantages, Kay et al. [3, 4] proposed the auto-regressive model, and the auto-regressive and moving average model, respectively. The sound transmission system has a strong Time-Varying line; therefore, the HSSs on the same stage processing cepstrum could cause data confusion in theory. If cepstrum were processed as an input of a neural network, the recognition results could be inaccurate [5]. The short-time Fourier transform is extensively applied for time–frequency analysis. But the type and length of analysis window critically influence on the analysis results. Wavelet Transform is prospective applications in the fields of biomedical signal processing [6, 7]. In order to ensure the effectiveness of information processing, accuracy and operation efficiency, the Discrete Wavelet Transform (DWT) and normalized Shannon information content were adopted to analyze the HSSs in this paper. By comparing the normal and abnormal HSSs, the eigenvalues associated with the coronary heart disease were extracted, and then input into the BPNN to identify and classify for the detection and adjuvant treatment of organic heart diseases.

19.2 The Heart Sound Signals Processing and Feature Extraction

Figure 19.1 shows the processing flow chart of pretreatment and feature extraction of HSSs.

Preprocessing aims at the removal of high-frequency components of the heart sound signals which result from the environment impact, and the selection of concentration energy of the first heart sound (S1) and the second heart sound (S2). By compulsory denoising, the high frequency coefficient was set to 0 in the

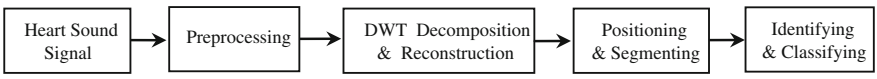


Fig. 19.1 Preprocessing and feature extraction of HSSs

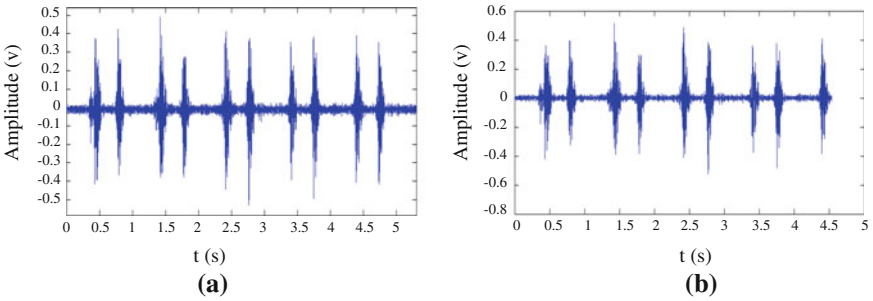


Fig. 19.2 Original signal of heart sound and denoise effect. **a** Original signals, **b** Denoised signals

wavelet decomposition, that is, all the high frequency portions were removed. This method is relatively simple, and the denoised signals are relatively smooth. The results are shown in Fig. 19.2. It indicates that the preprocessed signal still remained the main component of the original signals while the high-frequency murmurs were effectively eliminated.

The time–frequency diagram of Time-Varying (TV) HSSs can easily analyze the location of main frequencies. A 3D spectrum, which reveals the spectral characteristics of TV HSSs, is shown in Fig. 19.3. The time–frequency diagram in Fig. 19.3 figures that the main signal frequency locates in the band of 0–800 Hz. It indicates that the signals distribute by themselves energy. Frequency localization properties can apply to analyzing the TV spectra of signals. It is suitable for analyzing signals in details because of its non-sensitive to noises.

According to the characteristics of the HSSs, it needs to ensure the effectiveness of information processing, accuracy and operation efficiency, and facilitate to access the reconstructed results. The DWT algorithm was adopted to produce intensity envelopes of approximations and details of original phono- cardiographic signals [8]. The key to the discrete wavelet decomposition and reconstruction is the selection of wavelet basis functions and decomposition levels. Figure 19.4 is the denoised effect under different decomposition levels. From Fig. 19.4, the level 3 still contains many high frequency components; the level 4 and 5 are the ideal level in this study; level 6 and 7 can perfectly remove the noise, but some effective signals were removed. Figure 19.5 shows the denoising effect of different wavelet basis functions under the decomposition level 4. It indicates that the daubechies 6(db6) can not only denoise effectively, but also retain the effective portion of signals. So the db6 and level 4 were chosen to denoise the HSSs.

The signals (S) were decomposed into four layers (d1, d2, d3, d4) using db6 based on the Eq. (19.1):

$$S = d_1 + d_2 + d_3 + d_4 + a_j \tag{19.1}$$

where $a_j = \frac{\sum_{i=1}^{n} |d_{ji}|}{n}$ is the signal average in j level, d_{ji} is signal vectors in j level, n is the signal dimension. A high frequency feature was extracted. A single

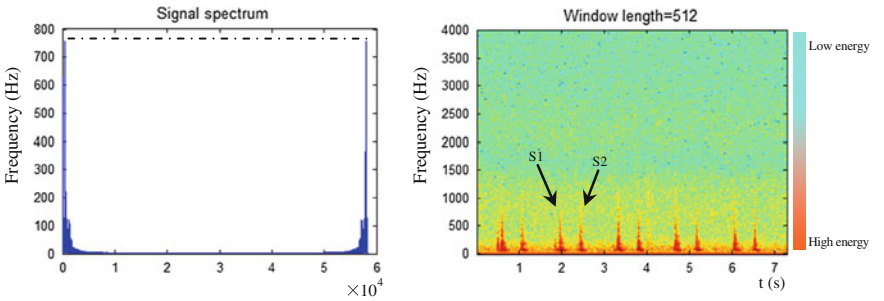


Fig. 19.3 Signal spectrum of normal heart sound

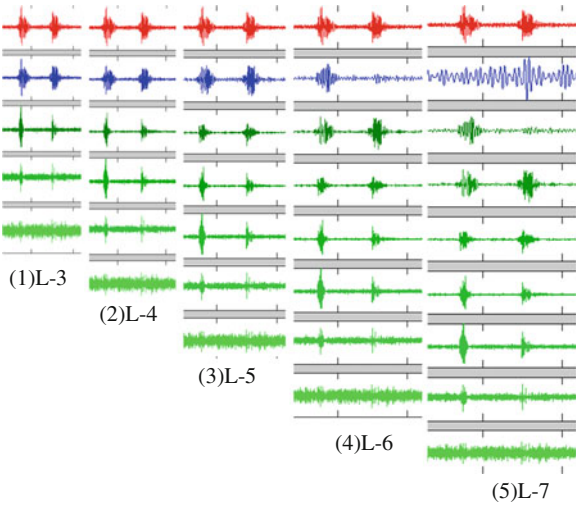


Fig. 19.4 Denoising effect under different decomposition levels

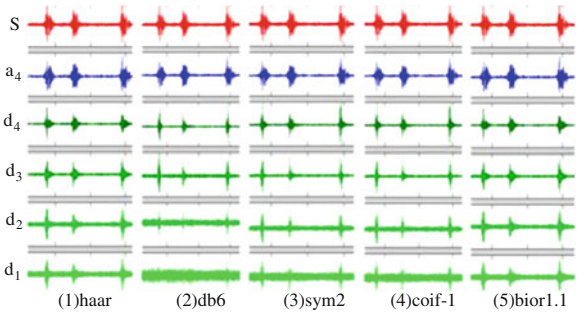


Fig. 19.5 Denoising effect of different wavelet basis functions

numerical value was used to express frequency feature value of signals in each layer.

19.3 Heart Sound Signals Classification Based on BP Neural Network

The characteristics of HSSs were identified based on the principle of pattern matching. The guide to the best match of the reference pattern is the recognition result for the HSSs. Figure 19.6 shows the identification process. The model classification of the characteristic signal of HSSs based on the BPNN

BPNN includes three steps [9, 10]: building, training and classification. The input of the BPNN is a 9-dimension vector. The HSSs are classified as four categories, which correspond to the simple aortic regurgitation, the mitral (valve) stenosis, the aortic regurgitation and the mitral valve insufficiency, respectively. Therefore, the number of input nodes of the BPNN is 9 and the number of output nodes of the BPNN is 4. The number of 3–12 hidden layer nodes was tested, respectively. The optimal node number was finally determined according to the result of training and error analysis. The number of hidden node was fixed as 10 in this paper.

The activation function of the BPNN adopts sigmoid function as the Eq. (19.2):

$$f(x) = \frac{1}{1 + e^{-x}} \tag{19.2}$$

The error of the p demo is calculated with the Eq. (19.3):

$$E_p = \frac{1}{2}\sum_{j=1}^{m} (y_{pj} - o_{pj})^2 \tag{19.3}$$

where, y_p is the actual output, o_p is the expected output. The error of the total sample is $E = \sum_p E_p$.

The feature matrix [da_1, da_2, da_3, da_4, aa_4, E_{S1}, E_{S2}, T_i, T_j] was input into the BPNN. Herein, da_i is the high frequency component in the i layer; E_{S1}, E_{S2} is the information content of S1 and S2, which is calculated with the Eq. (19.4); T_i is the interval between S1 and S2; T_j is the interval between S2 and next S1. T_i, T_j can be measured from the segmentation figure using Shannon entropy and time threshold.

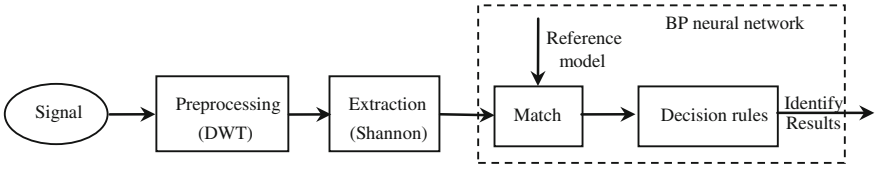


Fig. 19.6 Identification process of heart sounds

$$E_s = -\frac{1}{N}\sum_{i=1}^{n} hs^2(i) \cdot \log(hs^2(i)) \tag{19.4}$$

where, *hs* is one of the signal segment, *N* is the length of each segment.

The desired output value was initialized according to the identification category of the HSS. For example, when the identification category is 1, the desired output vector is [1 0 0 0]. It means the input signals are the first category signal, namely the simple aortic regurgitation.

19.4 Results and Discussions

The HSSs were compulsorily denoised and decomposed into 4 layers using the db6 wavelet. The frequency domains were divided into 4 bands, namely 0–138, 138–275, 275–551 and 551–1102 Hz. The coefficient of each layer is a characteristic value. Decomposition and reconstruction of normal and abnormal HSSs are shown in Fig. 19.7. The energy of normal HSSs concentrates in the range of 0–138 Hz as Fig. 19.7a. Thus, it is useful to determine the murmurs through the different energy of the signals. There have much noise locating in the high frequency band (138–1102 Hz) as Fig. 19.7b. Generally, it could arise from the outside interference or disorder of heart itself. According to the clinic research results, HSSs of CHD patients contain more high frequency components, especially in the band of 551–1102 Hz as Fig. 19.7b. The processing method proposed benefits to determine some early signs of CHD.

The purpose of the heart sound segmentation is to ascertain the position of main signal components, including the first heart sound(S1), the systole, the second heart sound(S2) and the diastolic, which are the basis of feature extraction and pattern recognition. Also, they are the prerequisite and foundation of non-invasive diagnosis of coronary heart disease [11]. The Shannon entropy and time gate were utilized to extract the HSS envelope. Figure 19.8 plots the segmentation results

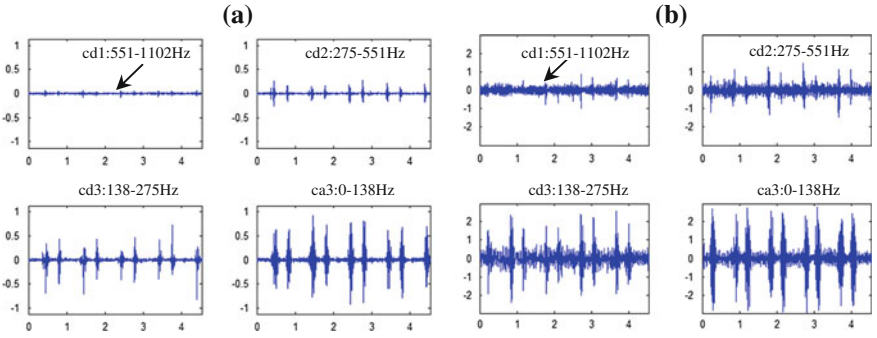


Fig. 19.7 Decomposition and reconstruction of normal and abnormal HSS

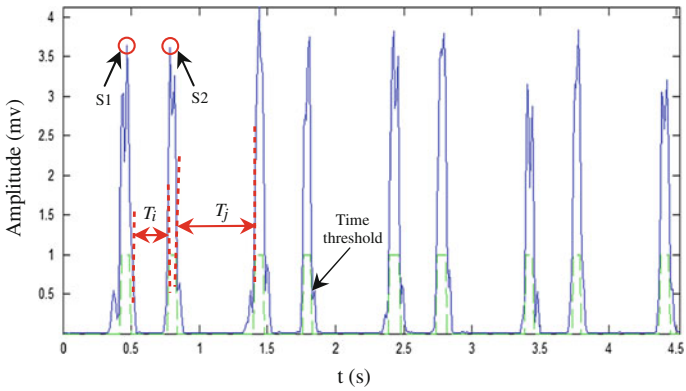


Fig. 19.8 Segmentation results using Shannon entropy and time threshold

using Shannon entropy and time threshold. From Fig. 19.8, the combination of the Shannon entropy and time gate can accurately segment the range of heart sounds and quickly locate the position of S1, S2. Comparing the healthy person, the CHD patients have higher energy in HSS, and the interval time has an obvious variation. The Shannon entropy was regarded as the energy eigenvalues of HSSs. Also, the interval T_i and T_j are regarded as eigenvalues of HSSs.

Sixty heart sound data were derived from the West China courseware resources. Among them, 45 heart sound data were used as training set. During training, the learning rate is 0.1; the error of training terminated is 0.01. In order to verify the BPNN trained, other 15 heart sound data of the simple aortic regurgitation, the valve stenosis, the aortic regurgitation and the mitral valve insufficiency were input into the BPNN, respectively. The testing results show that 11 cases were identified as the simple aortic regurgitation, 12 cases were identified as the mitral valve stenosis, 13 cases were identified as the aortic regurgitation and 14 cases were identified as the mitral valve insufficiency. Table 19.1 shows the recognition rate of the four categories of HSSs using the BPNN built in this paper. To the method of Self-Organizing Feature Map with Support Vector Machine for the normal and abnormal HSSs identification, its accurate rate is just approximately 85.1 % [8].

Table 19.1 The recognition rate of heart sound signals

Category of heart sound signal	Simple aortic regurgitation (%)	Mitral (valve) stenosis (%)	Aortic regurgitation (%)	Mitral valve insufficiency (%)
Recognition rate	73.33	80.00	86.67	93.33

19.5 Conclusion

The DWT db6 wavelet decomposition and reconstruction was used to analyze the HSSs. It effectively eliminates the high-frequency noise. It is simple to extract the HSS envelope by combining the Shannon entropy and time threshold, which is effective to segment and locate the HSSs. A BPNN with 9-10-4 configuration was built to recognize and classify the four categories HSSs. The recognition rate of the simple aortic regurgitation, the aortic regurgitation, the mitral valve stenosis and mitral valve insufficiency were 73.33, 80.00, 86.67 and 93.33 % respectively.

The processing method proposed is simple and accurate to analyze the HSSs. However, there are other early symptoms of coronary heart disease. For improving the recognition rate and reliability of early diagnosis of coronary heart disease, it's important to integrate other physiological information, such as pulse wave velocity, etc., into the novel processing method proposed in this paper.

References

1. Yang, Y., Xu, H.: The strategy of early diagnosis on coronary heart disease. Chin. J. Front. Med. Sci. **2**(1), 33–35 (2010)
2. Barschldorff, D., Ester, S., Dorsel, T., et al.: A new phonographic technique for congenital and acquired heart disease using neural networks. Biomed. Technik. **35**(11), 271–279 (1990)
3. Akay, M.: Time-frequency analysis of the turbulent sounds caused by femoral artery stenosis in dogs using wavelet transform. Engineering in Medicine and Biology Society. In: 14th Annual International Conference of IEEE (1992)
4. Akay, M.: Neural networks for the diagnosis of coronary artery disease. Neural Netw. **2**, 419–424 (1992)
5. Durand, L.-G., Guo, Z., Sabbah, H.N., Stein, P.D.: Comparison of spectral techniques for computer-assisted classification of spectra of heart sounds in patients with porcine bioprosthetic valves. Med. Biol. Eng. Comput. **31**, 229–236 (1993)
6. Messer, S.R., Agzarian, J., Abbott, D.: Optimal wavelet denoise for phonocardiograms. Microelectron. J. **32**, 931–941 (2001)
7. Khadra, L., Matalgah, M., El-Asir, B., Mawagdeh, S.: The wavelet transform and its applications to phonocardiogram signal analysis. Med. Inform. **16**, 271–277 (1991)
8. Liang, H., Sakari, L.: A heart sound segmentation Algorithm using wavelet decomposition and reconstruction. In: Proceeding of 19th International Conference IEEE/EMBS (1997)
9. Chen, T., Xing, S., Guo, P.: The research of non-invasive method of coronary heart disease based on neural network and heart sound signals. In: ICIECS (2009)
10. Ibrahim, T., Ahmet, A., Erdogan, I.: An intelligent system for diagnosis of the heart valve diseases with wavelet packet neural networks. Comput. Biol. Med. **33**(4), 319–331 (2003)
11. Akay, Y.M.: Automated noninvasive detection of coronary artery disease using wavelet-based neural networks. Engineering in Medicine and Biology Society. Engineering Advances: New Opportunities for Biomedical Engineers. In: Proceedings of the 16th Annual International Conference of the IEEE (1994)

Chapter 20
Using More Initial Centers for the Seeding-Based Semi-Supervised K-Harmonic Means Clustering

Lei Gu

Abstract In the initialization of the traditional semi-supervised k-means, the mean of some labeled data belonging to one same class was regarded as one initial center and the number of the initial centers is equal to the number of clusters. However, this initialization method using a small amount of labeled data also called seeds which are not appropriate for the semi-supervised k-harmonic means clustering insensitive to the initial centers. In this paper, a novel semi-supervised k-harmonic means clustering is proposed. Some seeds with one same class are divided into several groups and the mean of all data is viewed as one initial center in every group. Therefore, the number of the initial centers is more than the number of clusters in the new method. To investigate the effectiveness of the approach, several experiments are done on three datasets. Experimental results show that the presented method can improve the clustering performance compared to other traditional semi-supervised clustering algorithms.

Keywords Semi-supervised clustering · K-harmonic means clustering · K-means clustering · Seeds

20.1 Introduction

Data clustering, known as one pattern recognition technique, has been used in a wide variety of fields. Clustering is a division of data into homogeneous groups called clusters. Each group consists of objects having the larger similarity between

L. Gu (✉)
The Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai, China
e-mail: gulei@njupt.edu.cn

L. Gu
School of Computer Science and Technology, Nanjing University of Posts and Telecommunication, Nanjing, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_20,

themselves than objects of other groups [1, 2]. Different measure and criteria of similarity lead to the various clustering algorithms such as the c-means [3], fuzzy c-means [4], fuzzy c-mediods [5], partitioning round mediods [6], neural gas [7] and hierarchical algorithms.

The traditional unsupervised k-harmonic means clustering method, proposed by Zhang et al. [8, 9], is similar to the k-means clustering and minimize the harmonic average from all points in the data set to all cluster centers [10]. Although the k-harmonic means clustering is insensitive to the initial centers, it often obtains the local optimal solutions. So some improved k-harmonic means clustering is proposed such as the k-harmonic means with the hard and soft membership function [11].

A small amount of labeled data is allowed to be applied to aiding the clustering of unlabeled data in semi-supervised clustering unlike the unsupervised clustering, and so a significant increase in the clustering performance can be obtained by the semi-supervised clustering [12]. The popular semi-supervised clustering methods are composed of two categories called the similarity-based and search-based approaches respectively [13]. It is noticeable that semi-supervised k-means clustering by seeding had been proposed recently [12]. This presented method introduced the clustering method viewed as the semi-supervised variants of k-means called Seed-KMeans (SeedKM). SeedKM can apply some labeled data called seeds to the initialization of the k-means. Therefore, like the k-means, the SeedKM is very sensitive to the initial centers.

In the SeedKM, the average of all seeds with one same class is regarded as one initial center and the number of all initial centers is equal to the number of clusters. However, this semi-supervised clustering initialization cannot use for the k-harmonic means clustering because the k-harmonic means clustering is insensitive to the initial centers. Although one semi-supervised k-harmonic means clustering approach had been presented, but this method only applies some labeled data to affecting the semi-supervised clustering process and its initialization is the same with the unsupervised k-harmonic means clustering using the number of initial centers equalling the number of clusters [14]. Therefore, in this paper, a novel semi-supervised k-harmonic mean clustering (NSeedKHM) is proposed. One feature of the new presented NSeedKHM is that multiple initial centers more than the number of clusters is used for the NSeedKHM. In this paper, another semi-supervised variant of k-harmonic means clustering by seeding called the Seed-KHM is given in the experiments. The SeedKHM applied the same initialization algorithm with the SeedKM. In order to assess the performance of the proposed NSeedKHM method, some experiments are done on one artificial dataset and two real datasets. Experimental results show that the NSeedKHM can obtain the better clustering performance compared with the SeedKHM and SeedKM.

The remainder of this paper is organized as follows. Section 20.2 reports the traditional unsupervised k-harmonic means clustering algorithm. In Sect. 20.3, the novel semi-supervised k-harmonic means clustering method NSeedKHM is formulated. Experimental results are shown in Sects. 20.4, and 20.5 gives our conclusions.

20.2 The Unsupervised K-Harmonic Means Clustering

Compared to the k-means clustering, the feature of the k-harmonic means clustering is insensitive to the initialization of the centers [10].

Step 1. Assume that the whole dataset $X = \{x_1, x_2, \cdots, x_N\}$ has N unlabeled samples in the d-dimensional space R^d, X can be divided into K different clusters, and c_j $(j = 1, 2, \cdots, K)$ represents the center of each cluster. The procedure of the unsupervised k-harmonic means clustering as follows [10]:Acquire K initial centers $c_j\,(j = 1, 2, \cdots, K)$ for the k-harmonic means clustering, and $M^* = 0$.

Step 2. According to the following function $M(X)$, compute its value M. In Eq. (20.1), q is a parameter and let $q \geq 2$. In Sect. 20.4, we let $q = 3$.

$$M(X) = \sum_{i=1}^{N} \frac{K}{\sum_{j=1}^{K} \frac{1}{\left\|x_i - c_j\right\|^q}} \tag{20.1}$$

Step 3. Based on the following equation, get each element T_{ij}, $(i = 1, 2, \cdots, N,\ j = 1, 2, \cdots, K)$ of the matrix T.

$$T_{ij} = \frac{\left\|x_i - c_j\right\|^{-q-2}}{\sum_{j=1}^{K} \left\|x_i - c_j\right\|^{-q-2}} \tag{20.2}$$

Step 4. Obtain the weight L_i of each data x_i according to the following Eq. (20.3).

$$L_i = \frac{\sum_{j=1}^{K} \left\|x_i - c_j\right\|^{-q-2}}{\left(\sum_{j=1}^{K} \left\|x_i - c_j\right\|^{-q}\right)^2} \tag{20.3}$$

Step 5. Update each cluster center c_k using the following Eq. (20.4).

$$c_j = \frac{\sum_{i=1}^{N} T_{ij} L_i x_i}{\sum_{i=1}^{N} T_{ij} L_i} \tag{20.4}$$

Step 6. If $|M^* = M| > \varepsilon$, then let $M^* = M$ and go to Step 2; otherwise go to Step 7.
Step 7. For each data x_i, assign it to cluster j^* by the following Eq. (20.5) and end

$$j^* = \arg \max_{j=1,2,\cdots,K} U_{ij} \tag{20.5}$$

20.3 The Proposed NSeedKHM

In the last Section, the traditional unsupervised k-harmonic means clustering is introduced. Its semi-supervised variant called the NSeedKHM is demonstrated in this Section. In the NSeedKHM, a small amount of labeled data also called seeds is

allowed to be applied to aiding and biasing the clustering of unlabeled data unlike the unsupervised k-harmonic means clustering.

Firstly, the generation of seeds is given. Given the number of clusters M and a nonempty set $X = \{x_1, x_2, \cdots, x_N\}$ of all unlabeled data in the d-dimensional space R^d, the clustering algorithms can partition X into K clusters. Let W, called the seed set, be the subset of X and for each x_t $(x_t \in W)$, the label be given by means of supervision [12]. We assume that W can be divided into K groups on the basis of data labels and each subgroup should be no empty set for the implementation of one-class support vector machine. Therefore, we can obtain a K partitioning $\{W_1, W_2, \cdots W_K\}$ of the seed set W.

Secondly, the NSeedKHM semi-supervised clustering method is outlined as follows:

Step 1. After each W_p $(p = 1, 2, \cdots, K)$ is partitioned into E subgroups randomly, a $K \cdot E$ partitioning $\{W_{11}, W_{12}, \cdots W_{1E}, W_{21}, W_{22}, \cdots W_{2E}, \cdots W_{K1}, W_{K2}, \cdots W_{KE}\}$ of the seed set W can be obtained. ($E \geq 1, E$ is one integer)

Step 2. Set $S = \left\{X^{du} | d = 1, 2, \cdots, K; u = 1, 2, \cdots, E\right\}$ where $\forall X^{du} = W_{du}$, $d = 1, 2, \cdots, K$ and $u = 1, 2, \cdots, E$.

Step 3. For each subset X^{du} of S, compute C_{du} using the following equation $(d = 1, 2, \cdots, K\ ,\ \ u = 1, 2, \cdots, E)$:

$$C_{du} = \frac{1}{G} \sum_{r=1}^{G} x_r \tag{20.6}$$

where $\forall x_r \in X^{du}$ and G is the number of all data belonging to X^{du}.

Step 4. Use $C_{du}(d = 1, 2, \cdots, K\ ,\ \ u = 1, 2, \cdots, E)$ for the initial centers and run the traditional unsupervised k-harmonic means clustering based on the KE initial centers. Notice that the Eq. (20.5) is changed into the following Eq. (20.7) here.

$$j^* = \left\lceil \frac{\left(\arg \max_{j=1,2,\cdots,KE} U_{ij} \right)}{E} \right\rceil \tag{20.7}$$

Step 5. End the NSeedKHM.

Finally, the following points need to be explained about the NSeedKHM:

(a) When $E = 1$ the NSeedKHM becomes the SeedKHM. The SeedKHM is similar to the SeedKM and applies the seeds to the initialization. In the SeedKHM, the number of the initial centers is equal to the number of clusters.

(b) When $E \geq 2$, multiple initial centers more than the number of clusters is used for the NSeedKHM. In Sect. 20.4, let $E = 2$ in all experiments.

(c) In Step 1 of the NSeedKHM, each W_{du} $(d = 1, 2, \cdots, K\ ,\ u = 1, 2, \cdots, E)$ should be no empty set.

20.4 Experimental Results

To demonstrate the effectiveness of the NSeedKHM, we compared it with two semi-supervised clustering methods, such as SeedKM and SeedKHM, on one artifical dataset and two UCI real datasets [15], referred to as DUNN, Ionosphere and Haberman respectively. The DUNN dataset shown in Fig. 20.1 contains 90 cases with 2-dimensional feature from two classes. The Ionosphere dataset and the Haberman dataset collect 351 34-dimensional cases from two classes and 306 3-dimensional cases belonging to two classes respectively. All experiments were done by Matlab on WindowsXP operating system.

For the SeedKM, SeedKHM and NSeedKHM, we randomly generated $P\,\%(P = 4, 6, 8, 10, 12, 14, 16, 18, 20)$ of the dataset as seeds on each UCI dataset, and we randomly generated $P\,\%$ $(P = 10, 12, 14, 16, 18, 20)$ of the dataset as seeds on the DUNN dataset. Since true labels are known, clustering accuracies $Q\,\%$ on unlabeled data, the remaining $(100 - P)\,\%$ of the dataset could be quantitatively assessed. Therefore, the clustering accuracies $G\,\%$ of the whole dataset consisting of unlabeled data and labeled seeds could be calculated by $Q\,\% \cdot (100 - P)\,\% + P\,\%$. On each dataset, the SeedKM, SeedKHM and NSeedKHM were run 20 times for the different P and we report in Figs. 20.2, 20.3 and 20.4 the average accuracies $G\,\%$ of the whole dataset obtained over these 20 runs. Furthermore, let $q = 3$, $\varepsilon = 10^{-5}$ and $E = 2$ in the NSeedkHM.

As shown in Figs. 20.1, 20.2 and 20.3, we can see that the NSeedKHM achieves the best clustering performance compared with the SeedKM and SeedKHM. Although the SeedKM and SeedKHM can improve the clustering accuracies, there are the drastic distanctions between them when an equal amount of labelled data is used. For example, on the DUNN dataset with the number of seeds, which is 12 % of the whole dataset, clustering accuracies both the SeedKM and SeedKHM are less than 78 % while the corresponding accuracy of the

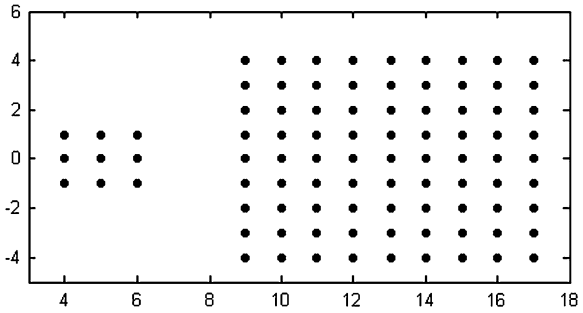


Fig. 20.1 The DUNN dataset

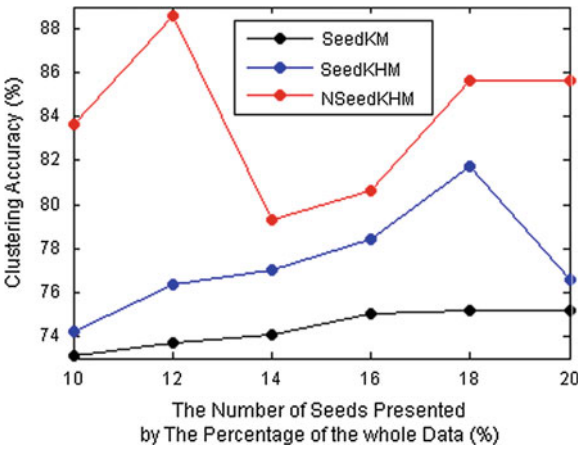


Fig. 20.2 Comparison of clustering accuracies on the DUNN dataset

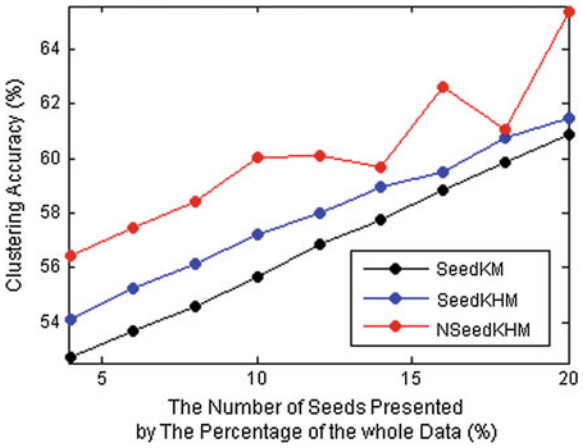


Fig. 20.3 Comparison of clustering accuracies on the ionosphere dataset

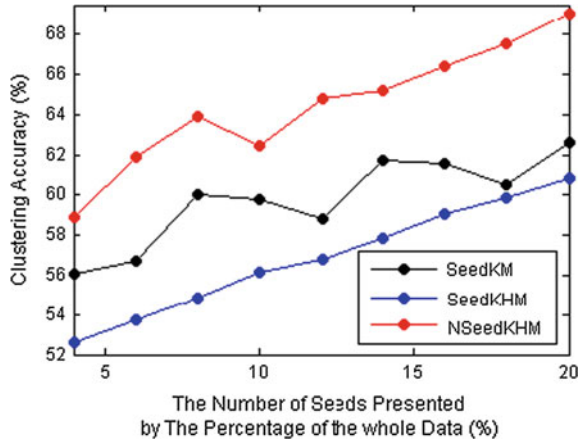


Fig. 20.4 Comparison of clustering accuracies on the Haberman dataset

NSeedKHM is more than 88 %. Moreover, both the SeedKHM and NSeedKHM belong to two semi-supervised variants of the unsupervised k-harmonic means clustering. However, from Fig. 20.3, we can see that the clustering performance of the SeedKHM is worse than the SeedKM but the NSeedKHM can obtain the best clustering accuracies and show good advantage over them.

20.5 Conclusion

In this paper, the presented method called NSeedKHM applies multiple initial centers to the clustering initialization. In the SeedKM and SeedKHM, the number of initial centers is equal to the number of clusters. So there is a difference between the NSeedKHM and them. The number of the initial centers is more than the number of clusters in the NSeedKHM. Experimental results show that the proposed NSeedKHM can lead to better clustering performance compared with other semi-supervised clustering algorithms such as the SeedKM and SeedKHM.

Acknowledgments This research is supported by the Open Foundation of the Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, China (No.2011-01). This research is also supported by the Scientific Research Foundation of Nanjing University of Posts and Telecommunications (No.NY210078).

References

1. Jain, A.K., et al.: Data clustering: a review. ACM Comput. Surv. **31**(3), 256–323 (1999)
2. Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Trans. Neural Netw. **16**(3), 645–678 (2005)
3. Tou, J.T., Gonzalez, R.C.: Pattern Recognition Principles. Addison-Wesley, London (1974)
4. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
5. Krishnapuram, R., et al.: Low complexity fuzzy relational clustering algorithms for web mining. IEEE Trans. Fuzzy Syst. **9**(4), 595–607 (2001)
6. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
7. Matinetz, T.M., et al.: Neural-gas network for vector quantization and its application to time-series prediction. IEEE Trans. Neural Netw. **4**(4), 558–568 (1993)
8. Zhang, B., Hus, M., Dayal, U.: K-harmonic means- a data clustering algorithm. Technical Report HPL-1999-124, Hewlett-Packard Laboratories (1999)
9. Zhang, B., Hsu, M., Dayal, U.: K-harmonic means. In: Proceedings of International Workshop on Temporal, Spatial and Spatio-temporal Data Mining, Lyon, France (2000)
10. Yang, F.Q., Sun, T.L., Zhang, C.H.: An efficient hybrid data clustering method based on k-harmonic means and particle swarm optimization. Expert Syst. Appl. **36**(6), 9847–9852 (2009)
11. Hammerly, C., Elkan, C.: Alternatives to the k-means algorithm that find better clusterings. In: Proceedings of the 11th International Conference on Information and Knowledge Management, pp. 600–607 (2002)

12. Basu, S., Banerjee, A., Mooney, R.J.: Semi-supervised clustering by seeding. In: Proceedings of the Nineteenth International Conference on Machine Learning, pp. 27–34 (2002)
13. Grira, N., Crucianu, M., Boujemaa, N.: Active semi-supervised fuzzy clustering. Pattern Recogn. **41**(5), 1834–1844 (2008)
14. Runkler, T.A.: Partially supervised k-harmonic means clustering. In: Proceedings of IEEE Symposium on Computational Intelligence and Data Mining, pp. 96–103 (2011)
15. UCI Machine Learning Repository. http://www.ics.uci.edu/∼mlearn/MLSummary.html

Chapter 21
Analysis and Optimization of CFS Scheduler on NUMA-Based Systems

Hongyun Tian, Kun Zhang, Li Ruan, Mingfa Zhu, Limin Xiao, Xiuqiao Li and Yuhang Liu

Abstract Non Uniform Memory Access (NUMA) architecture becomes more and more popular as it has better scalability than Uniform Memory Access (UMA). However, all previous work on the operation system scheduler assumed that the underlying system is UMA. As a result, the performance degrades on NUMA machines due to lacking of consideration to the underlying hardware. Researchers discover that the Completely Fair Scheduler (CFS) does not work smoothly on NUMA machines and even interfere performance relative to the O (1) scheduler. In this paper researchers investigate the causes for the performance decline and devise an architecture aware task-bound approach for NUMA system, which can help the CFS scheduler works efficiently on NUMA platforms. The evaluation shows that the approach can upgrade the system performance by more than 60 % on average. The research has great significance to the development and popularity of domestic operating system.

Keywords NUMA · CFS scheduler · Operating system · High-performance computer

H. Tian (✉) · K. Zhang · L. Ruan (✉) · M. Zhu · L. Xiao · X. Li · Y. Liu
State Key Laboratory of Software Development Environment, Beijing, China
e-mail: sympathyh@gmail.com

L. Ruan
e-mail: ruanli@buaa.edu.cn

H. Tian · L. Ruan
School of Computer Science and Engineering, Beihang University, Beijing, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_21,

21.1 Introduction

UMA architecture has been widely used in kinds of computer architectures. As shown in Fig. 21.1, all cores access to the same memory node according to the bus line. However, the memory access according to the bus line will be sharply increases as the number of cores per processor increases. As a result, bus contention turns to be the bottleneck of the system. New multicore systems increasingly use the NUMA architecture due to its better decentralized and scalable nature than UMA. There are multiple memory nodes in the NUMA systems. Each node has its own memory controller, compute nodes use Hyper Transport Bus to connect with each other. Each core can access to the memory on its own node and other nodes with different access latency, the memory access to the local node (*local access*) can be faster than the remote node (*remote access*). The bus contention is diminished but schedule strategy needs to be carefully decided to avoid remote memory access.

Linux is a leading operating system on servers and other big iron systems [1]. The task scheduler is a key part of Linux operating system and Linux continues to evolve and innovate in this area. A lot of good schedulers have been implemented by the kernel developers. O(1) scheduler and CFS scheduler are two most popular schedulers among them.

The O(1) scheduler is a multi-queue scheduler, each processor has a operation queue, but it cannot detect the node layer on NUMA systems. As a result, it cannot guarantee the process in scheduling keep running on the same node. Therefore, Eirch Focht developed a node affinitive NUMA scheduler based on the O(1) scheduler. But the O (1) scheduler needs large mass of code to calculate heuristics and became unwieldy in the kernel [2]. Ingo Molnar then developed the CFS based on some of the ideas from Kolivas's RSDL scheduler. CFS has been a part of the Linux since kernel 2.6.23. The purpose of CFS is to make sure that all the processes need run time could get an equal and fair share of processing time. It makes a progress in the fairness of assigning runtime among tasks but unfortunately it didn't take the under hardware layer into account. Processes may need to remote

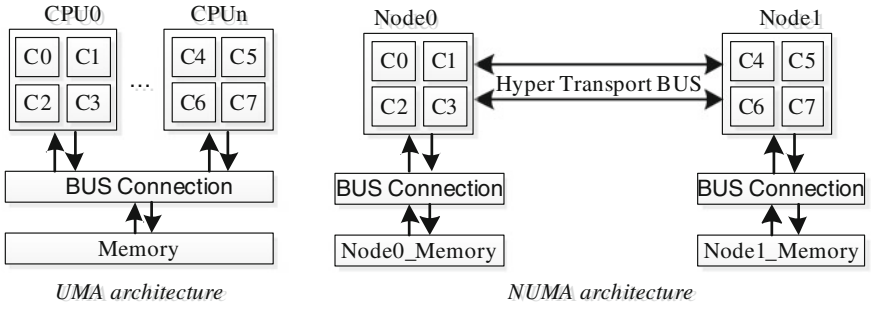


Fig. 21.1 Schematic overview of UMA and NUMA systems

access to its memory frequently, which can cause the performance degrade sharply. As a result, it cannot work well on the NUMA platform compared with the O(1) scheduler with NUMA patch.

We discover that the CFS scheduler not only fails to managing processes effectively on NUMA systems but even hurts performance when compared to the O(1) scheduler with Eirch Focht's NUMA patch. Our experiment setup on an NUMA system based on Loongson CPU, we use LMbench to evaluate the pipe bandwidth and latency, the test score shows that CFS scheduler will degrade as much as 40 % relative to the O(1) scheduler with NUMA patch.

The focus of our study is to investigate why CFS fails to work smoothly on NUMA platforms and devise the architecture aware task-bound approach that would help CFS work efficiently on NUMA platforms. The rest of this paper is organized as follows. Section 21.2 demonstrates why CFS scheduler fails to work well on NUMA systems. Section 21.3 presents our improved measure. Section 21.4 evaluates the task-bound approach. Section 21.5 discusses the related work before we make a conclusion about our research in Sect. 21.6 and present our acknowledgment in last section.

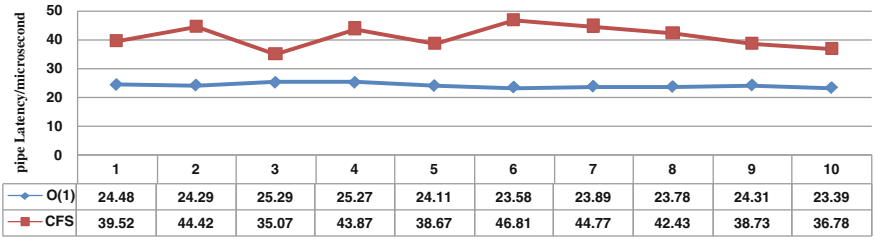
21.2 Motivation

The focus of this section is to experimentally demonstrate why CFS fails to work smoothly on NUMA platforms. We quantify the effects of performance degradation with benchmarks from the LMbench benchmark suite. We perform experiments on a Dual way NUMA server equipped with a Loongson3 processor per node running at 1 GHz, and 4 GB of RAM per node. The kernel of the operating system is Linux 2.6.36.1. Figure. 21.1 schematically represents the architecture of the dual way server.

To quantify the effects of performance degradation caused by the CFS scheduler, we run the bw_pipe (a tool to test the pipe communication bandwidth) and lat_pipe (a tool to test the pipe communication latency) sub items of LMbench to test the pipe communication bandwidth and latency. Besides, we look into the schedule pattern of the benchmark process use the linux command top.

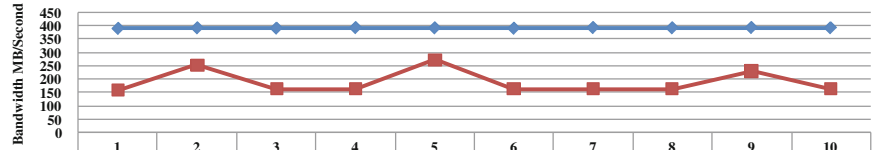
As we depicted in Figs. 21.2 and 21.3, the test scores show that the pipe latency with CFS scheduler grows 67.9 % on average while the bandwidth degrades 51.6 % compared with the NUMA patched O(1) scheduler. That is really bad! Besides, the test scores with CFS scheduler show strong randomness feature while the test scores with O(1) scheduler are far more stable.

To quantify the cause of the big difference between O(1) NUMA scheduler and CFS scheduler, we use top to look into the schedule pattern of the benchmark process during the test. Finally we find out that the test processes were scheduled among the 8 cores randomly by the CFS scheduler, while the O(1) NUMA scheduler always try to let the process running on the same core during the test. These results demonstrate a very important point that the CFS scheduler cannot



	1	2	3	4	5	6	7	8	9	10
O(1)	24.48	24.29	25.29	25.27	24.11	23.58	23.89	23.78	24.31	23.39
CFS	39.52	44.42	35.07	43.87	38.67	46.81	44.77	42.43	38.73	36.78

Fig. 21.2 Contrast test with lat_pipe in LMbench



	1	2	3	4	5	6	7	8	9	10
O(1)	389.27	391.28	390.25	392.18	391.23	389.78	392.47	391.39	392.59	391.69
CFS	158.4	251.38	164.13	164.3	271.47	161.89	164.39	163.75	230.21	164.18

Fig. 21.3 Contrast test with bw_pipe in LMbench

work efficiently due to lacking of consideration of the hardware architecture, the CFS scheduler does not distinguish the cores on the NUMA nodes and just assign tasks randomly to the cores.

Now that we are familiar with causes of performance degradation on NUMA systems, we are ready to explain why CFS scheduler fails to work efficiently on NUMA platforms. The main idea behind the CFS is maintaining balance fairness. To determine the balance, the CFS maintains the amount of time provided to a given task which called the virtual runtime, and the CFS maintains a virtual runtime ordered red-black tree (see Fig. 21.4) rather than run queue as has been done in prior Linux scheduler. The CFS scheduler always choose the process on the most left node of the RB tree and choose a free core to run the process.

Suppose that there are several processes p0, p1, p2, p3, p4, p5, p6, p7, p8, p9 in the system, the orders of the virtual runtime among these processes are $p9 > p8 > p7 > p6 > p5 > p4 > p3 > p2 > p1 > p0$. The CFS scheduler tries to choose the smaller virtual runtime process to run first. As a result, p0 to p7 are assigned to run on the 8 cores while p8 and p9 still in the RB tree waiting to be scheduled. In the next clock interrupt, the p7 finishes its work and then be deleted from the RB tree while the p0 process was moved to the end of the tree as its runtime increases, the p8 and p9 then get the chance to run on the cores. But in the come clock interrupt, another process, take p5 for example, finishes its work and been deleted from the tree. Then at this clock interrupt, the p0 is reassigned to core 5. The problem of remote memory access coming out as p0's memory is on the memory of node 0.

To summarize, CFS scheduler ignores the under layer of the hardware and finally causes the performance degradation. It fails to eliminate remote memory

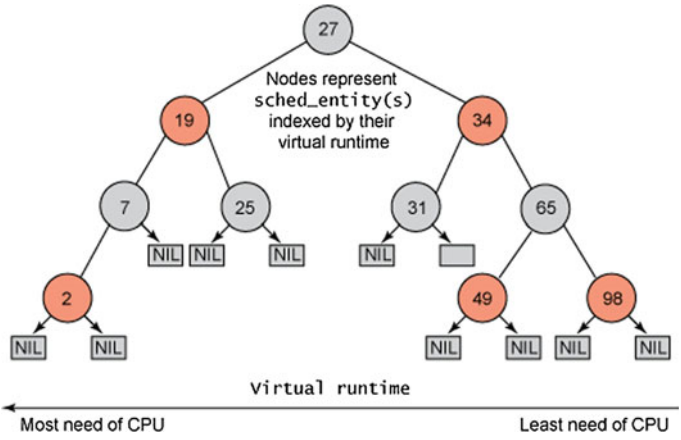


Fig. 21.4 Example of a *red-black* tree [3]

access and even introduces remote latency overhead. To solve this problem, we devise the task-bound approach to avoid remote memory access automatically.

21.3 Implementation

In order to devise the automatic task-bound approach exploiting NUMA architectures, it needs some kind of description to the system. For instance the Advanced Configuration and Power Interface Specification (ACPI), it provides the distance between hardware resources on different NUMA nodes [4]. But the ACPI does not define how this table is filled, and furthermore the ACPI does not make sense for MIPS processors.

Here we propose the concept of system topology matrix, the system topology matrix only needs to know how many cores are there in the system. Dirk proposed a similar concept of system distance matrix [5], but their matrix needs information from the system initialization and cannot be filled automatically.We implement the system topology matrix in the kernel after the kernel get the number of the cores in the system, for example, if the system has 8 cores, then we create a double dimensional array with ST_matrix [8][8] (see Fig. 21.5) to express the topology of the system. ST_matrix [0][1] means the distance between the core0 and core1, we normalized the data such that the cores on the same node results in a value of 1.

We use an average latency test to measure the communication latency between eight threads running on all eight cores, each test process is bounded on a core. For high performance technical computing application, the connect latency and the memory bandwidth frequently are the critical performance bottleneck, thus optimizing application code for connect latency and memory bandwidth is very important. Fig. 21.6 shows the results of our latency tests. The measured matrix

	Core0	Core1	Core2	Core3	Core4	Core5	Core6	Core7
Core0	0	1	1	1	1	1	1	1
Core1	1	0	1	1	1	1	1	1
Core2	1	1	0	1	1	1	1	1
Core3	1	1	1	0	1	1	1	1
Core4	1	1	1	1	0	1	1	1
Core5	1	1	1	1	1	0	1	1
Core6	1	1	1	1	1	1	0	1
Core7	1	1	1	1	1	1	1	0

Fig. 21.5 Initialization of the system topology matrix ST_matrix [8][8]

	Core0	Core1	Core2	Core3	Core4	Core5	Core6	Core7
Core0	0	1	1	1.5	4	5	4	6
Core1	1	0	1.5	1	4	5	5	6
Core2	1	1.5	0	1	5	6	4	5
Core3	1.5	1	1	0	5	6	4	5
Core4	4	4	5	5	0	1	1	1.5
Core5	5	5	6	6	1	0	1.5	1
Core6	4	5	4	4	1	1.5	0	1
Core7	6	6	5	5	1.5	1	1	0

Fig. 21.6 Normalized score of system topology matrix

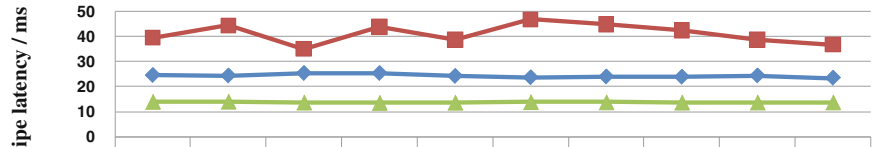
depict huge distance differences between remote nodes and we reset the cores to two nodes, core0 to core3 to node0 while core4 to core7 to node1.

After we get the system topology information according to the topology test module, we reset the cpuset of the system and ergodic the processes in the system once to bound them to the different nodes use the *schedule_set_affinity*(). Any process created after the test module will be automatically bound to a node. Then the CFS scheduler can schedule these processes on the node and remote memory is eliminated.

21.4 Evaluation

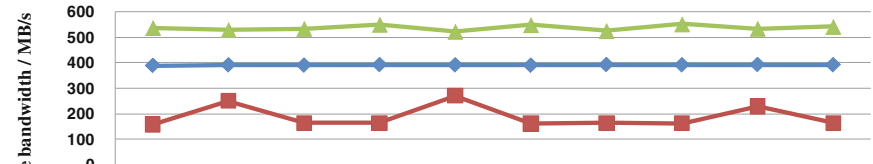
In this section we evaluate our architecture aware task-bound approach with the same environment we used in Sect. 21.2. We evaluate the benchmark with the task-bound approach on and off.

The task-bound approach can help CFS scheduler magically on the NUMA platform according to our tests. As depicted in Figs. 21.7 and 21.8, the test scores are far more excellent than the original CFS scheduler and are also better than the O(1) scheduler with NUMA patch. The pipe latency has been reduced by 66 % on average compared with the CFS scheduler, and also smaller than the average of the O(1) scheduler about 42.9 %. The pipe bandwidth has been upgraded by more than 196 % on average relative to the CFS scheduler and also 37 % higher than the O(1) scheduler on average. Besides, our test scores with task-bound approach are



	1	2	3	4	5	6	7	8	9	10
O(1)	24.48	24.29	25.29	25.27	24.11	23.58	23.89	23.78	24.31	23.39
CFS	39.52	44.42	35.07	43.87	38.67	46.81	44.77	42.43	38.73	36.78
CFS+TB	14.03	14.09	13.73	13.61	13.69	13.9	13.82	13.72	13.76	13.78

Fig. 21.7 Contrast test with lat_pipe in LMbench



	1	2	3	4	5	6	7	8	9	10
O(1)	389.27	391.28	390.25	392.18	391.23	389.78	392.47	391.39	392.59	391.69
CFS	158.4	251.38	164.13	164.3	271.47	161.89	164.39	163.75	230.21	164.18
CFS+TB	537	528.95	533.65	549.88	522.37	548.32	525.09	551.29	534.11	541.51

Fig. 21.8 Contrast test with bw_pipe in LMbench

very stable. Our evaluation demonstrates that our task-bound is significantly useful to help the CFS scheduler work efficiently on NUMA systems.

21.5 Related work

Research on NUMA related system optimizations dates back many years. Many research make efforts to address the computation and related memory on the same node [5, 6–8, 9]. None of the previous efforts, however, addressed automatic sort the system source and bound the task to the subsystem.

Li et al. [10] analyzed the O(1) scheduler and introduced a hierarchical scheduling algorithm based on NUMA topology. Their algorithm depends on the topology information provided by the system initialization and can not be used on the CFS scheduler anymore. Our algorithm is based on the architecture-aware module and can be ported to other platforms.

Blagodurov et al. [11] promoted the concept of resource conscious and presented a contention-aware scheduler, they identified threads complete for shared resources of a memory domain and placed them into different domain while put the independent processes on the same node, they tried to keep processes and their

memory on the same memory domain. Kamali in his master thesis [12] demonstrated the influence of remote memory access to the NUMA systems.

Dirk et al. [9, 13] proposed a platform-independent approach to describe the system topology, they use a distance matrix to provide system information, but their implantation depends on the user-defined strategies, only expert users can take advantage of their approach. Bosilca [5] proposed a framework as a middleware of MPI to tune types of shared memory communications according to the locality and topology.

21.6 Conclusion

Researchers have discovered that the original CFS scheduler fails to work efficiently on NUMA platforms due to lacking of consideration to the underlying hardware. Remote memory access occurring when the scheduler assigns tasks fairly on all the nodes. To address this problem, researchers devise the architecture aware task-bound approach. The evaluation shows that task-bound approach is of signality to the efficient work of CFS scheduler on NUMA systems. The research has a great significance to the development and popularity of domestic operating system.

Acknowledgments Our research is sponsored by the National "Core electronic devices high-end general purpose chips and fundamental software" project under Grant No.2010ZX01036-001-001, the Hi-tech Research and Development Program of China (863 Program) under Grant NO.2011 AA01A205, the National Natural Science Foundation of China under Grant NO.60973007, the Doctoral Fund of Ministry of Education of China under Grant NO.20101102110018, the Beijing Natural Science Foundation under Grant NO.4122042, the fund of the State Key Laboratory of Software Development Environment under Grant NO.SKLSDE-2012ZX-07.

References

1. Burkhardt, H.: KSR. June 2011 | TOP500 Supercomputing Sites (2011)
2. Jones, T.: Inside the Linux scheduler——The latest version© Copyright IBM Corporation (2006)
3. Jones, T.: Inside the Linux 2.6 Completely Fair Scheduler© Copyright IBM Corporation (2009)
4. Hewlett-Packard, Intel, Toshiba.: Advanced configuration and power interface (2011)
5. Ma, T., Bosilca, G., Bouteiller, A., Dongarra, J.J.: Locality and topology aware intra-node communication among multicore CPUs. In: Proceedings for EuroMPI, pp. 265–274 (2010)
6. Li, T., Baumberger D, et al.: Efficient operating system scheduling for performance-asymmetric multi-core architectures. In: Proceedings of Supercomputing (2007)
7. Azimi, T., et al.: Thread clustering: sharing-aware scheduling on multiprocessors. In: Proceedings of Eurosys (2007)
8. Corbalan, J., Martorell, X.: Evaluation of the memory page migration influence in the system performance. In: Proceedings of super computing, pp. 121–129 (2003)

9. Blagodurov, S., et al.: User-level scheduling on NUMA multicore systems under Linux. ACM Trans. Comput. Syst. **28**(4), Article 8 (2010)
10. Li, X.: NUMA scheduling algorithm based on affinitive node. Comput. Eng. 32(1), 99–101 (2006)
11. Blagodurov, S., Zhuravle, S., et al.: A case for NUMA-aware contention management on multicore systems. In: PaCT, pp. 557–558 (2010)
12. Kamali, A: Sharing Aware Scheduling on Multicore Systems. Simon Fraser University, Burnaby (2010)
13. Schmidl, D.: Towards NUMA support with distance information. In: IWOMP'11 Proceedings of the 7th International Conference on OpenMP, pp. 69–79. Springer, Berlin, Heidelberg © (2011)

Chapter 22
Web Crawler for Event-Driven Crawling of AJAX-Based Web Applications

Guoshi Wu and Fanfan Liu

Abstract This paper describes a novel technique for crawling Ajax-based applications through "event-driven" crawling in web browsers. The algorithm uses the browser context to analyse the DOM, scans the DOM-tree, detects elements that are capable of changing the state, triggers events on those elements and extracts dynamic DOM content. For illustration, an AJAX web application is utilized as an example to explain the approach. Additionally, the authors implement the concepts and algorithms discussed in this paper in a tool. Finally, the authors report a number of empirical studies in which they apply their approach to a number of representative AJAX applications. The results show that their method has a better performance often with a faster rate of state discovery. The "event-driven" crawling can effectively and accurately crawl dynamic content from Ajax-based applications.

Keywords AJAX · Event-driven crawling · Web crawler

22.1 Introduction

Web applications have been undergoing a significant change in recent years. More and more applications are dynamic and interactive: Javascript applications, Asynchronous JavaScript and XML (AJAX) [1] applications, Rich Internet Applications are already handling much of the information on the web, providing a high level of user interactivity. Highly visible examples include Google Mail [2] and Google Docs [3].

G. Wu (✉) · F. Liu
School of Software Engineering of BUPT, Beijing University of Posts and Telecommunications, Beijing, China
e-mail: renjianbaoxingtuan@gmail.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_22,

Crawling AJAX-based applications is more difficult than crawling traditional multi-page web applications. In traditional web applications, states are explicit, and correspond to pages that have a unique URL. In AJAX applications, however, the state of the user interface is determined dynamically, and through changes in the DOM [4] that are only visible after executing the corresponding Javascript code.

Current search engines, such as Google and Yahoo, fail to index these applications correctly since they ignore forms and client-side scripting. The web content behind forms and client-side scripting is referred to as the hidden-web. Although there has been extensive research on crawling and exposing the data behind forms, crawling the hidden-web has gained very little attention so far.

In this paper, we propose an approach called "event-driven" crawling, it is based on a crawler which can exercise client-side code, identify clickable elements that change the state and trigger these events to automatically walk through different states of a high dynamic AJAX site.

The paper is further structured as follows: In Sect. 22.2, we present a detailed discussion of our crawling algorithm and technique. Experimental results are shown in Sect. 22.3, while Sect. 22.4 contains the conclusions and future work.

22.2 A Method for Crawling AJAX

In this section, we propose a generic solution for AJAX Crawl.

22.2.1 User Interface States

EBay [5] is an E-commerce site including AJAX parts. Figure 22.1 displays schematically the eBay GUI for a product. The eBay interface for a given product includes product details and comments from the users. The comments are loaded from the server using AJAX and changed through two buttons (next and previous) or through a menu with the page number (1, 2, etc.), but the URL of the page remains the same.

Therefore, the state changes in one AJAX page can be modeled by recording the paths (events) to these DOM changes to be able to navigate the different states [6]. For that purpose, we define a transition graph as follows:

Definition 22.1 A transition graph G for an AJAX site X is a directed graph, denoted by a 3 tuple $< r, V, E >$ where:

1. R is the root node representing the initial state after X has been fully loaded into the browser.
2. V is a set of nodes representing the states. A state is a DOM tree. Each $v \in V$ represents a run-time state in X.

Fig. 22.1 eBay: Comments load using AJAX

3. E is a set of edges between states. Each (v1,v2) ∈ E represents a clickable element connecting two states if and only if state v2 is reached by executing an event on the clickable element in state v1.

The transition graph is best explained in using Fig. 22.2, which models the *next* and *previous* events invoked on the corresponding buttons of the eBay application. State 1 is the initial state. The edges between states are identified with labels (explained in Sect. 22.2.2) of the element to be clicked. Thus, clicking on the// DIV[1]/A[2] element in State 1 leads to State 2.

In addition, we can see that State 2 can be reached either by clicking the next arrow from State 1 or the previous arrow from State 3. In other words, several events can lead to the same state. This brings the issues of duplicate states and infinite loops. In order to avoid regenerating states, we simplify the graph by minimizing the number of redundant edges to form a directed acyclic graph. As shown in Fig. 22.3, the previous arrows are removed from the transition graph.

22.2.2 Clickable Elements

When Javascript is used, the application reacts to user events: click, doubleclick, etc. Figure 22.4 is a highly simplified example, showing different ways in which the next page can be opened. The AJAX example code shows that it is not just the

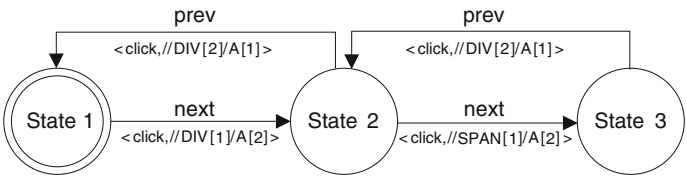


Fig. 22.2 The transition graph visualization

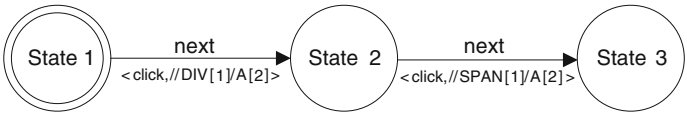


Fig. 22.3 The minimized transition graph visualization

hypertext link element that forms the doorway to the next state. As can be seen, a LI element (line 6) can have a click event attached to it so that it becomes a clickable element.

Definition 22.2 An element that has event-listeners attached to them can be denoted by a 2 tuple < t, x > where:

(1) t is the event type attached to the element: click, doubleclick, mouseover, etc.
(2) x is an XPath expression used to locate the clickable element that can cause the state transition.

As an example, a DIV element (line 3 in Fig. 22.4) can be represented by < click,//DIV [1] > .

22.2.3 Modeling AJAX Web Sites

As opposed to traditional Web, an AJAX Web site contains both static and dynamic content. Each page contains hyperlinks to other web pages as shown in Fig. 22.5. The difference to the traditional Web is that the user may trigger events in the same page (such as *next* and *prev*) which generate new states. The transitions caused by the events may be called AJAX links [7]. As opposed to this, traditional Web Sites are characterized just by a graph of pages, connected by hyperlinks.

The following components, shown in Fig. 22.6, participate in the construction of the crawling architecture based on the model of an AJAX web site:

Embedded Browser: The embedded browser provides a context for accessing the underlying engines and runtime objects, such as the DOM and JavaScript.

```
1  <a href=" javascript:OpenNextPage();">
2  <a href="#" onClick="OpenNextPage();">
3  <div onClick="OpenNextPage();">
4  <a href="next.html" >
5  <a class="next"/>
6  <li class=" next "><a href="#">next</a></li>
7  <!-- jQuery function attaching events to elements having attribute class="next" -->
8  $(".next").click(function() {
9  $("#content"). load("next.html");
10 });
```

Fig. 22.4 Different ways of attaching events to elements

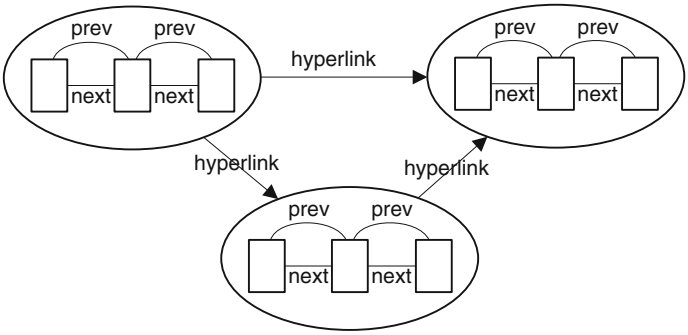


Fig. 22.5 Model of an AJAX Web site: AJAX pages, hyperlinks and AJAX states

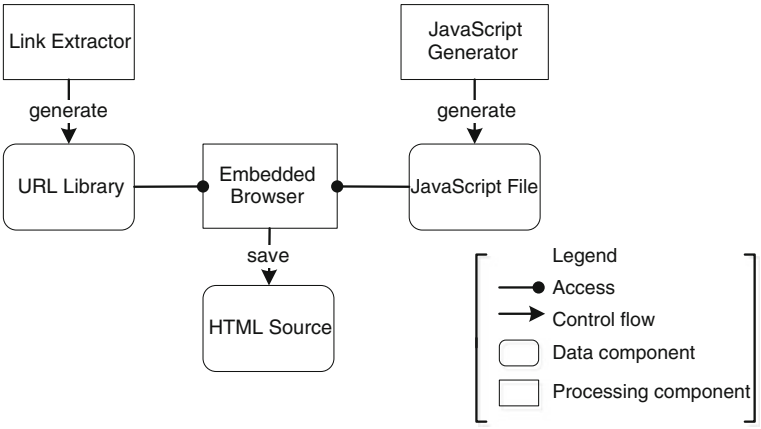


Fig. 22.6 Processing view of the crawling architecture

Hyperlink Extractor: The hyperlink extractor is used to collect hyperlinks which are shown in Fig. 22.5. It works like a traditional crawler and is responsible for building the static hyperlink graph.

URL Library: The URL Library is a data component maintaining all of the hyperlinks discovered by hyperlink extractor.

Javascript Generator: The JavaScript generator is used create Javascript files automatically.

Javascript File: The JavaScript file has access to the embedded browser's DOM. It controls the browser's actions and is responsible for finding clickable elements at runtime and triggering DOM events.

Take eBay as an example. First, the hyperlink extractor reads a seed URL and follows all of the links on the page. The result of crawling is an URL library, which contains all hyperlinks in eBay. The extractor uses a breadth-first approach. Usually, there is a limited number of different hyperlinks that can be extracted. Otherwise, a maximum depth limit can be set. Second, clickable elements must be

marked in a sample AJAX page (e.g., next button of the eBay application) in order to create corresponding Javascript file by Javascript generator. Finally, the URL library that was previously built and the Javascript file are possessed by the embedded browser. The browser then starts crawl procedure and saves the content of each AJAX page as seen in the browser, in exactly its specific state at the time of crawling.

22.2.4 Crawling Algorithm

The algorithm used by these components is shown in Table 22.1. The main procedure (lines 1–4) takes care of initializing the various components and processes involved. The actual, recursive, crawl procedure starts at line 8.

The first step of crawling is to read the initial DOM of the document at a given URL from the URL library (line 6). Crawling AJAX process starts after this initial state has been built (line 8). The DOM object in the initial state is transformed into the corresponding HTML string representation and saved on the file system (line 12). The generated static HTML file represents structure and content of the AJAX application as seen in the browser, in exactly its specific state at the time of crawling.

Table 22.1 The algorithm of crawling AJAX

.

ALGORITHM 1: Crawling AJAX	
	input: URL,clickable c
1	Function main(URL)
2	global *browser* = initEmbeddedBrowser()
3	global *linker* = initHyperlinkExtractor ()
4	Set *U* = *linker*.getHyperlinks(*URL*)
5	for *url* ∈ *U* do
6	*dom*=*browser*.getDOM(*url*)
7	*s* = state(*dom*)
8	crawl(*s*)
9	end for
10	end Function
11	Function crawl(State *s*)
12	*s*.saveAsHTML()
13	*xe* = getXpathExpr(*c*)
14	if (findClickables(*xe*)) then
15	*type* = getEventType(*c*)
16	*browser*.fireEvent(*c*, *type*)
17	waitForDynamicContentToLoad()
18	*newDom* = *browser*.getDOM()
19	*s_1*= state(*newDom*)
20	crawl(*s_1*)
21	end if
22	end Function

The algorithm will find the given clickable element by its XPath expression over the DOM (line 13, 14). After the element has been found, the browser triggers corresponding event on the element (line 15, 16). Whenever the DOM changes, a new state is created and the crawler continues with the new state (line 18–22).

The differences between our algorithm and other algorithm are listed as follows:

1. A static model of an AJAX web site is built by the hyperlink extractor before the actual crawl procedure starts. New state is created and is added it to the model dynamically in crawling process.
2. Using XPath expression, the clickable causing the state transition can be located on the DOM tree before the actual crawl procedure starts, thus bringing two important effects: First, it makes the crawler more efficiently and accurately in detecting clickables, instead of finding a series of candidate elements. Second, it guarantees that state reached after firing an event will always be a new state. This avoids exploring the same state multiple times, making it no longer necessary to compute the differences between two states by means of an enhanced *Diff* algorithm [8] or computing a hash of the content of state [7].
3. Our algorithm makes use of a timer to set delay, waiting for the dynamic AJAX content fully loaded from server, which is not mentioned in other algorithms.

22.3 Experimental Results

We have implemented the concepts presented in this paper in a tool and we applied it to different types of AJAX sites as shown in Table 22.2. In this section, we provide an empirical assessment of some of the key properties of our crawling technique. In particular, we address the accuracy (are the results correct?) and performance, focusing in particular on the performance gains results from crawling.

The results are displayed in Table 22.3. The table lists key characteristics of the sites under study, such as the average DOM size and the total number of detected states.

The performance measurements were obtained on a laptop with Intel P8400 processor 2.26 GHz, with 4 GB RAM and Windows Vista.

Assessing the correctness of the crawling process is challenging, because there is no strict notion of “correctness” with respect to state equivalence. Consequently, an assessment in terms of precision (percentage of correct states) and recall (percentage of states recovered) is impossible to give [9]. To address these concerns, we take a random sampling method. For C1, we select 30 hyperlinks in a total number of 307 hyperlinks randomly. We check whether all the states that can be generated by these hyperlinks are detected by crawler, and for each state, we check if all the AJAX contents are fetched.

Table 22.2 Experiment cases and examples of their clickable elements

Type	Case	AJAX site	Clickable elements
Electronic commerce	C1	http://www.360buy.com/	 next
	C2	http://www.taobao.com/	 next
			<a class = "tb-tab-anchor" data-index = "1" hidefocus = "true" href = "#reviews">
Twitter	C3	http://weibo.com/	next
	C4	http://t.163.com/session	<li class = "js-btn js-next">
			<a href = "javascript:;" > next
Blog	C5	http://blog.sohu.com/	 next
	C6	http://blog.sina.com.cn/	 next

Table 22.3 Results of running crawler on 6 AJAX applications

Case	Hyperlinks	Detected states	Average DOM size (kb)	Crawling rate (kb/s)
C1	307	3650	124	45.4
C2	352	3342	162	64
C3	309	2917	143	43.6
C4	353	3059	102	38.5
C5	224	1021	113	35.1
C6	220	798	98	33

Our results are as follows:

For C1 and C2, the crawler finds all the expected clickables and states with a precision and recall of 100 %.

For C3 and C4, the crawler was able to find 97 % of the expected clickables and reaches a precision of 98 %.

For C5 and C6, the crawler finds all the expected clickables and reaches a precision of 99 %.

It is clear that the running time of the crawler increases linearly with the size of the input (total DOM size). Note that the performance is also dependent on the CPU and memory of the machine, as well as the speed of the server and network properties of the case site.

We compare the performance of our tool with a prototype of major commercial software for testing web applications, a crawling tool for AJAX applications (DataScraper) [10]. The results were obtained using the AJAX sites shown in Table 22.2.

Only our algorithm successfully discovered most of all the states for the applications. The commercial product prototype could not achieve a complete crawl for any of the applications. That is because it did not apply a specific strategy for crawling AJAX but blindly executed the events on a page once without handling DOM changes which may add or remove events to the DOM. DataScraper could crawl some completely but not most of them, its precision is lower than our tool. Table 22.4 shows the number of states discovered by each tool.

Table 22.4 Number of states discovered by DataScraper, commercial software and our tool

Case	Total states	States discovered by		
		Commercial	DataScraper	Our tool
C1	120	46	120	120
C2	104	44	100	104
C3	89	27	82	86
C4	76	23	70	74
C5	58	16	54	58
C6	67	19	64	67

22.4 Conclusion and Future Work

In this paper, the authors have presented a new crawling approach based on the idea of "event-driven" crawling. Their solution aims at crawling dynamic content from AJAX sites efficiently and accurately. Experimental results show that the solution is correct and the crawler performs very well on a set of experimental sites. Furthermore, strengthening the tool by extending its functionality, improving the accuracy, performance, and stability are directions the authors foresee for future work.

References

1. Garrett, J.: Ajax: a new approach to web applications. Adaptive path. http://www.adaptivepath.com/publications/essays/archives/000385.php (2005)
2. Google Mail.: http://www.gmail.com
3. Google Docs.: http://docs.google.com
4. W3C.: Document Object Model (DOM). http://www.w3.org/DOM/ (2005)
5. EBay.: http://www.ebay.com
6. Mesbah, A., Bozdag, E., Deursen, A.v.: Crawling AJAX by inferring user interface state changes. In: Proceedings of the 8th IEEE International Conference on Web Engineering (ICWE'08), IEEE Computer Society, pp. 122–134 (2008)
7. Matter, R.: AJAX Crawl: Making AJAX Applications Searchable. Master's Thesis. ETH, Zurich (2008)
8. Chawathe, S.S., Rajaraman, A., Garcia-Molina, H., Widom, J.: Change detection in hierarchically structured information. In: SIGMOD'96: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, pp. 493–504. ACM Press (1996)
9. Mesbah, A., Deursen, A.v., Lenselink, S.: Crawling Ajax-based web applications through dynamic analysis of user interface state changes. ACM Trans. Web **6**(1), 1–30 (2012)
10. DataScraper.: http://www.gooseeker.com

Chapter 23
On the Universal Approximation Capability of Flexible Approximate Identity Neural Networks

Saeed Panahian Fard and Zarita Zainuddin

Abstract This study presents some class of feedforward neural networks to investigate the universal approximation capability of continuous flexible functions. Based on the flexible approximate identity, some theorems are constructed. The results are provided to demonstrate the universal approximation capability of flexible approximate identity neural networks to any continuous flexible function.

Keywords Flexible approximate identity · Flexible approximate identity activation functions · Flexible approximate identity neural networks · Uniform convergence · Universal approximation

23.1 Introduction

One of the most important issues in theoretical studies for neural networks is concerned with the universal approximation capability of feedforward neural networks. There have been many papers related to this topic over the past 30 years [1].

A few authors [2–5] recently deal with the concept of approximation of non-linear functions by approximate identity neural networks (AINNs). These networks are based on the widely-known sequences of functions named approximate identities [6].

Flexible approximate identity neural networks (FAINNs) are the generalization of AINNs. These networks use flexible approximate identity as activation functions with a traditional multilayer architecture. Lately, new model of feedforward neural networks called the generalized Gaussian radial basis function neural networks has been proposed which is shown in [7]. These neural networks are special case of the FAINNs.

S. Panahian Fard (✉) · Z. Zainuddin
School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Georgetown, Pulau Penang, Malaysia
e-mail: saeedpanahian@yahoo.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_23,

The main goal of this study is to investigate the universal approximation capability of FAINNs to any continuous flexible function. Based on a convolution linear operator in the real linear space of all continuous flexible functions, some theorems are presented. These theorems verify the approximation capability of FAINNs.

This paper is organized as follows. In Sect. 23.2, as the main technical tool, the definition of flexible approximate identity is given. And basic definitions and theorems are introduced. The main result is presented in Sect. 23.3. Conclusions are drawn in Sect. 23.4.

23.2 Preliminaries

The definition of flexible approximate identity which will be used in Theorem 1 is presented as follows.

Definition 1 Let $A = A(a_1, \ldots, a_m), a_i \in \mathbb{R}, i = 1, \ldots, m$ be any parameters. $\{\varphi_n(x, A)\}_{n \in N}, \varphi_n(x, A) : R \to R$ is said to be a flexible approximate identity if the following properties hold:

(1) $\int_{\mathrm{R}} \varphi_n(x, A)dx = 1;$

(2) Given ε and $\delta > 0$, there exists N such that if $n \geq N$ then

$$\int_{|\mathrm{x}| > \delta} |\varphi_{\mathrm{n}}(\mathrm{x}, \mathrm{A})|\mathrm{dx} \leq \varepsilon.$$

Now, we will be able to give the following theorem in order to construct the hypothesis of Theorem 3 in the next section.

Theorem 1 *Let* $\{\varphi_n(x, A)\}_{n \in N}, \varphi_n(x, A) : \mathrm{R} \to \mathrm{R}$ *be a flexible approximate identity. Let* f *be a function on* $C[a, b]$. *Then* $\varphi_n * f$ *uniformly converges to* f *on* $C[a, b]$.

Proof Let $x \in [a, b]$ and $\varepsilon > 0$. There exists a $\delta > 0$ such that $|f(x) - f(y)| < \frac{\varepsilon}{2||\varphi||_1}$ for all $y, |x - y| < \delta$. Let us define $\{\varphi_n * f\}_{n \in N}$ by $\varphi_n(x, A) = n\varphi(nx, A)$. Then,

$$\begin{aligned}
\varphi_{\mathrm{n}} * \mathrm{f}(\mathrm{x}) - \mathrm{f}(\mathrm{x}) &= \int_{\mathrm{R}} \mathrm{n}\varphi(\mathrm{ny},\mathrm{A})\{\ \mathrm{f}(\mathrm{x} - \mathrm{y}) - \mathrm{f}(\mathrm{x})\}\ \mathrm{dy} \\
&= \left(\int_{|\mathrm{y}| < \delta} + \int_{|\mathrm{y}| < \delta} \right) \mathrm{n}\varphi(\mathrm{ny}, \mathrm{A})\{\mathrm{f}(\mathrm{x} - \mathrm{y}) - \mathrm{f}(\mathrm{x})\}\mathrm{dy} \\
&= \mathrm{I}_1 + \mathrm{I}_2,
\end{aligned}$$

where $I_1 + I_2$ are as follows:

$$\begin{aligned}|I_1| &\leq \int\limits_{|y|<\delta} n\varphi(ny,A)\{f(x-y)-f(x)\}dy \\ &< \frac{\varepsilon}{2||\phi||_1} \int\limits_{|y|<\delta} n\varphi(ny,A)dy \\ &= \frac{\varepsilon}{2||\phi||_1} \int\limits_{|t|<n\delta} \varphi(t,A)dt \\ &\leq \frac{\varepsilon}{2||\phi||_1} \int\limits_{R} \varphi(t,A)dt = \frac{\varepsilon}{2}.\end{aligned}$$

For I_2, we have

$$\begin{aligned}|I_2| &\leq 2||f||_{C[a,b]} \int\limits_{|y|\geq\delta} n|\varphi(ny,A)|dy \\ &= 2||f||_{C[a,b]} \int\limits_{|t|\geq n\delta} |\varphi(t,A)|dt.\end{aligned}$$

Since

$$\lim_{n\to+\infty} \int\limits_{|t|>n\delta} |\varphi(t,A)|dt = 0,$$

there exists an $n_0 \in N$ such that for all $n \geq n_0$,

$$\int\limits_{|t|\geq n\delta} |\varphi(t,A)|dt < \frac{\varepsilon}{4||f||_{C[a,b]}}.$$

Combining I_1 and I_2 for $n \geq n_0$, we get

$$||\varphi_n * f(x) - f(x)||_{C[a,b]} < \varepsilon.$$

We use the following (cf. [8]) in order to prove Theorem 3 which is given as the main result in the Sect. 23.3:

Definition 2 Let $\varepsilon > 0$. A set $V_\varepsilon \subset C[a,b]$ is called ε-net of a set V, if $\tilde{f} \in V_\varepsilon$ can be found for $\forall f \in V$ such that $\left\|f - \tilde{f}\right\|_{C[a,b]} < \varepsilon$.

Definition 3 The ε-net is said to be finite if it is a finite set of elements.

Theorem 2 *A set V in $C[a,b]$ is compact iff $\forall\, \varepsilon > 0$ in R there is a finite ε-net.*

Now, we present the universal approximation capability of FAINNs in the next section.

23.3 Main Result

The main aim of this section is to investigate the conditions for the universal approximation capability of FAINNs to any continuous flexible function. Now, the following theorem is proposed to show the universal approximation capability of FAINNs.

Theorem 3 *Let $C[a,b]$ be linear space of all continuous functions on the real interval $[a,b]$,* and *$V \subset C[a,b]$ a compact set. Let $A = A(a_1,\ldots,a_m), a_i > 0, i = 1,\ldots,m$ be any parameters, $\{\varphi_n(x,A)\}_{n\in N}, \varphi_n(x,A) : \mathrm{R} \to \mathrm{R}$ be a flexible approximate identity. Let the family of functions $\left\{\sum_{\mathrm{j}=1}^{\mathrm{M}} \lambda_{\mathrm{j}}\varphi_{\mathrm{j}}(\mathrm{x},\mathrm{A})|\lambda_{\mathrm{j}} \in \mathrm{R}, \mathrm{x} \in \mathrm{R}, \mathrm{M} \in \mathrm{N}\right\}$, be dense in $C[a,b]$, and given $\varepsilon > 0$. Then there exists $\mathrm{N} \in \mathrm{N}$ which depends on V and ε but not on f, such that for any $f \in V$, there exist weights $c_k = c_k(f,V,\varepsilon)$ satisfying*

$$\left\| f(x) - \sum_{i=1}^{N} c_k\varphi_k(x,A) \right\|_{C[a,b]} < \varepsilon$$

Moreover, every c_k is a continuous function of $f \in V$.

Proof The method of proof is analogous to that of Theorem 1 in [9]. Because V is compact, for any $\varepsilon > 0$, there is a finite $\frac{\varepsilon}{2}$-net $\left\{f^1,\ldots,f^M\right\}$ for V. This implies that for any $f \in V$, there is an f^j such that $\|f - f^j\|_{C[a,b]} < \frac{\varepsilon}{2}$. For any f^j, by assumption of the theorem, there are $\lambda_i^j \in R, N_j \in \mathrm{N}$, and $\varphi_i^j(x,A)$ such that

$$\left\| f^j(x) - \sum_{i=1}^{N_j} \lambda_i^j\varphi_i^j(x,A) \right\|_{C[a,b]} < \frac{\varepsilon}{2}. \tag{23.1}$$

For any $f \in V$, we define

$$F_-(f) = \left\{ j \middle| \|f - f^j\|_{C[a,b]} < \frac{\varepsilon}{2} \right\},$$

$$F_0(f) = \left\{ j \middle| \|f - f^j\|_{C[a,b]} = \frac{\varepsilon}{2} \right\},$$

$$F_+(f) = \left\{ j \middle| \|f - f^j\|_{C[a,b]} > \frac{\varepsilon}{2} \right\}.$$

Therefore, $F_-(f)$ is not empty according to the definition of $\frac{\varepsilon}{2}$ -net. If $\tilde{f} \in V$ approaches f such that $\|f - f^j\|_{C[a,b]}$ is small enough, then we have $F_-(f) \subset F_-(\tilde{f})$ and $F_+(f) \subset F_+(\tilde{f})$. Thus $F_-(\tilde{f}) \cap F_+(f) \subset F_-(\tilde{f}) \cap F_+(\tilde{f}) = \emptyset$, which implies $F_-(\tilde{f}) \subset F_-(f) \cup F_0(f)$. We finish with the following.

$$F_-(f) \subset F_-(\tilde{f}) \subset F_-(f) \cup F_0(f). \tag{23.2}$$

Define

$$\begin{aligned} d(f) &= \left[\sum\nolimits_{j \in F_-(f)} \left(\frac{\varepsilon}{2} - \|f - f^j\|_{C[a,b]}\right)\right]^{-1} \quad \text{and} \\ f_h &= \sum_{j \in F_-(f)} \sum_{i=1}^{N_j} d(f)\left(\frac{\varepsilon}{2} - \|f - f^j\|_{C[a,b]}\right)\lambda_i^j \varphi_i^j(x, A) \end{aligned} \tag{23.3}$$

Then $f_h \in \sum_{j=1}^{M} \lambda_j \varphi_j(x, A)$ approximates f with accuracy ε:

$$\begin{aligned} &\|f - f_h\|_{C[a,b]} \\ &= \left\| \sum_{j \in F_-(f)} d(f)\left(\frac{\varepsilon}{2} - \|f - f^j\|_{C[a,b]}\right)\left(f - \sum_{i=1}^{N_j} \lambda_i^j \varphi_i^j(x, A)\right)\right\|_{C[a,b]} \\ &= \left\| \sum_{j \in F_-(f)} d(f)\left(\frac{\varepsilon}{2} - \|f - f^j\|_{C[a,b]}\right)\left(f - f^j + f^j - \sum_{i=1}^{N_j} \lambda_i^j \varphi_i^j(x, A)\right)\right\|_{C[a,b]} \\ &\le \sum_{j \in F_-(f)} d(f)\left(\frac{\varepsilon}{2} - \|f - f^j\|_{C[a,b]}\right)\left(\|f - f^j\|_{C[a,b]} + \left\|f^j - \sum_{i=1}^{N_j} \lambda_i^j \varphi_i^j(x, A)\right\|_{C[a,b]}\right) \\ &< \sum_{j \in F_-(f)} d(f)\left(\frac{\varepsilon}{2} - \|f - f^j\|_{C[a,b]}\right)\left(\frac{\varepsilon}{2} + \frac{\varepsilon}{2}\right) = \varepsilon. \end{aligned} \tag{23.4}$$

In the following step, We prove the continuity of c_k. For the proof, we use (23.2) to obtain

$$\begin{aligned} &\sum_{j \in F_-(f)} \left(\frac{\varepsilon}{2} - \|\tilde{f} - f^j\|_{C[a,b]}\right) \\ &\le \sum_{j \in F_-(\tilde{f})} \left(\frac{\varepsilon}{2} - \|\tilde{f} - f\|_{C[a,b]}\right) \\ &\le \sum_{j \in F_-(\tilde{f})} \left(\frac{\varepsilon}{2} - \|\tilde{f} - f^j\|_{C[a,b]}\right) \\ &\quad + \sum_{j \in F_0(f)} \left(\frac{\varepsilon}{2} - \|\tilde{f} - f^j\|_{C[a,b]}\right). \end{aligned} \tag{23.5}$$

Let $f \to \tilde{f}$ in (23.5), then we have

$$\sum_{j\in F_-(\tilde{f})}\left(\frac{\varepsilon}{2}-\left\|\tilde{f}-f^j\right\|_{C[a,b]}\right)\to\sum_{j\in F_-(f)}\left(\frac{\varepsilon}{2}-\left\|f-f^j\right\|_{C[a,b]}\right) \tag{23.6}$$

This obviously demonstrates $d(\tilde{f})\to d(f)$. Thus, $\tilde{f}\to f$ results

$$d(\tilde{f})\left(\frac{\varepsilon}{2}-\left\|f-f^j\right\|_{C[a,b]}\right)\lambda_i^j\to d(f)\left(\frac{\varepsilon}{2}-\left\|f-f^j\right\|_{C[a,b]}\right)\lambda_i^j. \tag{23.7}$$

Let $N=\sum_{j\in F_-(f)}N_j$ and define c_k in terms of

$$\begin{aligned} f_h &= \sum_{j\in F_-(f)}\sum_{i=1}^{N_j} d(f)\left(\frac{\varepsilon}{2}-\left\|f-f^j\right\|_{C[a,b]}\right)\lambda_i^j\varphi_i^j(x,A)\\ &\equiv \sum_{k=1}^{N} c_k\varphi_k(x,A) \end{aligned}$$

From (23.7), c_k is a continuous functional of f. This completes the proof.

23.4 Conclusion

Some class of feedforward neural networks with a traditional multilayer architecture has been constructed to obtain an approximation of any flexible continuous function. By employing the flexible approximate identity, Theorem 1 is established. This theorem constructs the hypothesis for Theorem 3. In Theorem 3, it has been proved that if a flexible approximate identity neural networks with a hidden layer is dense in C [a,b], then for a given compact set $V \subset C$ [a,b] and an error bound ε, one can approximate any continuous flexible function $f \in V$ with the accuracy ε.

References

1. Ismailov, V.E.: Approximation by neural networks with weights varying on a finite set of directions. J. Math. Anal. Appl. **398**, 72–83 (2012)
2. Hahm, N., Hong, B.I.: An approximation by neural networks with a fixed weight. Comput. Math. Appl. **47**, 1897–1903 (2004)
3. Li, F.: Function approximation by neural networks. In: Proceedings 5th International Symposium on Neural Networks, pp. 348–390 (2008)
4. Turchetti, C., Conti, M., Crippa, P., Orcioni, S.: On the approximation of stochastic processes by approximate identity neural networks. IEEE Trans. Neural Netw. **9**, 1069–1085 (1998)
5. Zainuddin, Z., PanahianFard, S.: Double approximate identity neural networks universal approximation in real Lebesgue spaces. Lect. Notes Comput. Sci. **7663**, 409–415 (2012)

6. Wheeden, R.L., Zygmund, A.: Measure and Integral. Marcel Dekker, New York (1977)
7. Navaro, F.F., Martínez, C.H., Monederi, J.S., Gutiérrez, P.A.: MELM-GRBF: a modified version of the extreme learning machine for generalized radial basis function neural networks. Neurocomputing **74**, 2502–2510 (2011)
8. Lebedev, V.: An Introduction to Functional Analysis and Computational Mathematics. Brikhäuser, Boston (1997)
9. Wu, W., Nan, D., Li, Z., Long, J., Wang, J.: Approximation to compact set of functions by feedforward neural networks. In: Proceedings 20th International Joint Conference on Neural Networks, pp. 1222–1225 (2007)

Chapter 24
A Spectral Clustering Algorithm Based on Particle Swarm Optimization

Feng Wang

Abstract The shortcoming of traditional spectral clustering algorithm is its dependence on initial value. This paper proposes a spectral clustering algorithm based on the particle swarm optimization, considering the characteristic of the good global and local optimization capability and the randomization of initial population. According to the example analysis, the spectral clustering algorithm based on the particle swarm optimization has overcome the shortcoming of excessive dependence on initial value of the traditional spectral clustering algorithm. The accuracy of the cluster is improved.

Keywords Spectral clustering algorithm · F-measure · Particle swarm optimization algorithm

24.1 Introduction

Spectral clustering algorithm is a new class of clustering algorithm established on the basis of the spectral graph theory to transfer the problem of clustering into the optimal partition problem of the graph. The essence of algorithm is to achieve the reduction of dimension process through Laplacian Eigenmap [1]. The main calculation steps include: (1) calculating the similarity between the original set of data points; (2) calculating the feature values and eigenvectors of the similarity matrix; (3) selecting some eigenvectors to cluster with original data.

Spectral clustering algorithm has overcome the shortcomings such as sensitivity to the sample shape and susceptibility of stopping at local optimal solution [2]. However, as the research on spectral clustering algorithm is still in its early stage,

F. Wang (✉)
School of Information Engineering, Lanzhou University of Finance and Economics, Lanzhou, China
e-mail: wfeng@lzcc.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_24,

the algorithm can still be greatly improved. The improvement of the algorithm can be divided into six aspects: establishment of the similarity matrix, determination of the number of cluster, selection of the feature value, selection of the Laplacian matrix, expediting of the clustering process and the improvement of the initial value dependence.

As Particle swarm optimization algorithm has clear advantage in searching the global optimal, the use of particle swarm algorithm to solve the Laplacian matrix selection can overcome the excessive dependence on an initial value of traditional spectral clustering algorithm, and improve the clustering accuracy.

24.2 Particle Swarm Optimization Algorithm

Particle swarm optimization algorithm, PSO for short, was formally proposed by Kennedy and Eberhart in 1995 IEEE international neural network conference published a paper entitled "Particle Swarm Optimization". Particle swarm optimization algorithm is a global optimization algorithm based on swarm intelligence to simulate the migration and social behavior of birds in the feeding process through the cooperation and competition between individuals [3].

In PSO algorithm, each bird in the space is a potential solution of the optimization problem, called particle. Fitness value of each particle is measured by fitness function. Each particle has a speed that determines the direction and distance they fly. The particles then follow the current optimal particle search in the solution space. The algorithm first obtains a group of random particles as the initial solution, and then each particle updates itself by following two extreme values, the global extreme and the individual extreme. The global extreme indicates the current optimal solution the population has found; it is the local extreme if all the particles within the neighborhood of the selected particle have reached to the optimal solution. The individual extreme indicates the optimal solution the particle has found. Through this iteration, particles change their position and speed according to the global or local extreme so as to achieve the optimal solution [4].

Assuming that $R_{N\times D}$ is a target search space indicates a N D-dimensional particle swarm, the i-th particle is: $X_i = (x_{i1}, x_{i2}, \ldots, x_{iD}),\ i = 1, 2, \cdots, N$. The flying speed of the i-th particle is: $V_{\mathrm{i}} = (v_{i1}, v_{i2}, \ldots, v_{iD}),\ i = 1, 2, \cdots, N$. The optimal position searched by the i-th particle so far is called individual extreme, credited as $p_{best} = (p_{i1}, p_{i2}, \cdots, p_{iD}),\ i = 1, 2, \cdots, N$. The optimal position searched by the whole particle swarm so far is called the global extreme, credited as $g_{best} = (p_{g1}, p_{g2}, \cdots, p_{gD})$. Updating the position and speed according to the following formula based on the individual extreme and global extreme:

$$v_{id} = w \cdot v_{id} + c_1 r_1 (p_{id} - x_{id}) + c_2 r_2 \left(p_{gd} - x_{id}\right) \tag{24.1}$$

$$x'_{id} = x_{id} + v_{id} \tag{24.2}$$

$$w = w_{\max} - run\frac{(w_{\max} - w_{\min})}{sumrun} \tag{24.3}$$

In which the range of speed v is $[v_{\min}, v_{\max}]$, the selection of $v_{\max}$ usually given by experience, and general setting for the 10–20 % of the problem space, c_1 and c_2 are learning factor used to adjust the step size of best position movement of the particles point relative to itself or the neighborhood [5]. The experience value of c_1 and c_2 is $\phi = c_1 + c_2 \leq 4.0$, and are usually taken as $c_1 = c_2 = 2$. w is the inertia weight [6]; formula (24.3) shows that w decreases linearly with the increase of the number of iterations. Some literatures suggest that with the increase of update variable [7], the value of w should decrease linearly from 0.9 to 0.4 while r_1 and r_2 are random numbers within the range of [0, 1]. Formula (24.1) consists of three parts: the first part $w \cdot v_{id}$ is "inertia", indicating that the particles' trend of movement is in accordance with the original direction and speed; it reflects the inherent habits of particles; the second part is $c_1 r_1 (p_{id} - x_{id})$, indicating that the particles are approximating to their own historical best position, reflects the memory of the particles; the third part is $c_2 r_2 (p_{gd} - x_{id})$, indicating the particles are to the trend of approximating to the best position of the population or neighborhood, reflects the collaboration or sociality.

In the optimization process, PSO shows the following characteristics: (1) fewer parameters need to be adjusted, the program is easier to describe and implement; (2) the particles own a random variable speed; (3) particle itself has a "memory" ability, and "collaboration" between particles. Because of these characteristics, compared to the other optimal algorithm, the PSO algorithm converges fast, can find local and global optimum and can avoid degradation of optimization to a certain extent.

24.3 Spectral Clustering Algorithm Based on PSO

Spectral clustering algorithm is strongly sensitive to the data input sequence, different input sequence makes the similarity matrix and Laplacian matrix difference, the reason for this is the K means algorithm of the spectral clustering algorithm is dependent on the initial data, and in addition it is easy to fall into local optimum [8, 9]. In consideration of the advantages of the PSO algorithm in global and local optimization, as well as the particle stochastic population characteristics, introducing the PSO algorithm based on the traditional spectral clustering algorithm. The input matrix of the algorithm is the matrix T consisting n eigenvectors of the selected Laplacian matrix, the specific processes are as follows:

(1) Initialize the particle swarm. Randomly initialize the position and velocity of the particles and the classification of matrix T;
(2) Calculate the clustering center and fitness;

$$Center_{jp} = \frac{\sum_{i-1}^{n} \omega_{ij} T_{ip}}{\sum_{i=1}^{n} \omega_{ij}}, \; j = 1, 2, \cdots, k \tag{24.4}$$

Indicate the clustering center of class j. In which:

$$\omega_{ij} = \begin{cases} 1 & \text{if the } i\text{-th data point belongs to class } j \\ 0 & \text{others} \end{cases} \tag{24.5}$$

$$Fittness = \sum_{j=1}^{k} \sum_{i=1}^{n} \omega_{ij} \sum_{p=1}^{h} \left[\left(T_{ip} - Center_{jp} \right) \Big/ \sum_{i=1}^{n} \omega_{ij} \right]^2 \tag{24.6}$$

(3) Determine the individual optimum p_{best} and the group optimumg_{best};
(4) Update the inertia weight according to formula (24.3), update the position and speed of particles according to formula (24.1) and (24.2);
(5) Re-clustering with the clustering center of the position of particles;
(6) Compare and judge the termination condition, if the termination condition is satisfied, then the clustering result is the optimal clustering results, if it is unsatisfied, return to (2).

24.4 Example Analysis

This paper selects three data sets from UCI database, which are Iris data set, Balance data set and Stalog (heart) data set. Iris is the data set of classification of iris, Balance is the data set of scale balanced, and Heart is the data set of the diagnosis of heart disease. The basic information of the data set shows in Table 24.1.

It can be seem in Table 24.1 that the Iris data set has 3 classifications, 5 attributes and one of them is the target attribute which indicates the classification of iris, as well as 150 objects in total; Balance data set has 3 classifications, 5 attributes and one of them is the target attribute which indicates the equilibrium of the balance, as well as 625 objects in total; Heart data set has 2 classifications, 14 attributes and one of them is the target attribute which indicates whether the presence of heart disease, as well as 270 objects in total.

Table 24.1 The composition of the data sets

Data set	Number of samples	Classification	Number of attributes
Iris	150	3	5
Balance	625	3	5
Heart	270	2	14

Table 24.2 Initial classification of the three data sets

Data set	N_i			
	N_1	N_2	N_3	Total
Iris	50	50	50	150
Balance	49	288	288	625
Heart	149	121		270

Notes N_i indicates the number of all of the objects in class i of the data set

Table 24.3 Composition of the cluster of 4 clustering methods

Data set	Algorithm	N_{ij}			
		N_{11}	N_{22}	N_{33}	Total
Iris	SC	50	60	40	150
	P-SC	50	61	39	150
Balance	SC	132	219	274	625
	P-SC	195	268	162	625
Heart	SC	118	152		270
	P-SC	121	149		270

Notes This table indicates the number of all of the objects in cluster j, *SC* indicates the traditional spectral clustering algorithm, *P-SC* indicates the spectral clustering algorithm based on PSO

Without loss of generality, using some of the known target attributes as a class, weed out the target attribute during cluster analysis, and then compare and analyze the result of cluster analysis and the known classification, so as to determine the effectiveness of the clustering algorithm. Therefore, this paper chooses clustering validity evaluation method in the following examples to evaluate with F-measure method which is also an external evaluation method.

In the F-measure evaluation method, the range of F value is [0, 1], and greater value represents better clustering results, the results in Table 24.5 are calculated from the data in Tables 24.2, 24.3, 24.4. Table 24.5 shows that the clustering effect of the spectral clustering algorithm based on PSO is better than the traditional

Table 24.4 Number correctly clustering object in each cluster of two kinds of spectral clustering algorithms

Data set	Algorithm	N_{ij}			
		N_{11}	N_{22}	N_{33}	Total
Iris	SC	50	47	37	134
	P-SC	50	49	37	136
Balance	SC	21	201	187	409
	P-SC	21	224	150	395
Heart	SC	68	70		138
	P-SC	70	69		139

Notes N_{ij} indicates the number of all of the objects in cluster j and class i

Table 24.5 Comparison of the two spectral clustering algorithm clustering effect

Data set	Algorithm	F			
		$F(1)$	$F(2)$	$F(3)$	F
Iris	SC	1	0.8545	0.8222	0.8922
	P-SC	1	0.8829	0.8315	0.9048
Balance	SC	0.2320	0.7929	0.6655	0.6902
	P-SC	0.1721	0.8058	0.6667	0.6920
Heart	SC	0.5094	0.5128		0.5109
	P-SC	0.5185	0.5111		0.5152

spectral clustering algorithm. (Iris $F_{p-sc} = 0.9048 > F_{sc} = 0.8922$; Balance: $F_{p-sc} = 0.6920 > F_{sc} = 0.6902$; Heart: $F_{p-sc} = 0.5152 > F_{sc} = 0.5109$)

24.5 Conclusion

The shortcoming of the traditional spectral clustering algorithm is its dependence on initial value. This paper presents an improved spectral clustering algorithm, a spectral clustering algorithm based on the particle swarm optimization. This method is not sensitive to the initial value, so getting the initial value is not sensitive to the spectral clustering algorithm. Example analysis shows that, compared to the traditional spectral clustering algorithm, the spectral clustering algorithm based on PSO has overcome dependence on the initial value of the spectral clustering algorithm. The clustering effect is hence better.

References

1. Cai, X., Dai, G., Yang, L.: Overview of spectral clustering algorithm. Comput. Sci. **35**(07), 14–18 (2008)
2. Zhang, X., Qian, X.: Immune spectral clustering algorithm for image segmentation. J. Softw. **21**(9), 2196–2205 (2010)
3. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of the 4th IEEE International Conference on Neural Networks, pp. 1942–1948. IEEE Service Center, Piscataway (1995)
4. Langdon, W.B., Poli, R.: Evolving problems to learn about particle swarm and other optimizers. Proc. CEC-2005. **1**, 81–88 (2005)
5. Eberhart, R, Shi, Y.: Particle swarm optimization: Developments, applications and resources . Proc. IEEE Congr. Evol. Comput. **1**(1), 81–86 (2001)
6. Eberhart, R., Shi, Y., Kennedy, J.: Swarm Intelligence. Morgan Kaufmann, San Mateo (2001)
7. Wang, W., Tang, Y.: Current situation and prospect of particle swarm optimization. J. Zhejiang Univ. Technol. **35**(2), 136–141 (2007)
8. Wang, L., Bo, L., Jiao, L.: Density-sensitive semi-supervised spectral clustering. J. Softw. **18**(10), 2412–2422 (2007)
9. Su, S., Wang, J., Fang, J.: Overview applications and research on particle swarm optimization algorithm. Comput. Technol. Dev. **17**(5), 248–250 (2007)

Chapter 25
A Framework for Porting Linux OS to a cc-NUMA Server Based on Loongson Processors

Kun Zhang, Hongyun Tian, Li Ruan, Limin Xiao, Yongnan Li and Yuhang Liu

Abstract In order to make the Linux operating system available on a cache coherence NUMA (cc-NUMA) server based on Loongson processors, a family of general-purpose MIPS64 CPUs developed by the Institute of Computing Technology in China, this paper proposes a framework for porting Linux operating system to this cc-NUMA server. Researchers present the overall port scheme after analyzing the framework of the Linux kernel and the architecture of the hardware platform, and then they discuss the transplantation in details with processor-level transplantation, memory management transplantation, interrupt and trap transplantation. The performance evaluation shows that the whole system works stable and the ported operating system could reach about 30 % of the theoretical peak value of floating-point calculation. The method could port Linux OS to the target board successfully and can be used on other platforms. The research has great significance to the development of the domestic Loongson processor and the cc-NUMA platform based on Loongson processors.

Keywords High performance computer · cc-NUMA · Loongson · Linux kernel

25.1 Introduction

Loongson is a family of general-purpose MIPS64 CPUs developed by the Institute of Computing Technology (ICT) in China. The Loongson-3B processor is an 8-coreprocessor with 1 GHz frequency [1]. Non-Uniform Memory Architecture

K. Zhang (✉) · H. Tian · L. Ruan · L. Xiao · Y. Li · Y. Liu
State Key Laboratory of Software Development Environment, Beihang University, Beijing, China
e-mail: zhangkun2441@126.com

K. Zhang · L. Ruan (✉)
School of Computer Science and Engineering, Beihang University, Beijing, China
e-mail: ruanli@buaa.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_25,

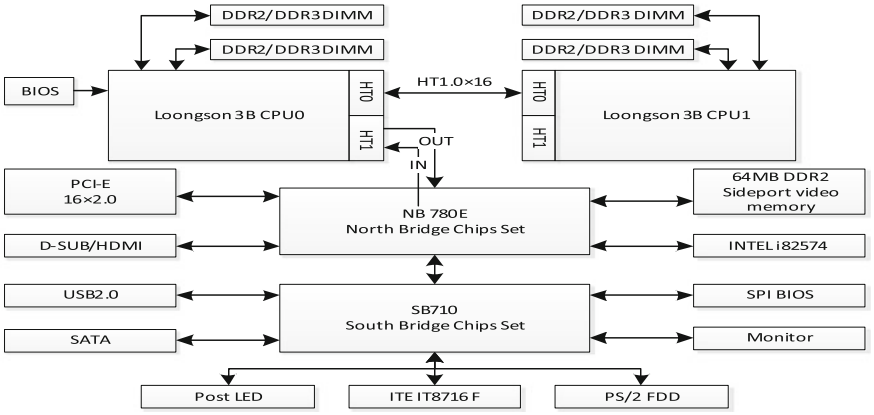


Fig. 25.1 The logic building-block of the cc-NUMA system based on Loongson CPU

(NUMA) has been more and more popular in the field of high performance computer as it has better scalability than Uniform Memory Architecture (UMA) [2]. The NUMA system has the feature that for any given region of physical memory, some processors are closer to it than the other processors.

cc-NUMA is a kind of NUMA system. cc-NUMA server based on Loongson processors is devised by the Institute of Computer Architecture of BeiHang University. Figure 25.1 shows the logic building-block view of the target platform. There are two nodes on the board with a processor per node. As depicted in Fig. 25.1, the CPU0 can access to its own memory faster than access to the memory on the other node. Frequently remote memory access would degrade the system performance seriously. So we need to avoid remote memory access in the porting scheme. CPU0 on the node0 is the boot CPU of the system and the other processors need to be initialized by it. Therefore, we need to solve this problem during the system initialization. Besides, the North Bridge chipset connects the peripheral component and South Bridge chipset with the boot CPU. Then we need to program the under-layer functions of the PCI device handler to make them work properly.

The rest of this paper is organized as follows. Section 25.2 analyzes the Linux kernel and puts forward an overall porting scheme. Section 25.3 discusses the transplantation in detail and we evaluate the modified kernel in Sect. 25.4. Some related work is introduced in Sect. 25.5 befor the conclusion in Sect. 25.6.

25.2 Linux Kernel Analysis and Overall Porting Scheme

Linux is one of the most widely ported operating system kernels. In this section, we first analyze the kernel and then put forward an overall porting scheme. As shown in Fig. 25.2, the Linux kernel lies between the user space and the under-layer hardware.

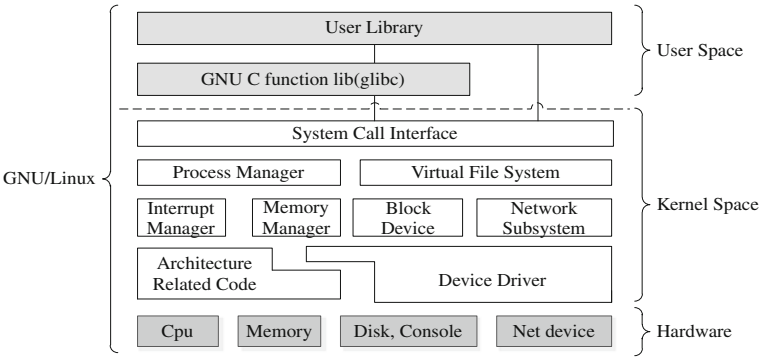


Fig. 25.2 A simplified view of Linux operating system

Kernel is the key part of the operating system with different modules responsible for monitoring and management of the entire computer system [3]. As presented in Fig. 25.2, the Linux operating system can be divided into three layers. Kernel layer lies under the system call interface and can be further divided into architecture independent module and architecture relevant module. Architecture independent module is common to all Linux support architectures. Architecture relevant module is called as Board Support Package (BSP), which is relevant to the hardware architecture and the processors [4].

The most important modules in the kernel are Process management module (PMM), memory management module (MMM), virtual file system (VFS) and network interface module (NIM) [5]. These modules cooperate with each other to make the kernel runs properly. PMM chooses the next-running process during the clock interrupt and it is also responsible for the load balance of the system [6]. As introduced in Sect. 25.1, it must be careful to choose the next-running process because of the problem of remote memory access. To make sure the scheduler works well on the cc-NUMA architecture, some information of the under-layer needs to be provided to the scheduler.

The MMM is responsible for deciding the region of memory that each process can use and determining what to do when not enough memory is available [7]. The memory management module can be logically divided into two parts, architecture-independent module and architecture-relevant module. The architecture-relevant part contains codes of memory initialization and some handler functions of memory management. The virtual file system provides a unified interface for all devices, and hides the specifics of the under-layer hardware. To perform useful functions, processes need to access to the peripherals connected to the North Bridge and South Bridge, which are controlled by the kernel through device drivers. The device driver module is device specific [8]. We don't need to rewrite the device driver code, as the AMD Inc. has already provided the driver for the peripheral devices.

We get the overall porting scheme according to our analysis to the kernel code as presented in Fig. 25.3. Details will be discussed in the next section.

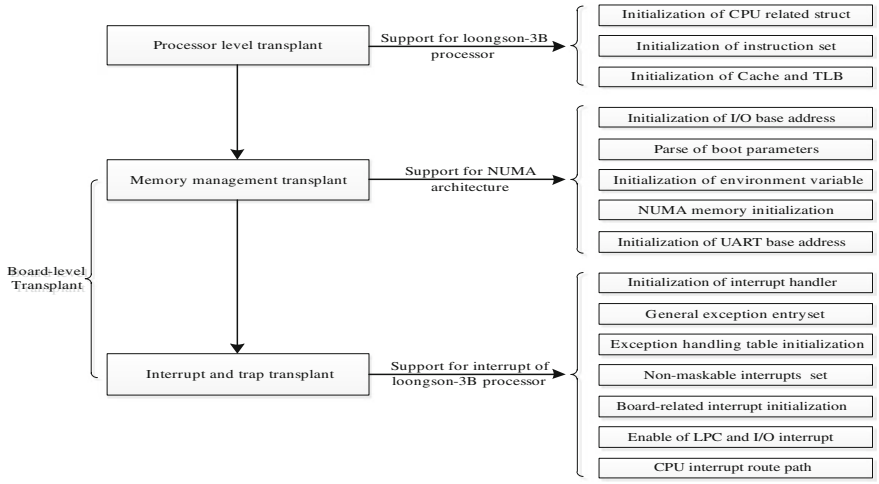


Fig. 25.3 The overall porting scheme

25.3 Kernel Porting

In order to port Linux to the Loongson cc-NUMA platform, we take the stable version of Linux 2.6.36 as our orignal edition. As schematically described in Fig. 25.3, we will concretely discuss the kernel porting in detail in this section.

25.3.1 Processor-Level Transplantation

Processor-level transplantation includes the initialization of CPU related structure, cache volumes and TLB volumes. Specifically, it contains the following steps:

1. Initialize the CPU related structure. Such as the processor_id of Loongson-3B, the machine type, the instruction cache, the data cache and the second cache.
2. Get the architecture type of the CPU according to the processor_id. Loongson-3B processor is based on MIPS 4KC.
3. Get the cpu_id, fpu_id and the type of the CPU based on the front two steps.
4. Check the virtual memory bits with EntryHi register to prepare for the memory management transplantation, Table 25.1.

25.3.2 Memory Management Transplantation

Memory is the key component of the system, process can't work without memory. The memory management transplantation mainly contains 5 parts.

Table 25.1 Example code of processor-level transplantation

```
#define PRID_IMP_LOONGSON3A        0x6305
#define PRID_IMP_LOONGSON3B        0x6306 //definition of the processor_id
#define enable_fpu()               \
do {                               \
    if (cpu_has_fpu)               \
        __enable_fpu();            \
} while (0);
```

1. Initialize the base address of I/O space. The Loongson-3B processor unified the whole physical space. As a result, the I/O memory address is a part of them.
2. Parse the boot command. The default command is "console = tty, 115200, root = /dev/sda1". We parse the boot command to determine the frequency of the console and the path of the root file system. In addition, some other environment variables can be passed by the boot command.
3. Parse the environment variables. These parameters are transferred by the bootloader according to hardware registers, including the frequency of bus clock, cpu clock and the size of the memory, high memory size et al.
4. Initialize memory subsystem supporting for NUMA architecture. As we introduced in Sect. 25.1, it needs to avoid remote memory access as much as possible to be efficiently. We distinguish the memory between NUMA nodes with the memory size of each NUMA node. Each NUMA node has its own memory and the kernel can be conscious about the system memory.
5. Initialize the UART base address. The UART port is very important for the system debug as it can print out the debug information for the developer. The UART base address varies between the different processors, which should be set according to the datasheet.

25.3.3 Interrupt and Trap Transplantation

Interrupt subsystem is the essential composition of multithread system. This part can be divided into three parts as follows.

1. Initialize the system interrupts. First, the kernel gets the active cpu_id from the CP0_coprocessor, then it sets the interrupt registers to mask all the interrupt flags and clear the interrupts hanged up. Then the kernel enables the LPC and I/O interrupt controller and other interrupt related registers.
2. Map the interrupt handler to the irq number. We need to rewrite the hardware related interrupt handlers such as shown in Table 25.2.

Table 25.2 Example code of interrupt and trap transplantation

```
void __init mach_init_irq(void)
{
        lpc_interrupt_route(); //Route the LPC interrupt to Core0 INT0
        ht_interrupt_route(); // Route the HT interrupt to Core0 INT1
        mips_cpu_irq_init()// Route the cpu related interrupt to Core0
        init_i8259_irqs(); // Route the serial interrupt on the south bridge to Core0
}
```

3. Set the trap base address and initialize the exception handling table. Some under layer related handlers need to be rewrite.

25.4 Evaluation

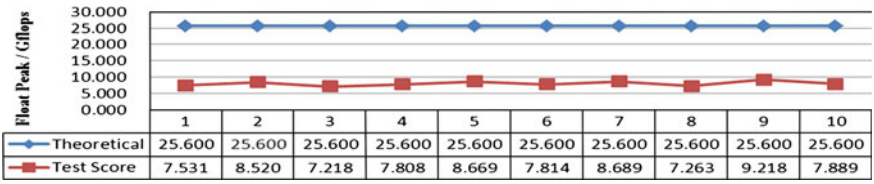
In this section, we evaluate our ported system with *stressapptest 1.0.4* to test the memory subsystem. Besides, we use *linpack* to test the peak value of floating-point calculation to evalute the system performance. The *linpack* test case is "mpiexec –np 8./xhpl" with HUGE_TLB configured in kernel. The test score of *stressapptest* showed in Fig. 25.4 demonstrates that the memory subsystem works stable. While the test score of *linpack* in Fig. 25.5 shows that our ported system can reach about 30 % of the theoretical peak value of floating-point calculation.

25.5 Related Work

It has been a long time of work to port Linux OS to other platforms such as ARM, MIPS. Hu Jie [9] has transplanted Linux OS to the ARM platform. As of Loongson platform, Cheng xiao-yu [10, 11] has transplanted the μC/OS to the Loongson based platform, and Qian Zheng-jiang [12] discussed about the development of Linux distribution on Loongson platform. Besides, the ICT has made a lot of work

```
~/stressapptest-1.0.4_autoconf/src$ ./stressapptest -M 1024
Log: root @ loongson_debian on Wed Sep 1 12:38:35 PDT 2012 from open source release
Log: 2 nodes, 2 cpus.
Stats: Found 0 hardware incidents
Stats: Completed: 4.00M in 1.29s 3.09MB/s, with 0 hardware incidents, 0 errors
Stats: Memory Copy: 4.00M at 3.90MB/s
```

Fig. 25.4 Evaluation with *stressapptest 1.0.4*



	1	2	3	4	5	6	7	8	9	10
Theoretical	25.600	25.600	25.600	25.600	25.600	25.600	25.600	25.600	25.600	25.600
Test Score	7.531	8.520	7.218	7.808	8.669	7.814	8.689	7.263	9.218	7.889

Fig. 25.5 Evaluation with *linpack*

on the Loongson-3A processor. Our research can make up for the current situation about lacking of research to the Loongson-3B platform.

25.6 Conclusion

In this paper, Researchers analyzed the Linux kernel architecture and the hardware platform based on Loongson processor. Besides, they discussed porting Linux to Loongson NUMA platform concretely. Processor-level transplantation, memory management transplantation and interrupt related transplantation were introduced in detail. The evaluate score shows that the ported system runs smoothly on the Loongson platform. The research provides an example of porting Linux to Loongson platforms and can be easily used on other platforms. The research has great significance to the development of domestic Loongson processor and the cc-NUMA platform based on Loongson processors.

Acknowledgments Our research is sponsored by the Hi-tech Research and Development Program of China (863 Program) under Grant NO.2011AA01A205, the National Natural Science Foundation of China under Grant NO.61232009, the Doctoral Fund of Ministry of Education of China under Grant NO.20101102110018, the Beijing Natural Science Foundation under Grant NO.4122042, the fund of the State Key Laboratory of Software Development Environment under Grant NO.SKLSDE-2012ZX-07.

References

1. Weiwu, H., Ru, W., Baoxia, F., et al.: Physical implementation of the eight core Loongson-3B microprocessor. J. Comput. Sci. Technol. **26**(3), 520–527 (2011)
2. Bolosky, W., Fitzgerald, R.: The development of computer architecture in HPC. In: Proceedings of the 19th International Conference on Parallel Architectures, pp. 557–561
3. Sanjeev, K.: Reliability estimation and analysis of linux kernel. Int. J. Comput. Sci. Technol. **12**(2), 529–533 (2011)
4. Bowman, T.: Linux as a case study: its extracted software architecture [EB/OL]. http://pIg.uwater-Ioo. ca/∼ itbowman/papers
5. Jones, T.: Inside the Linux 2.6 Completely Fair Scheduler. [EB/OL]. http://www.ibm.com/deve-loperworks/linux/library/l-completely-fair-scheduler
6. Galvin, S.: Operating System Concepts, 4th edn. pp. 458–460 (1994)

7. Eranian, S.: Virtual Memory in the MIPS Linux Kernel, pp. 320–331. Prentice Hall PTR, Upper Saddle River (2005)
8. Choi, J., Baek, S., Shin, S.Y.: Design and implementation of a kernel resource protector for robustness of Linux module programming. In: Proceedings of the 2006 ACM Symposium on Applied Computing, pp. 1477–1481 (2006)
9. Jie, H., Genbao, Z.: Research transplanting method of embedded Linux kernel based on ARM platform. International Conference of Information Science and Management Engineering. Xi'an, China, pp. 424–432 (2010)
10. Qian, Z., Fujian, W., Boyan, L.: Transplant Method and Research of μC/OS_II on Loongson paltform. 2011 Fourth International Conference on Intelligent Computation Technology and Automation, pp. 291–301. Xi'an, China (2011)
11. Xiao yu, C., Duyan, B., Ye, C et al.: Transplantation of μC/OS on loongson processor and its performance analysis. Comput. Eng. **05**(02), 372–379 (2009)
12. Zhengjiang, Q., Jin yi, C.: Development of Linux release based on Loongson mipsel architecture. J. Chang Shu Inst. Technol. **22**(10), 87–91 (2008)

Chapter 26
Optimization for the Locations of B2C E-Commerce Distribution Network Based on an Improved Genetic Algorithm

Guanshi Li and Dong Wang

Abstract To solve the problems of high cost and unreasonable location layout that the self-built logistics system of a B2C E-commerce company has, a B2C logistics distribution optimization solution is proposed. The solution established a mathematical model based on the construction costs of the regional distribution centers, the operation costs of the whole logistics distribution network, the transportation costs of the whole supply chain and the penalty cost for the situations of delivery overtime. The model considers the factors of economies of scale and standard service level. As the model has NP-Hard complexity, a mixed genetic and simulated annealing algorithm is proposed to solve the problem. And at last, a case of a B2C e-commerce company verifies the correctness of the whole theory.

Keywords B2C e-commerce · Distribution center location · Improved genetic algorithm

26.1 Introduction

In recent years, the E-commerce market of China experiences a rapid development (shown in Fig. 26.1). The online shopping market in China reached 756.6 billion and had a year-on-year growth of 44.6 % in 2011. And the number of E-commerce user had reached 194 million with an annual consumption of 3,901 per capita. The B2C E-commerce in China had shared the same experience. The total B2C market transactions are 240.07 billion and the year-on-year growth is 130.8 % in 2011. At the same time, the competition in B2C market is very fierce. The traditional monopoly situation of the IT industry has not formed yet. More and more

G. Li (✉) · D. Wang
Logistics Information Technology and RFID Laboratory, Shanghai Jiao Tong University, Shanghai, China
e-mail: liguanshi0706@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_26,

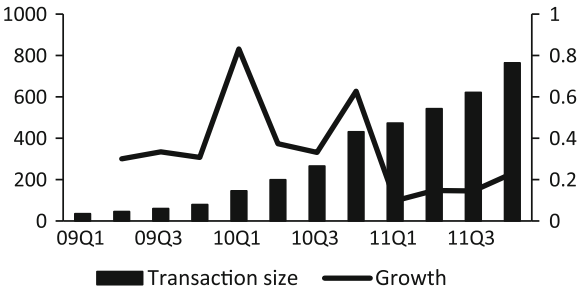


Fig. 26.1 The B2C E-commerce development in China

companies have realized the importance of logistics services as it is the key factor in the fierce competitive market. Through the reasonable development of logistics distribution system, these enterprises can expand their market share and find the best balance between the costs and their services. The Fig. 26.1 shows the development of B2C E-Commerce in recent years.

In recent years, a lot of experts and scholars have studied the sector of E-commerce logistics services. These researches mainly concentrate in these areas: the research of customer loyalty based on logistics services [1–3], E-commerce reverse logistics [4] and E-commerce urban distribution [5, 6].

This paper is focused on the study of E-commerce logistics distribution network design and optimization. It establishes a mathematics model to simulate the actual situation and designs an improved genetic algorithm. An actual case of logistics distribution network optimization is used to demonstrate the correctness of the whole theory.

26.2 B2C Distribution Network Optimization

26.2.1 Current Situation

The B2C self-built logistics systems in China don't have a long development history. The problems of low degree of specialization and incomplete network of logistics services are obvious. As a result the cost of the self-built logistics system is high and the layout of the logistics distribution system is unreasonable. Therefore, most of the B2C companies cannot afford efficient logistics services within low cost.

26.2.2 The Design of B2C Distribution Network

In order to minimize the total cost of the logistics distribution network while guarantee or improve the current customer service level, it is necessary for the

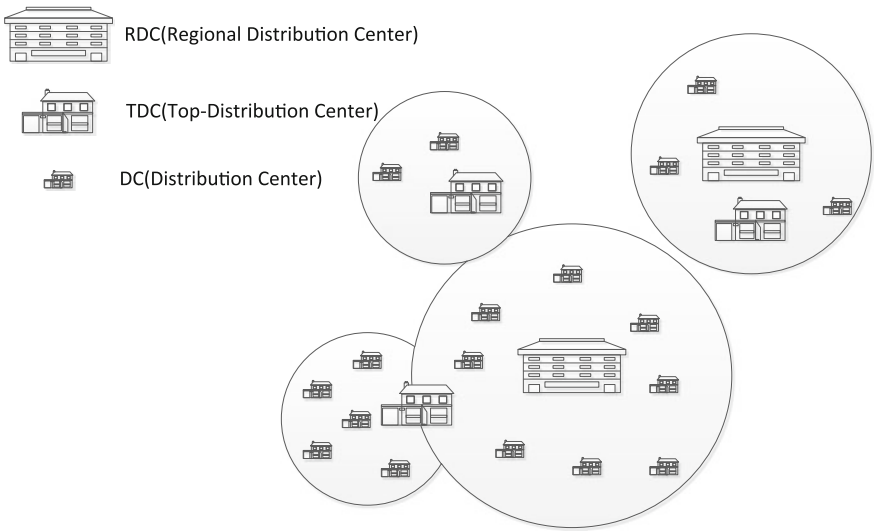


Fig. 26.2 The B2C distribution network

Regional Distribution Center (RDC) to aggregate the B2C online order. And the RDC should coverage all the requirements in his service region. There are a lot of researches in this area [7, 8], but most of them are based on random demand and only consider the transportation costs of the logistics distribution network. It is not appropriate to use them in the actual situation as they also ignore the factors of types of goods and scale of economies. The Fig. 26.2 shows the self-built distribution network of B2C E-Commerce.

In order to reduce the logistics costs while improve the logistics service, the paper designs the architecture that describes the logistics flows of B2C orders from RDC to the DC. The designed logistics distribution network has three different functional levels. They are RDC, Top-Distribution Center (TDC) and Distribution Center (DC). The RDCs are in charge of storage, sorting and regional distribution. The TDCs are in charge of shipment transiting. The DCs are in charge of the urban distribution. The Fig. 26.3 shows the distribution architecture of B2C E-Commerce.

26.2.3 OptimizationModel

There are four parts of costs that should be considered to establish that network: the construction costs of RDCs, the operational costs of the RDCs, the costs of regional transportation and the total penalty costs of the whole system.

The model is established based on the below assumptions: 1. Each TCD only connect with one RDC; 2. The RDCs have minimum limitation of the order disposal amount and total working area; 3. The RDCs can only be selected from its

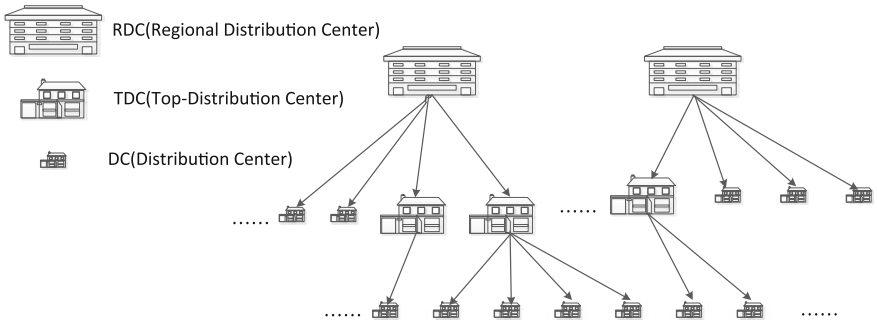


Fig. 26.3 The B2C distribution architecture

candidates; 4. Customers will turn to other B2C companies at a certain probability when the company couldn't deliver their orders on time; 5. The working areas of RDC are leased.

The mathematics model is based on the annual order amount of RDC. The way RDC aggregates its order is:

$$R_m = \sum_{n=1}^{N} \omega_{mn} R_n + r_m \tag{26.1}$$

N is the set of TDC candidates; R_m is the annual order amount of RDC m and R_n is the annual order amount of TDC n; r_m is the annual order amount of RDC m before aggregating; ω_{mn} is the 0–1 indicator whether RDC m and TDC n is connected.

The working area of RDC can be classified as the storage area, the sorting area and the outbound area. The equations to estimate them are:

$$S_{m1} = \sum_{i=1}^{I} \frac{h_i x_i y_i R_m}{v_i z_i} \tag{26.2}$$

$$S_{m2} = \frac{R_m}{\pi} \tag{26.3}$$

$$S_{m3} = \frac{R_m}{\tau} \tag{26.4}$$

$$S_m = \frac{S_{m1} + S_{m2} + S_{m3}}{\varphi} = \frac{\sum_{i=1}^{I} \frac{h_i x_i y_i R_m}{v_i z_i} + \frac{R_m}{\pi} + \frac{R_m}{\tau}}{\varphi} \tag{26.5}$$

S_m is the total working area of RDC m; S_{m1} is the storage area of RDC m; S_{m2} is the sorting area of RDC m; S_{m3} is the outbound area of RDC m; x_i is the average storage area for each i-th order; y_i is the average floor space of the shelves that storage the commodities of the i-th order; v_i is the storage redundancy coefficient of the commodities of the i-th order; z_i is the average area of shelves that storage

the commodities of the i-th order; h_i is the proportion of the i-th order that RDC handled; π is the average disposal amount of RDC's sorting area; τ is the average disposal amount of RDC's outbound area; φ is the storage redundancy coefficient of RDC.

The construction costs of all the RDCs are:

$$W = \sum_{m=1}^{M} \delta_m (C_{m1} S_m + \sum\nolimits_{i=1}^{I} \frac{h_i x_i R_m C_{m2i}}{v_i z_i} + C_{m3} \frac{R_m}{\pi} + d_m) \tag{26.6}$$

W is the construction costs of all the RDCs; δ_m is the 0–1 indicator whether candidate m is selected as a RDC; C_{m1} is the average land lease cost of RDC m; C_{m2i} is the average cost of shelves that storage the i-th order at RDC m; C_{m3} is the average construction cost of the sorting area of RDC m; d_m is the fixed cost of RDC m.

The operational costs of all the RDCs are:

$$U = \sum_{m=1}^{M} \delta_m (k_1 S_{m1} + k_2 S_{m2} + k_3 S_{m3} + k_4 S_m) \tag{26.7}$$

U is the operating cost of all the RDC; k_1 is the average operating cost of the storage area of RDC; k_2 is the average operating cost of the sorting area of RDC; k_3 is the average operating cost of the outbound area of RDC; k_4 is the average management cost of RDC.

The costs of regional transportation are:

$$T = \Sigma_{m=1}^{M} \Sigma_{n=1}^{N} \Sigma_{i=1}^{I} \delta_m \omega_{mn} h_i R_n \mu_i D_{mn} \tag{26.8}$$

T is the cost of regional transportation; D_{mn} is the travel distance between RDC m and TDC n; μ_i is the average transportations cost of the i-th order.

The total penalty cost of the whole system is:

$$P = \Sigma_{m=1}^{M} \Sigma_{n=1}^{N} \delta_m \omega_{mn} \theta_n \rho p R_n \tag{26.9}$$

P is the total penalty cost of the whole system; p is the average price of the order; θ_n is the 0–1 indicator whether TDC's services are overtime; ρ is the average percentage of customer lost when services are overtime.

Based on the above theory, the objective function of the whole model is:

$$MinZ = W + U + T + P \tag{26.10}$$

26.3 Improved Genetic Algorithm

The nonlinear mixed-integer programming model the paper established is a typical NP-Hard problem, which cannot be resolved by a routine method in a reasonable time. It is necessary to adopt a heuristic algorithm to solve the model.

This paper use the a new method combined the genetic algorithm [9–11] with the simulated annealing algorithm [12–14]. As GA is one of the heuristic algorithms that is available in network optimization and the simulated annealing algorithm is used to improve the phenomenon of premature convergence of GA.

26.3.1 The Framework of the Algorithm

The concrete steps of the improved algorithm proposed by this paper are (show):

(1) The initialization process. It is helpful to import a good set of group in the algorithm initialization process. It is reasonable to use an initial group that is generated by the random function in the condition of lacking the specific information where a good feasible solution would be. Initialize the value of Population (i) where i = 0, the size of the population *Size*, the initial temperature, the termination temperature, the temperature schedule *S* and the max number of iterations in each temperature.
(2) Calculate the objective value *Z(J)* of each chromosome in the population. Take the objective value of each chromosome as its fitness value. Search the chromosome *J'* whose fitness value is the minimum in the population. Record its information.

The improved genetic crossover operation is: Select two chromosomes of the parent group based on the genetic selection criterion and generate two new chromosomes through the regular genetic crossover operation. Use the Metropolis Criterion to identify whether replace the original chromosome with the new chromosome.

The improved genetic mutation operation is: Select one chromosome of the population based on the genetic selection criterion and applies the regular genetic mutation operation to get one new chromosome. Use the Metropolis Criterion to identify whether replace the original chromosome with the new chromosome.

(3) Sort the chromosomes in the population by their fitness value. Apply the genetic operations on the entire population while using the Metropolis Criterion of Simulated Annealing Algorithm to identify whether accepts the newly generated chromosome. If the newly generated chromosome is not accepted, the original chromosome will still be used at the next iteration.
(4) Identify whether the number of iterations reaches the max iteration number at current temperature. If the value has reached that number, then the temperature goes down to next level. If the temperature reaches the termination temperature, then end the algorithm. Calculate the objective value of the whole population and find out the chromosome with the minimum value *J*. Export that result. Otherwise algorithm jumps to step (26.2).

The Fig. 26.4 is an algorithm frame of the above algorithm.

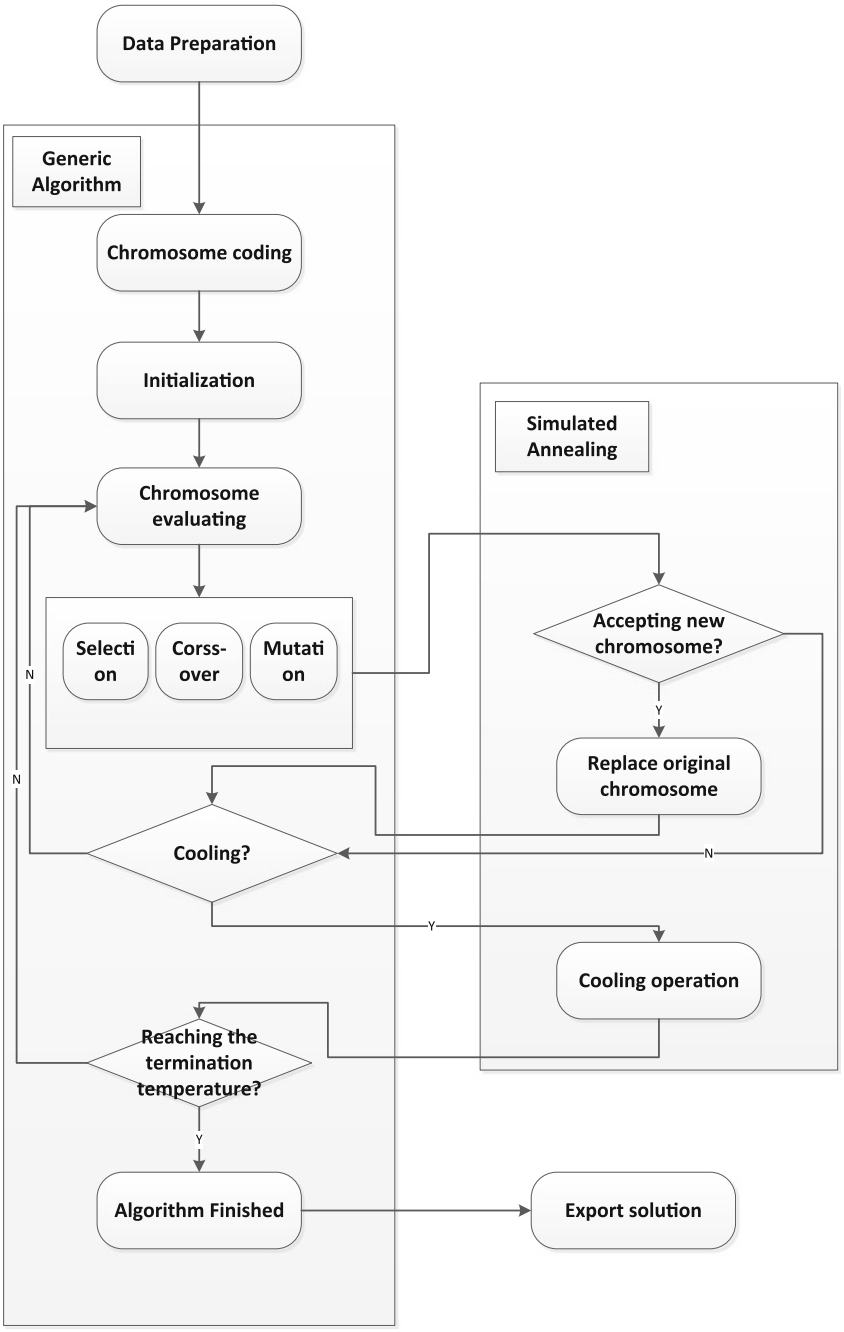


Fig. 26.4 The B2C distribution architecture

26.3.2 Unification Coding of Chromosome

The chromosome unification coding mechanism is:

$$[R_1, R_2, \ldots R_m / RDC(0,1) | X_1, X_2, \ldots X_n / RDCindex] \tag{26.11}$$

The concrete description of the coding: The first part shows the selection of the RDCs, the coding example is that one represents the corresponding candidate is chosen as a RDC. The second part shows the connection information between RDCs and TDCs, the coding example is that two represents the corresponding candidate is connected with the RDC which is coded two.

26.4 Example and Result Analysis

The above model and algorithm were applied to construct an actual B2C distribution network of a leading B2C E-Commerce company in China. The network has to be established form 10 RDC candidates and 45 TDC candidates. All the data the algorithm needed such as the distance matrix between logistics facilities candidates, the operating and fixed costs of logistics facilities candidates, the commodity category list and the history online transaction data, are provided by that corporation and the national statistical authorities. The algorithm is implemented in Java and the coding environment is Linux.

Through the application of the hybrid Genetic Algorithm and Simulated Annealing Algorithm, the optimization solution of the distribution network is concluded,and it is shown in the Tables 26.1 and 26.2. In order to minimize the total costs of the distribution network while improve the customer services, corporation should establish its RDCs in Beijing, Shanghai, Wuhan and Guangzhou. The service scope of each RDC is shown in Table 26.2.

The comparisons between the improved Genetic Algorithm and the typical Genetic Algorithm are also shown in the Table 26.1. The results are different. Compare to the typical Genetic Algorithm, the total cost of the whole distribution

Table 26.1 Optimized results and comparisons

Optimal solution	RDC amount	Iterations	RDC	Total orders	Total cost
After our algorithm	4	1000	Beijing	5519167	47970339.82
			Shanghai	6471813	
			Wuhan	4415770	
			Guangzhou	5300895	
After typical GA	4	1500	Beijing	5937628	49693427.49
			Nanjing	6149251	
			Wuhan	3974634	
			Shenzhen	5646132	

Table 26.2 RDC service scope

RDC	RDC scope
Beijing	Tianjin, Jinan, Yantai, Shijiazhuang, Qingdao, Weihai, Cangzhou, Weifang, Qinhuangdao, Chengde, Taiyuan, Linfen, Datong, Hohhot, Baotou
Shanghai	Nanjing, Hangzhou, Suzhou, Ningbo, Xiamen, Fuzhou, Hefei, Wuxi, Wenzhou, Xuzhou, Changzhou, Quanzhou, Yangzhou, Nantong, Linyi
Wuhan	Nanchang, Changsha, Changde, Zhengzhou, Luoyang, Kaifeng, Jiujiang, Zhuzhou, YiChang
Guangzhou	Nanning, Guilin, Liuzhou, Qinzhou, Haikou, Dongguan, Huizhou, Maoming, Zhanjiang, Zhuhai,Shantou, Shenzhen

network has been reduced by 3 %, although the iteration time of the improved Genetic Algorithm is less than the iteration time of the typical Genetic Algorithm. It is very obvious that the improved algorithm is more efficient than the typical Genetic Algorithm.

26.5 Conclusion

The B2C logistics distribution network contains many complex factors. It has become the core competency of a B2C E-commerce company. This paper establishes a mathematical optimization model based on the characteristics of B2C logistics distribution network and proposes an improved genetic algorithm to solve the problem. The algorithm has improved the typical genetic algorithm in efficiency and optimization results. And at last, the whole optimization solution is verified with an actual B2C logistics distribution example.

References

1. China Internet Network Information Center. The report of China's online shopping market research [EB/OL]. http://www.cnnic.net.cn/(2011)
2. Ramanathan, R.: The moderating roles of risk and efficiency on the relationship between logistics performance and customer loyalty in e-commerce. Transp. Res. Part E 950–962 (2010)
3. Lu, Y.: From virtual community members to C2C e-commerce buyers: Trust in virtual communities and its effect on consumers' purchase intention. Electron. Commer. Res. Appl. **9**, 346–360 (2010)
4. Thirumalai, S.: Customer satisfaction with order fulfillment in retail supply chains: implications of product type in electronic B2C transactions. J. Oper. Manage. **23**, 291–303 (2005)
5. Wang, Z.: A new location-inventory policy with reverse logistics applied to B2C e-markets of China. Int. J. Prod. Econ. **107**, 350–363 (2007)
6. Du, T.C.: Dynamic vehicle routing for online B2C delivery. Int. J. Manage. Sci. **33**, 33–45(2005)

7. Sterle, C.: Location-routing models and methods for freight distribution and infomobility in city logistics. UniversitadegliStudi di Napoli Federico II (2009)
8. Kiya, F.: Stochastic programming approach to re-designing a warehouse network under uncertainty. Transp. Res. Part E **48**, 919–936 (2012)
9. Lau, H.C.W.: A credibility-based fuzzy location model with Hurwicz criteria for the design of distribution systems in B2C e-commerce. Comp. Ind. Eng. **59**, 873–886 (2010)
10. Fei, F.: Study and application on the pivotal optimization techniques of the storage network in finished vehicle logistics. Shanghai Jiao Tong University (2009)
11. Taniguchi, E.: Optimal size and location planning of public logistics terminals. Transp. Res. Part E **35**, 207–222 (1999)
12. Leite, J.P.B.: Improved genetic operators for structural engineering optimization. Adv. Eng. Softw. **29**(7–9), 529–562 (1998)
13. Dong, M.: Shortest path based simulated annealing algorithm for dynamic facility layout problem under dynamic business environment. Expert Syst. Appl. **36**, 11221–11232 (2009)
14. Sahin, R.: A simulated annealing algorithm for solving the bi-objective facility layout problem. Expert Syst. Appl. **38**, 4460–4465 (2011)

Chapter 27
An Improved Multi-Objective Differential Evolution Algorithm with an Adaptive Crossover Rate

Huifeng Zhang, Jianzhong Zhou, Na Fang, Rui Zhang and Yongchuan Zhang

Abstract In order to properly solve multi-objective optimization problem, an improved multi-objective differential evolutionary algorithm with an adaptive crossover rate is proposed in this paper. To adjust the evolution adaptively, an adaptive crossover rate is integrated into the differential evolution. The new strategy can diverse pareto individuals and further to pareto front, which avoids the local convergence that traditional differential evolution always trapped in. In addition, combining with great ability of searching local optima of differential evolution, evolutionary speed and diversity can be simultaneously improved by the modified crossing operator. The simulation on these benchmark problems verifies the efficiency of the proposed algorithm with convergence metric and diversity metric, and the obtained results also reveal that the proposed method can be a promising alternative in solving multi-objective optimization problems.

Keywords Multi-objective optimization · Differential evolutionary algorithm · Pareto individuals · Pareto front

27.1 Introduction

In comparison to single objective optimization problems, multi-objective problems (MOPs) have series of solutions but not a unique solution, since the real requirements in application are unknown before making decisions. In particular, MOPs have several desirable characteristics: (i) conflicting objectives; (ii) intractably large and highly complex search space. In recent years, a number of different multi-objective evolutionary algorithms (MOEAs) have been proposed, such as Fonceca and Fleming's MOGA [1], Srinivas and Deb's NSGA [2], and

H. Zhang · J. Zhou (✉) · N. Fang · R. Zhang · Y. Zhang
Huazhong University of Science and Technology, Wuhan 430074, China
e-mail: prof.zhou.hust@263.net

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_27,

Horn's NPGA [3], etc. These MOEAs present the necessary additional operator to extend simple EA from solving single objective problem to a multi-objective problem, and fortunately they have been proved efficient and desirable in dealing with the MOPs due to their population based methodology and independence character of problem presentation.

Differential evolution (DE) [4], which is proposed by Price and Storn in 1997, it is a simple but powerful evolutionary algorithm with fewer parameters in comparison to EA. After it has been successfully used in solving single-objective optimization problems, some researchers have extended DE to solve some MOPs and obtained some success. Recently, DE has been developed to different styles in solving MOPs, such as Pareto differential evolution (PDE) [5], Pareto differential evolution approach (PDEA) [6], and adaptive differential evolution algorithm (ADEA) [7], these MODEs have been widely used in application and obtained satisfactory results. However, all these DEs based on multi-objective problems suffer from premature convergence at different degree as it does in single-objective optimization problem. In this paper, an adaptive evolution mechanism in differential evolution algorithm is proposed, and the modified crossover has been presented to diverse population of each generation as the differential evolution proceeds especially when it converges to local optima, which greatly improves the variety of archive population to avoid the premature convergence in solving different MOPs. The proposed MODE is also utilized on the benchmark problems with convergence metric and diversity metric, and the obtained results reveal that the improved MODE can be a promising method for solving MOPs.

The paper is organized as follows: some basic information about MOPs and the dominance relationship is introduced in Sect. 27.2, and in Sect. 27.3 we briefly describe the improved evolution mechanism, ultimately the numerical simulation further proves the effectiveness in Sect. 27.4, and outlines conclusion in Sect. 27.5.

27.2 Improved MODE

The MODE is similar to multi-objective evolutionary algorithm, it also has evolution strategy: selection, crossing and mutation, and dominance mechanism is also included in the MODE. Since the differential evolution strategy plays important role in the MODE, it is generally generated from three distinct individuals x_{r1}, x_{r2}, x_{r3}. randomly selected in the initial population, and the evolutionary operator is mathematically described in literature [8, 9].

27.2.1 Improved Crossover Rate

Since population diversity will decrease as differential evolution proceeds, crossover rate should be adjusted adaptively. In DE, the crossover rate is set as a

constant, which adjusts the population evolution at certain rate. When differential evolution falls into local area, and evolutionary will suffer premature problem.

$$p_c = \begin{cases} 4 * p * Minpc * (\dfrac{gen}{Maxgen}) - p * Minpc, & if\ gen > \dfrac{Maxgen}{2} \\ p * Minpc, & if\ gen < \dfrac{MaxGen}{2} \end{cases}$$

where p is predefined parameter, *Minpc* denotes the defined minimum crossover rate, *gen* represents the current generation number, *Maxgen* is the maximum generation number. The improved crossover rate can increase along with the population evolution, which means that the population diversity will increase especially when the search process falls into local area.

The pseudo code of improved evolution operator can be described as follows: Fig. 27.1.

27.2.2 Description of Improved MODE Algorithm

In comparison to other MOEAs, the proposed evolutionary strategy can adjust the population evolution well due to its adaptive crossover rate integrated into the crossover operator.

The overall flowchart of the proposed differential evolutionary algorithm is presented in Fig. 27.2:

27.3 Simulation

In this part, the proposed algorithm is compared against MODE and NSGA-II on some test functions in the experiments [10]. To evaluate these two goals, we use two widely used metrics: convergence metric γ and diversity metric Δ [11], which are proposed by Deb in 2002.

Fig. 27.1 The pseudo code of improved crossover operation

```
gen = 0;
p_c = p;
While (gen < MaxGen)
Begin
  If (gen < MaxGen / 2)
    p_c = p_c * Minpc
  else
    p_c = 4 * p_c * Minpc * (gen / MaxGen) - p_c * Minpc
  gen = gen + 1;
End
```

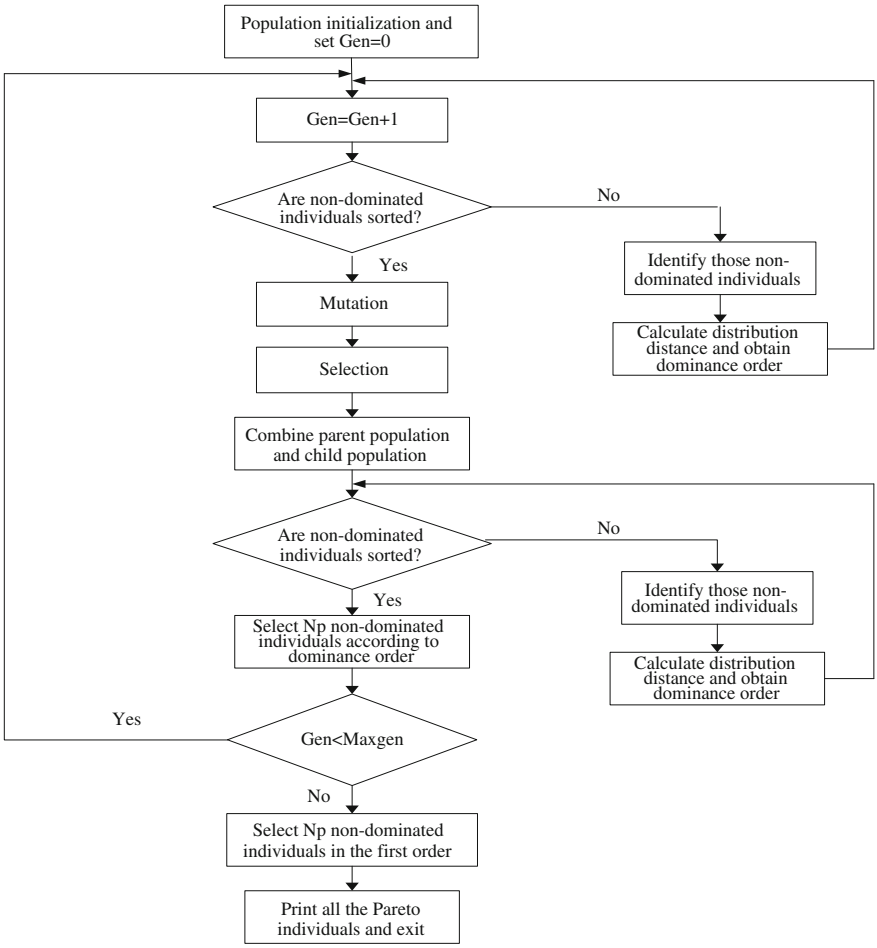


Fig. 27.2 The flowchart of the improved multi-objective differential evolution

In order to compare with other algorithms presented in previous years, the population size N_p is set to 100, the algorithm is run for 250 generations, and the size of archive set is set to 100.

In addition, to compare with other evolutionary algorithm, such as NSGA, SPEA2, PDEA, DEMO and ADEA, Tables 27.1 and 27.2 show convergence property and diversity distribution of obtained pareto front by those alternatives, it is clearly seen that the proposed differential evolutionary algorithm has better convergence and more uniformly distributed diversity than other methods. Furthermore, the variance of each problem is smaller, which also means that it can perform more stable than other methods.

Table 27.1 Convergence metric γ

Problem	ZDT1	ZDT3	ZDT4	ZDT6
NSGA-II	0.033482	0.1145	0.513053	0.296564
	0.00475	0.00794	0.11846	0.013135
SPEA2	0.023285	0.018409	4.9271	0.232551
	0	0	2.703	0.004945
PMODE	0.0058	0.02156	0.63895	0.02623
	0	0	0.5002	0.000861
DEMO/	0.001083	0.001178	0.001037	0.000629
(parent)	0.000113	0.000059	0.000134	0.000044
ADEA	0.002741	0.002741	0.1001	0.000624
	0.000385	0.00012	0.4462	0.00006
Improved MODE	0.000262	0.000535	0.007659	0.000294
	0.000117	0.000016	0.000134	0.000065

Table 27.2 Diversity metric Δ

Problem	ZDT1	ZDT3	ZDT4	ZDT6
NSGA-II	0.390307	0.73854	0.702612	0.668025
	0.001876	0.019706	0.064648	0.009923
SPEA2	0.154723	0.4691	0.8239	1.04422
	0.000874	0.005265	0.002883	0.158106
PDEA	0.298576	0.623812	0.840852	0.473074
	0.000742	0.000225	0.035741	0.021721
DEMO/	0.325237	0.309436	0.359905	0.442308
(parent)	0.030249	0.018603	0.037672	0.021721
ADEA	0.38289	0.52577	0.4363	0.3611
	0.001435	0.04303	0.11	0.0361
Improved MODE	0.236789	0.089484	0.256107	0.095743
	0.000178	0.001023	0.001278	0.004725

27.4 Conclusion

This study presents an improved evolutionary strategy to avoid local optimal by using a new crossover rate. In comparison to the traditional crossing operator, the improved crossover rate can adjust the population evolution adaptively to enlarge the search scale when differential evolution falls into local area. The efficiency of this improved MODE is evaluated by standard performance of some test problems, it is seen from the results that the improved MODE has stable performance in both convergence and diversity.

Acknowledgments This work is granted by the Public Welfare Industry of the Ministry of Water Resources (No.201001080), the special research fund for high school doctoral program (NO.20100142110012), and National Natural Science Foundation of China (NO.51107047).

References

1. Fonseca, C.M., Fleming, P.J.: Genetic algorithms for multi-objective optimization: Formulation, discussion and generalization. In: Forrest, S. (ed.) Proceedings of the Fifth International Conference on Genetic Algorithms, University of Illinois at Urbana-Champaign, pp. 416–423. Morgan Kauffman Publishers, San Mateo (1993)
2. Srinivas, N., Deb, K.: Multi-objective optimization using non-dominated sorting in genetic algorithms. Evol. Comput. **2**(3), 221–248 (1994)
3. Horn, J., Nafpliotis, N., Goldberg, D.E.: A niched Pareto genetic algorithm for multi-objective optimization. In: Proceedings of the First IEEE Conference on Evolutionary Computation, IEEE World Congress on Computational Intelligence vol. 1, pp. 82–87. Piscataway, New Jersey (1994)
4. Storn, R., Price, K.: Differential evolution—a simple and efficient Heuristic for global optimization over continuous spaces. J. Glob. Optim. **11**, 341–359 (1997)
5. Abbass, H.A., Sarker, R., Newton, C.: PDE. A Pareto-frontier differential evolution approach for mufti-objective optimization problems. In: Proceedings of the congress on evolutionary computation 2001 (CEC'2001), vol. 2, pp. 971–978. IEEE Service Center, Piscataway (2001)
6. Madavan, N.K.: Multi-objective optimization using a Pareto differential evolution approach. In: Proceeding of the Congress on Evolutionary Computation (CEC' 2002), vol. 2, pp. 1145–1150. IEEE Service Center, Piscataway (2002)
7. Qian, W., Li, A.J.: Adaptive differential evolution algorithm for mufti-objective optimization problems. Appl. Math. Comput. **201**(1–2), 431–440 (2008)
8. Storn, R., Price, K.: Differential evolution-a simple and efficient adaptive scheme for global optimization over continuous spaces. Technical report TR-95-012 1995. International Computer Science Institute, Berkeley (1995)
9. Storn, R.: On the usage of differential evolution for function optimization. NAFIPS'96, pp. 519–523 (1995)
10. Deb, K., Pratap, A., et al.: A fast and elitist multi-objective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. **6**, 182–197 (2000)
11. Deb, K., Jain, S.: Running performance metrics for evolutionary multi-objective optimization, Technical Report 2002004, KanGAL, Indian Institute of Technology, Kanpur 208016, India (2003)

Chapter 28
Research on Semantic Based Distributed Service Discovery in P2P Environments/ Networks

Feng Xu and Shusheng Zhang

Abstract In order to reduce the network communication cost of the discovery of distributed services and improve the search efficiency and recall performance, researchers put forward a distributed service discovery strategy based on semantic in P2P environment. Firstly, they constructed an ontology model to describe the types of services, and then located the starting point of the search process on the associated Peer node based on semantic service classification. Secondly, a double layer parallel service discovery method was put forward. In UDDI layer, they used the classic keyword matching to search services in UDDI center. In the semantic layer, researchers used the semantic query and reasoning on the service ontology model. Finally, the performance of the proposed method was verified by experiment.

Keywords P2P · Semantic · Ontology · The discovery of distributed services

28.1 Introduction

Along with the development of semantic web [1], ontology [2] and SOA technology, the mass information in the network environment and the resources, services that enterprises providing are encapsulated into the form of Web services for using. Facing the diversity and the complexity of expressions and structures of web services, how to find the appropriate Web service rapidly and accurately according to the user's needs in mass Web services has become the research hotspot. Efficient Web service discovery technology has become the key of using the Web service effectively.

F. Xu (✉) · S. Zhang
The Key Laboratory of Contemporary Design and Integrated Manufacturing Technology Ministry of Education, Northwestern Polytechnic University, Xi'an, China
e-mail: sag_angel@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_28,

Universal description, discovery and integration protocol (UDDI) [3] were used to support registration, description and discovery of web services. It uses the SOAP protocol to transmit messages, and the Web Services Description Language (WSDL) to describe the Web service and its interfaces. To be more precise, carefully depicting the web service's capacity, supporting the more precise matching between users' needs and web services' descriptions, researchers introduced the semantic web technology [4]. With the help of logic inference of ontology, the capacity of machines' understandable to the service description information is strengthened, and support the logical reasoning matching between users' needs and services' capability.

In addition, how to store, index, exchange the service metadata is also a research hotspot. In the process of the service discovery, the centralized registration and storage of the service metadata, may have the following drawbacks: (1) it is easy to generate a bottleneck of network transmission if the process involves too many network nodes; (2) when the error occurs on the UDDI server, the entire network web services' discovery will get in the blind area. (3) There are a large number of the providers of Web services, and they may be in different locations and organizations, which requires web services registering unified on a centralized UDDI server. In fact, it is not feasible. Because certain area may have its own UDDI server, or even a business, a group may also have their own private UDDI servers.

Therefore, we need to introduce P2P technology to handle the metadata exchange, so as to overcome the restriction to service discovery caused by traditional UDDI technology service metadata centralized registration, centralized storage [5, 6]. P2P is a distributed application architecture that partitions tasks or workloads among peers. A pure P2P network does not have the notion of clients or servers but only equal peer nodes that simultaneously function as both "clients" and "servers" to the other nodes on the network [7]. Just like a Unstructured P2P networks, which do not impose any structure on the overlay networks. Peers in these networks connect in an ad-hoc fashion based on some loose set of rules [8]. However, in structured P2P networks, peers are organized following specific criteria and algorithms, which lead to overlays with specific topologies and properties [9].

The METEOR-S project, using domain ontology to describe the UDDI centers in P2P environment semanticly, and on this base semi-automatic semantic Web service search mechanism was put forward [10]. Tsinghua University proposed a web service discovery mechanism based on the P2P and semantic [11, 12]: firstly, they make the P2P network structured and divide them by grouping; then they propose two layer search mechanism. Namely, finding the possible group that service may be in at first, then find the possible service node in the group.

Here, in order to improve the performances, like efficiency of service discovery, recall and so on, we put forward a distributed service discovery strategy based on semantic in P2P environment. Firstly, we constructed an ontology model to describe the types of services, and then located the starting point of the search process on the associated Peer node based on semantic service classification; Secondly, a double layer parallel service discovery method was put forward: on

UDDI layer, we used the classic keyword matching to search services in UDDI center; on the semantic layer, we used the semantic query and reasoning on the service ontology model. Finally, the performance of the proposed method was verified by experiment. Comparing with the existing methods, we pay more attention to reduce the searching hops. In the less hops, we find corresponding web services, the less search cost and the less network load we could realize and find the corresponding web service rapidly. In addition, we made full use of the characteristics of the P2P network; we didn't deliberately stressed that the P2P network must be form in structure. Therefore, the cost of network communication could be reduced. We divide the P2P UDDI servers according to the types of services division, and advocate the web service providers register the web service according to the service domain or the regional registration.

28.2 Semantic Classification of UDDI Oriented Service Category Ontology Model Construction

In order to avoid the high searching overhead and improve the efficiency of searching in the P2P environment, we will build an ontology model to describe the categories of web services. Based on this kind of semantic service category, we could locate the starting point on the associated Peer node, which can be targeted for searching distributed web services, and the blindness in the searching process for distributed web services will be reduced.

Generally, which UDDI server will be selected to register the web services on follows certain rule: Web services' providers will normally be in accordance with the principle of proximity, that means the service will be registered to the UDDI server nearby, such as the corresponding regional, national UDDI servers; In addition, the web services' provider may also select a UDDI server to register their web services according to the type, the industry, the field and so on, of web services; finally, there is another possibility, in accordance with the organization, enterprises and other types which the web services attach to, the web services' provider will select a UDDI registration, such as a group's UDDI server or an organization's UDDI server.

Considering the above, we construct the service category ontology model. Create the corresponding class of the UDDI classification criteria—the Service Classified class. This class has three subclasses, including Field, Region and Organization. At present, we have added some instances to the three classes. The instances didn't distinguish strictly according to their types for covering more instances. The Field class includes the Petrifaction, Sale, Service, Telecommunications, Sewage Waste Treatment, Architecture, Manufacturing, Hazardous Waste Disposal, Energy Source, Aviation, Traffic and other fields' instances; the Region class includes the Asia, South_America, North_America, Europe, Africa, Oceania, Antarctica and other regional instances; the Organization class includes the

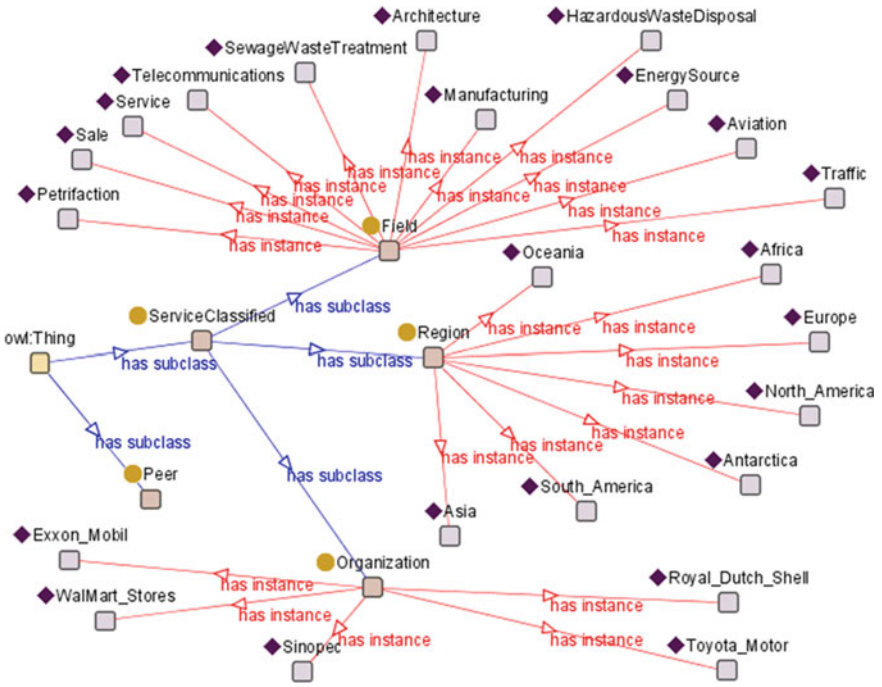


Fig. 28.1 Semantic classification of UDDI oriented service category ontology model

Exxon_mobil, WalMart_Stores, Sinopec, Toyota_Motor, Royal_Dutch_Shell and other organizations.

In addition, only classifying the services is not enough. We need to add the relation between the instance and the UDDI addresses storing corresponding services. So a class—Peer was put forward to express UDDI server. The Peer class has a property named URL, which represents the UDDI server's access address. In order to reflect the relation, we add a Haspeer property between the class Service Classified and class Peer, as shown in Fig. 28.1.

28.3 Double-Layer Parallel Service Discovery Method

The network flooding query consumes the huge costs of time and resources in the P2P environment. The costs occurred mainly in multi-hop query, because each Accumulation of the query hops' number, the costs will exponentially increase. So, we need to find the appropriate web service, as far as possible in a limited number of hops. Then there will be no case where the most of peers fall into the query. It will reduce the search costs and enhance its efficiency. If we need to search the web service in a finite number of hops, we must extend the breadth of

searching. That means, a more comprehensive search must be executed in certain query hop.

Here, we propose a double-layer parallel service discovery method. In the P2P environment, a peer stores the UDDI library and the semantic ontology base describing the web service. In UDDI layer, the UDDI center using the classic keyword matching method to discover the web services; in the semantic layer, we use semantic query and reasoning on the service category ontology model. Note that, the service described by the ontology may not be released in the UDDI library on the same peer. Because the ontology library and the UDDI library is two different things, has not to store in a peer together.

In UDDI layer, according to a set of standard based norms, that UDDI provided for the description and discovery of services, and the support of realization based on internet. As the corporation Indus logic's product Soap uddi includes the following several software packages: com.induslogic.uddi: defines the object in all the registration and discovery process of web services; com.induslogic.uddiserver.inquiry: supports web services discovery; com.induslogic.uddiserver.publish: supports the release of web services; com.induslogic.uddiserver.replication: used to replicate the registration information; com.induslogic.uddiserver.service: includes all the remote-procedure-call object; com. induslogic.uddiserver.soap: contains a direct call to the object RPC Router; com.induslogic.uddiserver.util: supports the database connection. When the program response to the request of service discovery, it use a remote procedure call method calls RPC Router, and the request will be transferred to related classes, such as com.induslogic.uddiserver.inquiry Find Service class, through Uddi Service invoke Appropriate Api method. Then through the get Date method, do the keywords based query to the database, the feedback result is an Envelope object.

In the semantic layer, using the Semantic Annotations for WSDL (SAWSDL) standard, we add the semantic information to the web services; establish the mapping relation between the Web services and the ontology. The mapping relations have been registered to the service registry center since the release of the web services. Then, the discovery of the web services can be realized by reasoning the ontology to get the corresponding concepts, which links to the web service's parameters. Finally, these matching results are integrated by the entire matching results of the web service. The principle diagram is as shown in Fig. 28.2.

28.4 Experiment Studying

To check the effectiveness of the proposed algorithm, using Java (the development language) and Eclipse (the development tools), using Apache AXIS, Jena, jUDDI, Racer and other three party kits, we developed a prototype system, whose functions includes P2P network communications (using flooding communication mechanism in the P2P network) function, ontology reasoning function, UDDI query function.

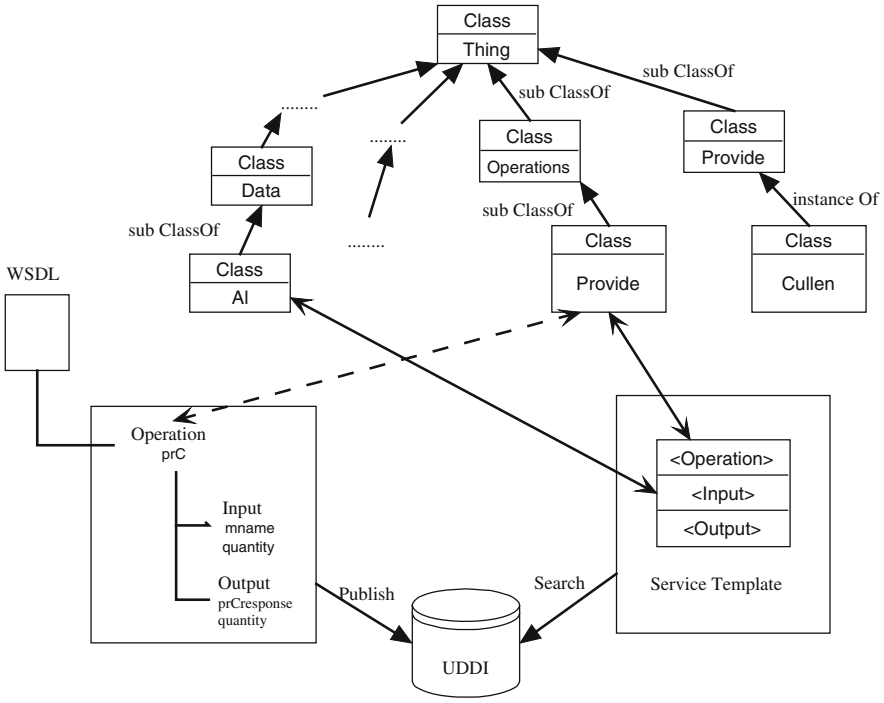


Fig. 28.2 Matching principle diagram

We compare our algorithm (SDSD) to the classic algorithm which directly search the distributed UDDI in the P2P environment using keywords based query method for services discovery (DSD), and focus on two performances, including the network communication cost (the maximum number of forwarding the queries to find a service, $H_{\max}$) and the recall (R).

Collect and construct the UDDI libraries and service semantic ontology libraries that the experiment needed. Divide the service and add the instances to the service category ontology. The experiment is 20 computers installed with Windows XP operating system in the lab. Each machine is configured with the CPU of 2G frequency, the memory of 2G. The experiment was performed 50 times, and we set the P2P nodes' connectivity matrix every times. The experiment results are shown in Fig. 28.3.

By inquiring and reasoning the service category ontology to find the general range at first, and then using the double-layer parallel service discovery method to search the web services, SDSD algorithm increases the search range and reduces the search depth effectively, so the Probability of one-time finding related services are large, the average of the maximum forwarding counts is only 0.7 times, far better than the DSD algorithm 8.2 times; in addition, due to the small scale of the network, and the quantity of the web services needed to be discovered is small, so

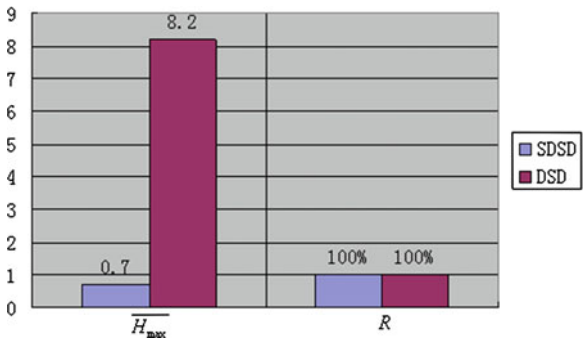


Fig. 28.3 The comparison between SDSD and DSD on H_{max} and R

the averages of the two algorithms' Recalls are both 100 %. From the comparison, we could conclude that, SDSD algorithm has excellent performances in experimental environment, but it still needs to be validated in practice further in future.

28.5 Conclusion

This paper presented a distributed service discovery strategy based on semantic in P2P environment to reduce the distributed service discovery network communication cost, improve search efficiency and recall. Firstly, researchers constructed an ontology model to describe the types of services, and then located the starting point of the search process on the associated Peer node based on semantic service classification; Secondly, a double layer parallel service discovery method was put forward. In UDDI layer, researchers used the classic keyword matching to search services in UDDI center; in the semantic layer, they used the semantic query and reasoning on the service ontology model. Finally, an experiment shows the excellent performances on the network communication cost and the recall.

References

1. Shadbolt, N., Hall, W., Berners-Lee, T.: The semantic web revisited. Intell. Syst. IEEE **21**(3), 96–101 (2006)
2. Fensel, D., Van_Harmelen, F., Horrocks, I. et al.: OIL: an ontology infrastructure for the Semantic Web. Intell. Syst. IEEE **16**(2), 38–45 (2001)
3. Walsh, A.E.: Uddi Soap, Wsdl. The Web Services Specification Reference Book. Prentice Hall Professional Technical Reference (2002)
4. McIlraith, S.A., Son, T.C., Zeng, H.L.: Semantic web services. Intell. Syst. IEEE **16**(2), 46–53 (2001)
5. Huang, L.: A P2P service discovery strategy based on content catalogues. Data Sci. J. **6**, 492–S499 (2007)

6. Verma, K., Sivashanmugam, K., Sheth, A.: METEOR-S WSDI: a scalable P2P infrastructure of registries for semantic publication and discovery of web services. Inf. Technol. Manage. **6**(1), 17–39 (2005)
7. Schollmeier, R.: A definition of peer-to-peer networking for the classification of peer-to-peer architectures and applications. Proceedings of the First International Conference on Peer-to-Peer Computing, IEEE (2002)
8. Shen, X.M., Yu, H., John, B et al.: Handbook of Peer-to-Peer Networking. Springer, New York (2009)
9. Kelaskar, M., Matossian, V., Mehra, P et al.: A Study of Discovery Mechanisms for Peer-to-Peer Application (2002)
10. Verma, K., Sivashanmugam, K., Sheth, A., et al.: METEOR-S WSDI: a scalable P2P infrastructure of registries for semantic publication and discovery of web services. J. Info. Technol. Manag. **6**(1), 17–39 (2005)
11. Chen, D.W., Xu, B., Cai, Y.R., et al.: A P2P basedweb service discovery mechanism with bounding deployment and publication. Chin. J. Comput. **28**(4), 615–626 (2005)
12. Du, Z.X., Huai, J.P.: Research and implementation of an active distributed web service registry. J. Softw. **17**(3), 454–462 (2006)

Chapter 29
Fast String Matching Algorithm Based on the Skip Algorithm

Wenqing Wu, Hongbo Fan, Lijun Liu and Qingsong Huang

Abstract String matching is a fundamental problem in computer science. In order to gain higher performance online exact single pattern string matching algorithms, the authors improved the Skip algorithm which is a comparison based exact single pattern string matching algorithm. By introducing the q-grams method in the sliding window and the comparing window of Skip respectively, the branch prediction failure and the average branch cost are reduced. Meanwhile, the greedy jump method is introduced in Skip. Greedy jump is a common accelerating method for string matching, while there are some waste reads for algorithms with unfixed jump distance. Since Skip has fixed jump distance, greedy jump is more suitable for Skip. Therefore, the authors presented the HGQS algorithm. Experiments results have indicated that HGQS has higher practical performance and it is more efficient in some cases than other known algorithms in many cases on the platform.

Keywords String matching · Q-grams · Greedy jump · HGQS · Design of algorithm

29.1 Introduction

String matching is a fundamental problem in computer science. Poor performance of string matching caused a serious impact on our research especially in the main research fields of our team which are the large-scale information retrieval and the mass medical information processing.

W. Wu · H. Fan · Q. Huang (✉)
Department of Computer Science Kunming University of Science and Technology, Yunnan Key Laboratory of Computer Technology Applications, Kunming, China
e-mail: kmustailab@hotmail.com

L. Liu
Department of Biomedical Engineering, Kunming University of Science and Technology, Kunming, China
e-mail: cloneiq@126.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_29,

Online exact single pattern string matching is the basis of string matching which means seeking all the occurrences of a pattern $P = P\ [0,\ldots,m-1]$ of length m in a text $T = T[0,\ldots,n-1]$ of length n over the same alphabet Σ of length σ,which the text has not be preprocessed to build some structure to fast locate the pattern, such as index etc. The research of this problem has special significance. In this paper, we concentrate on practical online exact single pattern string matching, and all algorithms in this paper are for searching an exact pattern.

Skip is a simple and efficient comparison based string matching algorithm. Its jump distance is fixed at m. Experimental results show that Skip gains good performance for large alphabet and very short pattern. The idea of Skip is quite different from the other type of comparison based algorithms, BM type algorithms [1]. Although The BM type algorithms have been researched intensively, Skip was ignored for a long time, and the related researches based on Skip are rare. For example, the article [2] combined Skip and Quick Search, although our experimental results this approach can not enhance the performance than original Skip.

In this article, we improved the Skip by introducing the q-grams method and the greedy jump method which were used in current high-speed algorithms. Therefore we presented a new serial of algorithms named HGQS. Experiments results indicated that HGQS is more efficient in some cases than other known algorithms in our platform.

29.2 Introduction of Skip

In the preprocessing phase, Skip creates a barrel based on the pattern, which is a linked list for each character of the alphabet to record each position of this character in the pattern. Skip uses the sliding window method like mostly algorithm. In the searching phase, Skip reads the last character of the sliding window, and gains the occurring positions of this character in the pattern by the barrel. Then some windows of the same length with the pattern are determined based on above positions. The string matching can be found only in these windows. These windows are called Match window.

Skip checks these Match windows with the naive matching. If a string matching is found, the position of the match window should be reported. If all match windows have been checked or there is none possible match window, the sliding window jumps m. The matching process of a sliding window is shown in Fig. 29.1, which the last character in the sliding window is assumed as x.

The C type pseudo-code of Skip is listed in **Code.1**.

Code.1 The Skip algorithm ***SKIP*** (*P, T, m, n*){
//preprocessing phase

1: **for** $c \in \Sigma$ **do** $z[c] \leftarrow NULL$;
2: **for** $i \leftarrow 0$ **to** $m-1$ **do**
3: {$ptr \leftarrow$ **new** List Node; **if** $ptr = NULL$ **then return** "Error.";

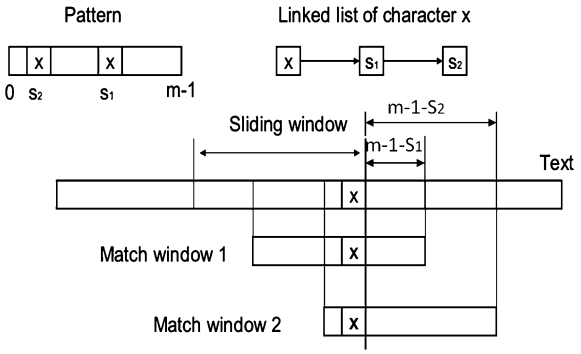


Fig. 29.1 The matching process of skip

4: $(ptr- > element) \leftarrow i$; $(ptr- > next) \leftarrow z[p_i]$; $z[p_i] \leftarrow ptr$;}

//searching phase

5: **for** $j \leftarrow m-1$ **to** $n-1$ **step** m **do**
6: **for** $(ptr \leftarrow z[t_j]$; $ptr \neq NULL$; $ptr \leftarrow (ptr- > next))$
7: **if** $P[0..m-1] = T[j-ptr- > element..j-ptr- > element+m-1]$ **then**
8: *Report find matching* $(j-ptr- > element)$;}*//end of Skip*

29.3 The *q*-grams Method

In the string matching phase, branch operations consume the most of computing time. The cost of branch operation is no-fixed. It is determined by the branch prediction failure rate. If branch prediction succeeds, the branch operation only needs one processor tick, otherwise, because the instructions have being in the pipeline are wrong, the processor should empty all of instructions have been in pipeline which results in a punishment of dozens of ticks. Reducing the branch prediction failure rate is priority to the optimization of algorithm performance.

In Skip, once only one character is aimed whether in the skipping of Sliding Window or in the comparing of Match Window. Thus the branch prediction failure rate and the average branch cost are very high for small alphabet and long pattern. To improve the performance of the algorithm, now we introduce the *q*-grams method into the two operations.

The *q*-grams technique is an important accelerating technique on pipeline processors. To date, many of the current fast algorithms use *q*-grams.

In *q*-grams, *q* consecutive characters are processed as a single character. For example, the string "Hello" is processed as "Hel-ell-llo" for 3-grams. This can make the alphabet perceived by the algorithm larger. Because a fairly large alphabet leads to a low rate of branch prediction failure, a high probability of jump

and a high average jump distance, q-grams can greatly enhance the performance. Another benefit of q-grams is that q-grams solutions are more flexible than the original one. Generally speaking, smaller alphabets or longer patterns demand bigger q.

After q-grams method introduced, for the string S of length m of the sliding window, it is processed to a string of length $m - q + 1$ in Σ^q. Then we create barrels by Σ^q with the method of Skip. While, both the branch prediction failure rate and the average count of the Match Windows needing to be checked in each sliding window are reduced exponentially. This is the q-grams method for the sliding window.

If we record the consequent first q characters of the match window in an integer by joining these characters with the shift operation, to compare this integer with the integer that is joined with the first consequent q characters with the same method of the pattern just need one operation. It also implemented the q-grams. The branch would be wrong predicted only if these q characters were all matched, therefore the branch prediction failure rate will be reduced exponentially. For the comparison of the other subsequent characters in the match window, naive matching can provide enough high performance because the previous integer comparison already filters most of the positions could not be matched.

By introducing the q-grams method in the sliding window and the match window, the branch cost can be significantly reduced. However, after the q-grams method introduced, characters read in each sliding window and match window were increased, and the jump distance of the sliding window was reduced to m-q + 1. To get higher performance, we should select the fair value of q by experiments.

Meanwhile, after the q-grams method introduced, the preprocessing space occupancy (the capacity of the table z) is increased exponentially too. When the space is bigger than the capacity of processor cache, the access performance would be deteriorated seriously. Therefore, we introduced a Hash method refers to Q-hash [3]. It can reduce the preprocessing space occupancy and improve the performance of the algorithm. The Hash function of this Hash method is (29.1),

$$hash(t_i, \ldots, t_{i+q-1}) = (t_i << (q-1) * s) + \ldots + (t_{i+q-2} << s) + t_{i+q-1} \quad (29.1)$$

which the value of s should be fulfilled that the space of the entire hash table is about half of the processor L1-Cache. It will increase the utilization of the Cache as much as possible.

The q-grams method introduced in the sliding window is called Q-grams-Sliding Windows method (Q-method), and the q-grams method introduced in the match window is called int and char-Hybird-Comparison-Matching Windows method (H-method). And we named the Skip algorithm which introduced the above methods as HQ-Skip algorithm.

29.4 The Greedy Jump Method

The greedy jump method is a common accelerating method which is presented in GSB2 [4]. Since the greedy jump method is used, GSB2 is faster than its basis algorithm, SBNDM2, in some cases.

When the core loop of an algorithm can reach its largest jump distance with a high probability, the greedy jump method assumes the current window can reach the largest jump distance and it is read that the characters in the following window that the current window would have been jumped with the largest jump distance. If the above assumption were not correct, the jumping and the reading based on the assumption would be ignored and the original matching method of the algorithm would be used to match the current window. Because the probability of reaching the largest jump distance is high, the greedy jump can double the jump distance of the core loop of algorithm with two times of read operation in most of the cases, which should reduce the average total number of operations in the core loop of an algorithm. Moreover, in the core loop of the greedy jump, the memory access and branch operations in the processing of different windows can be inherent parallelism on the superscalar processor because there is no date dependence in these two windows, thus some of the processing delay is hidden. Therefore, the performance of the algorithm should be improved after the greedy jump method is introduced.

For example, in the greedy jump of GSB2, the core loop of SBNDM2 is modified from **while** $(B[t_i] << 1)\ \&\ [t_{i-1}] = 0$ **do** $i \leftarrow i + m - 1$; to

while $(B[t_i] << 1)\&B[t_{i-1}] = 0\ \&\&\ (B[t_{i+m-1}] << 1)\&B[t_{i+m-2}] = 0$

do $i \leftarrow i + m * 2 - 2$.

On algorithms with unfixed jump distance, greedy jump reads some not utilized characters by the failure assumption. But the algorithms with fixed jump distance can overcome this lack to some degree because the characters have been read must be used in the next window. Although this time of read is ignored, these characters have been in the registers or high speed L1-Cacheand, and they can be gained very fast at the next time of read. Therefore, the serial algorithms of Skip are more suitable for greedy jump.

Let the Greedy jump method (G-method) introduce to HQ-Skip, the result algorithms are named HGQS. In the greedy jump of GSB2, it reads two windows once in a core loop. Actually, reading multiple windows can also fulfill the matching. Therefore, there are three parameters in HGQS: the q value in the q-grams of Q-method, another q value in the q-grams of H-method and the count of windows that the G-method read once in a core loop. If all the three parameters are assumed to 2, and mark the target algorithm as HGQS_$q2g2h2$, The code of HGQS_$q2g2h2$ is shown in **Code.2**.

Code.2 The HGQS_$q2g2h2$ algorithm *//preprocessing phase*

1: $s \leftarrow 8$; $q \leftarrow 2$; $h \leftarrow 2$; **if** $m < \max(q, h)$ **return** *error;*
2: *CheckInt* $\leftarrow (p_0 << 8) + p_1$; **for** $c \in \Sigma$ **do** $z[c] \leftarrow$ *NULL*;

3: **for** $i \leftarrow 0$ **to** $m - q$ **do**
4: $\{ptr \leftarrow$ **new** List Node; **if** $ptr = NULL$ **then return** "error";
5: $(ptr- > element) \leftarrow i$; $(ptr- > next) \leftarrow z[hash(p_i, p_{i+1})]$;
6: $z[hash(p_i, p_{i+1})] \leftarrow ptr$;$\}$
7: *Naive matching p* **from** t_{n-2m} **to** t_{n-1};
8: **for** $i \leftarrow 0$ **to** $m - 1$ **do** $\{ t_{n-m*2+i} \leftarrow t_{n-m+i} \leftarrow p_i\}$

//searching phase

9: **for** $j \leftarrow m - q$ **to** $n - 1$ **step** $m - q + 1$ **do**
10: {**while** $z[hash(t_j, t_{j+1})] = NULL$ **and** $z[hash(t_{j+m-1}, t_{j+m})] = NULL$
11: **do** $j \leftarrow j + m * 2 - 2$;
12: **for** $(ptr \leftarrow z[hash(t_j, t_{j+1})]$; $ptr \neq NULL$; $ptr \leftarrow (ptr- > next))$
13: {**if** $CheckInt = (y[j - ptr- > element] << 8) + y[j + 1 - ptr- > element]$
14: **then continue;**
15: **if** $P[2...m-1] = T[j + 2 - ptr- > element...j - ptr- > element + m - 1]$
16: **then** *Report* $(j - ptr- > element)$;$\}$*//end of HGQS_q2g2h2*

In Code.2, we used a cross-border protection method which was used frequently in the field. Several times the joined pattern appeared after the tail of the text indicated that there must have pattern matching before the occurrence of cross-border. Therefore the check operations of the cross-border in the core loop are not need. The cross-border protection zone must larger than the largest jump distance of the core loop after the greedy jump method was adopted.

29.5 Experiments Results

To show the performance of HGQS, we do the following comparative experiment based on SMART, which is the implement of article [5, 6].

The platform is Intel I7-2600k@3.4Ghz/z68/8 GB DDR3 RAM/Ubuntu 10.04LTS 64-bit desktop edition/g ++4.4.5 with -O3 optimize parameter. We tested under the DNA sequence *E.coli*, English test Bible.txt, natural language samples text world192.txt and the 20 MB length random texts with alphabet size is 16, which the first three texts are from SMART. For each matching condition, the 100 patterns in each pattern set were picked from the 100 prior random selected and non-overlapping positions of the text, and the average matching speed was recorded as the final result which the highest and the lowest 20 % results were ignored. The text had been read in memory to avoid the impact of disk and before each time of matching phase was timed by RDTSC[1](±30 CPU ticks). In this

[1] http://en.wikipedia.org/wiki/Time_Stamp_Counter.

Table 29.1 Experimental results for the random text of alphabet size 16, which unit is MB/s

m = 2		m = 8		m = 64		m = 256	
EBOM [12]	1337.1	. SBNDM2_2i32	5535.0	SBNDM4b_a64	14144	**HGQS_*q4g3h1***	**21011**
TVSBS_w6 [11]	1224.1	EBOM	5468.9	HGQS_*q*4*g*2*h*1	13806	QF_3_5i32	18551
HGQS_*q*2*g*3*h*2	1181.3	GSB2b_i32	5255.9	FSBNDM_w4a64	13336	EBOM	15968
GSB2b_i32 [4]	1151.4.	**HGQS_*q2g2h3***	5009.3	UFNDM4_a64	13266	BXS5_a64	14348

Table 29.2 Experimental results for DNA sequence, which unit is MB/s

m = 2		m = 8		m = 64		m = 256	
FSO_u61a64	953.5	. **HGQS_*q4g2h1***	2791.6	SBNDM6b_a64	12344	**HGQS_*q5g2h2***	**16255**
Shift-Or_i32	928.0	. SBNDM4b_i32	2732.9	**HGQS_*q5g2h3***	11559	QF_5_3i32	12479
Shift-And_i32	713.0	. UFNDM4_a64	2466.7	FSBNDMqq6f0	11504	SBNDM6_a64	11748
EBOM	640.8.	FSBNDMqq4f1	2397.3	UFNDM6_a64	11194	EBOM	9897.3

Table 29.3 Experimental result for English text, which unit is MB/s

m = 2		m = 8		m = 64		m = 256	
EBOM	1214.7	SBNDM2_2i32	4198.9	**HGQS_*q4g2h1***	**11755**	**HGQS_*q4g2h1***	**16214**
TVSBS_w6	1160.8	**.HGQS_*q3g3h3***	3922.6	SBNDM6b_a64	11543	QF_4_4i32	14318
GSB2b_i32	1067.1	FSBNDMqq3f1	3836.1	TVSBS_w6	11350	TVSBS_w6	13398
FSO_u61a64	1048.6.	EBOM	3812.9	UFNDM6_a64	11169	BXS5_a64	12800

Table 29.4 Experimental result for the natural language sample, which unit is MB/s

m = 2		m = 8		m = 64		m = 256	
TVSBS_w6	1407.5	SBNDM2_2i32	5070.9	**HGQS_*q4g2h1***	**12459**	**HGQS_*q4g2h1***	**17706**
EBOM	1311.0	EBOM	4749.9	TVSBS_w8	11958	QF_3_5i32	15723
GSB2b_i32	1134.3	GSB2b_i32	4608.4	SBNDM4b_a64	11509	TVSBS_w6	14380
HGQS_*q2g3h2*	**1107.6**	**HGQS_*q2g2h3***	4354.8	UFNDM4_a64	11457	BXS5_a64	13745

experiment, the CPU frequency was locked by cpufrequtils,[2] the network and the unrelated background service are closed to ensure the processor utilization was below to 3 %. This experiment environment is similar with the environment of the article [5, 6], but we used $-O_3$ parameter with auto branch prediction and we only timed the matching phase, which more suits for the study habits of the string matching field.

SMART has given the implements of 85 known algorithms, which covered most known algorithms as of May 2010. This experiment continued the work of SMART, complemented the FSO [7], BXS [8], QF [8], Q-Hash_4096 [9], SBNDM*q*b [9], BSDM*q* [10], FSBNDM-w*k* [11], SBNDM-w*k* [11], FS-w*k* [11] TVBS-wk [11], FSBNDM*q* [12], FSBNDM*q*b [12], GSB2 [12], GSB2b [12]etc., which are newer algorithms not be included by SMART. All bit parallel solutions are implemented with the 32-bit edition (i32) and the 64-bit edition (a64) and tested both of them. And just some parameters of the solutions are listed in SMART. This experiment complemented the unlisted parameters. E.g., only three conditions of Hash*q* ($q = 3, 5, 8$) were listed in SMART, we complemented the solutions of $q = 4, 6, 7$ etc. If an algorithm with different parameters were called different algorithms, there would be more than 300 algorithms compared and these algorithms covered most of known algorithms.

The experiment results of the optimize parameter and its average matching speed of the fastest four type algorithms (if some algorithms are only inconsistent with parameter such as q value of q-grams, and these algorithms are presented in one article, it is called that these algorithms belong to a type algorithm) are listed for random texts with alphabet size is 16, and DNA sequence, English test and the sample of natural language in from Tables 29.1, 29.2, 29.3, 29.4. Due to length constraints, for a type of algorithms, only the fastest one is listed and the unit of speed is MB/s. Tables 29.2, 29.3.

[2] http://wiki.archlinux.org/index.php/Cpufrequtils.

From the experimental results, it is indicated that HGQS is faster than other known algorithms for long patterns and for the DNA sequence with patterns of length 8 on our platform.

29.6 Conclusion

In this article, the authors introduced the q-grams method and the greedy jump method in Skip. Therefore, a serial of improved algorithms of Skip which is named HGQS are presented. The Experimental results have indicated that HGQS has higher practical matching performance and it is more efficient than existing algorithms for long patterns and for the DNA sequence with patterns of length 8.

Acknowledgments This paper is supported by the Yunnan Province Social Development Science and Technology Project of China (NO.2010CA016) and the Innovation Fund for Technology Based Firms of China (NO.10C26215305130). Thanks for the funding.

References

1. Boyer, R.S., Moore, J.S.: A fast string searching algorithm. Commun. ACM **20**(10), 762–772 (1977)
2. Naser, M.A.S., Rashid, N.A., Aboalmaaly, M.F.: Quick-skip search hybrid algorithm for the exact string matching problem. Int. J. Comput. Theory Eng. http://www.ijcte.org/papers/462-G1278.pdf (2012)
3. Lecroq, T.: Fast exact string matching algorithms. Info. Process. Lett. **102**(6), 229–235 (2007)
4. Peltola, H., Tarhio, J.: Variations of forward-SBNDM. In: 16th Prague Stringology Conference, PSC2011.http://www.stringology.org/event/2011/p02.html (2011)
5. Faro, S., Lecroq, T.: The exact string matching problem: a comprehensive experimental evaluation. Computing Research Repository. http://arxiv.org/abs/1012.2547 (2010)
6. Faro, S., Lecroq, T.: The exact online string matching problem: a review of the most recent results. ACM Comput. Surv. http://www-igm.univ-mlv.fr/∼lecroq/articles/acmsurv2013.pdf (2013)
7. Fredriksson, K., Grabowski, S.: Practical and optimal string matching. In: The 12th Symposium on String Processing and Information Retrieval, pp. 376–387. Springer, Berlin (2005)
8. Durian, B., Peltola, H., Salmela, L et al.: Bit-parallel search algorithms for long patterns. In: LNCS 6049: The 9th International Symposium on Experimental Algorithms, SEA 2010, vol. 2010, pp. 129–140, Springer, Berlin (2010)
9. Durian, B., Holub, J., Peltola, H et al.: Tuning BNDM with q-grams. In: Proceedings of the 11th Workshop on Algorithm Engineering and Experiments, ALENEX2009, vol. 2009, pp. 29–37. SIAM, New York (2009)
10. Faro, S., Lecroq, T.: A fast suffix automata based algorithm for exact online string matching. In: LNCS 7276: The 17th International Conference on Implementation and Application of Automatad (CIAA 2012), vol. 2012, pp. 149–158. Springer, Berlin (2012)

11. Faro, S., Lecroq, T.: A multiple sliding windows approach to speed up string matching algorithms. In: LNCS 7276: The 11th International Symposium on Experimental Algorithms, SEA2012, vol. 2012, pp. 172–183 (2012)
12. Faro, S., Lecroq, T.: Efficient variants of the backward-oracle-matching algorithm. Int. J. Found. Comput. Sci. **20**(6), 967–984 (2009). doi:10.1142/S0129054109006991

Chapter 30
Burst Signal Sorting Based on the Phase Continuity

Fangmin Yan, Ming Li and Ling You

Abstract This paper proposed a new algorithm for burst signal sorting. The proposed algorithm can be used to identify and locate TDMA users based on the continuity of carrier phase. In the proposed algorithm, the continuity of carrier phase between TDMA burst signals is evaluated according to their frequency deviations, initial phases and initial positions. Then burst signals are sorted based on their degree of continuity. The proposed algorithm is effective when researchers do not know the information which the burst carries. Some simulations and experiments in this paper show that the accurate rate of the proposed sorting algorithm is greater than 0.9 when the ES/N0 $>$ 6 dB. Specially, and the performance is stable when the frequency deviation changes.

Keywords Burst signal sorting · Phase continuity · Carrier phase · Frequency deviation · Initial position

30.1 Introduction

Burst-mode signal is widely used in the applications such as satellite communication and mobile communication. As an important part of the blind burst-signal processing, burst signal sorting technique has received much attention, especially when identifying and locating different TDMA users from the captured bursts. Some signal sorting methods have been proposed for radar signal in recent years [1–3], using the pattern recognition theory. For the communication signals, Huang utilized the small carrier frequency difference to sort the burst signals [4], but it can't work when the frequency difference was equal (or close). Our sorting method

F. Yan (✉) · M. Li · L. You
Science and Technology on Blind Signal Processing Laboratory, Chengdu, China
e-mail: yanfangmin07@gmail.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_30,

is focused on the carrier phase, aiming at using the inherent relation between the initial phase and the initial position and the frequency deviation of bursts to sort bursts which belong to different users from the same emitter.

The algorithm proposed in this paper does not rely on the information that the burst signal carries. In order to realize our sorting algorithm, we need to estimate three important parameters: the frequency deviation, the initial phase and the initial position. Luckily, there exist many methods that can accomplish these estimations. In this paper, we will adopt some proper methods and do some effective modifications to achieve better performance.

The rest of this paper is organized as follows. Section 30.2 gives a description of the burst signal model. In Sect. 30.3, we propose a new algorithm for the burst signal sorting. Section 30.4 analyzes the factor that influences the performance of the proposed algorithm. In Sect. 30.5, simulation results are provided and analyzed. Finally, we conclude the paper in Sect. 30.6.

30.2 Problem Formulation

Consider the transmission of TDMA burst signals emitted from some users. The signals are MPSK modulated with carrier frequency f_c. Each burst signal consists of N symbols, where the first L symbols are a fixed-length sequence which we call them "sync word", and the remaining $N - L$ symbols are random data. If the received burst signal is over-sampled by P, it can be modeled as

$$r_n = s_n \exp(j(2\pi f_c \cdot nT_s + \phi_{i0})) + z_n, n = 1, \ldots, PN \tag{30.1}$$

Here, $s_n = \sum_{k=-\infty}^{\infty} \exp(j\varphi_k)\, h(nT_s - kT - \tau)$ is the baseband signal, $\{\varphi_k\}$ are data symbols belonging to the MPSK alphabet $\{\exp(j2\pi m/M),\ m = 0, 1, \cdots, (M-1)\}$, $h(t)$ is the squared root raised cosine (SRRC) pulse waveform, τ is the timing error. T_s is the sampling period, $T = PT_s$ is the symbol period, ϕ_{i0} is the initial carrier phase of the ith user. z_n is zero-mean additive white Gaussian noise with independent real and imaginary components, each having a variance of $N_0/2E_s$. E_s and N_0 denote the symbol energy and the noise power spectral density ($SNR = E_s/N_0$), respectively.

The number of bursts is N_s. Each user owns at least one burst. Both the length of each burst and the interval between two adjacent bursts do not change during the transmission. Then, all the received bursts with frequency f_c can be expressed as

$$R(n) = \sum_{m=1}^{N_s} r(nT_s - mN'T) \tag{30.2}$$

where, $N' = N + N_0$, N_0 is the interval between two adjacent bursts.

In this paper, bursts from the same emitters are clustered based on $R(n)$.

30.3 Algorithm Description and Analysis

The purpose of our sorting algorithm is to cluster the bursts which belong to the same user. In general, no matter whether the emitter transmits data or not, the carrier of the user works regularly, i.e., the carrier phase of burst signals emitted from the same user are continuous.

Figure 30.1 shows the burst transmission model of multiple users that work at the same frequency. The user1 ensures the continuity of carrier phase when it works. Our sorting algorithm is on the basis of this fact.

30.3.1 Algorithm Theorem

Assuming the kth burst in $R(n)$ is $R_k(n)$, the initial position of $R_k(n)$ is μ_k, the frequency deviation of $R_k(n)$ is Δf_k and the initial phase of $R_k(n)$ is Pha_k. We consider that bursts from the same user have the same (or close) frequency deviation, so that their carrier phases are continuing which is shown in Fig. 30.2.

In Fig. 30.2, the continuity of carrier phase is expressed as

$$Pha = 2\pi\Delta ft + \phi_{i0} \tag{30.3}$$

If the parameters of the kth burst and the mth burst (k > m) satisfy (30.4),

$$Pha_k - [Pha_m + 2\pi\Delta f(\mu_k - \mu_m)] < \Delta \tag{30.4}$$

Then, we decide that the kth burst and the mth burst both belong to the ith user. Otherwise they must belong to different users. In (30.4), Δ called a threshold, is an experienced value.

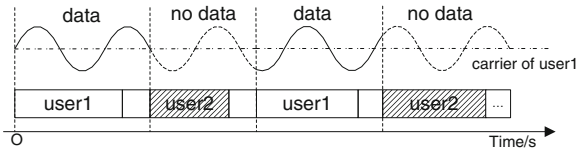


Fig. 30.1 Burst transmission model of multiple users

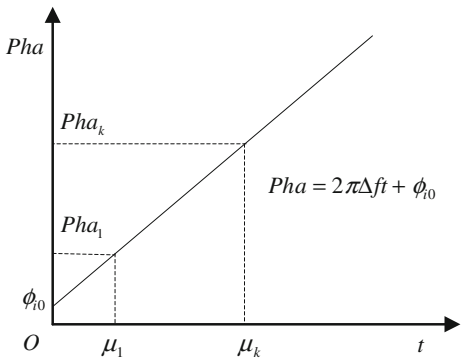


Fig. 30.2 Continuity of carrier phase

30.3.2 The Procedure of Sorting Algorithm

- Firstly, a proper DA burst detection method is used to detect the bursts $R(n)$ and estimate their initial position. Denote the set of all the bursts as A.
- Secondly, the frequency deviation of each burst is estimated.
- Thirdly, after the frequency deviation of each burst is compensated, the initial carrier phase of each burst can be estimated. All the initial carrier phases of the same user bursts should be continue which is shown in Fig. 30.2.
- Finally, when we have gotten all the parameters (the frequency deviation, the initial phase and the initial position) of bursts in A, we can utilize the inequality (30.4) to sort different bursts. In the process of sorting, we make all the other bursts compare with the first burst (the initial position is smallest). If one burst satisfies (30.4), it belongs to the user who sends the first burst and will be eliminated from A. Denote the set of bursts which belong to the user who send the first burst as B. Denote the rest of the bursts as A^1, $A^1 = A - B$. Then, A^1 is sorted by the same process. Repeating the above sorting process to A^1, we will get another set of bursts B^1 and another rest of bursts as A^2, $A^2 = A^1 - B^1$. The sorting process continues until the rest of bursts are empty.

Certainly, when we utilize the proposed algorithm to sort the bursts, we prefer fine estimations of parameters. In this paper, we adopt the average likelihood ratio test method to estimate the initial position of the burst [5]. The ML estimation method is used to estimate the frequency deviation of each burst [6]. We utilize a feed-forward timing estimation method to estimate the timing error [7]. For the initial phase estimation, a NDA method is utilized [8, 9].

The estimation methods for parameters are very ripe now. Our sorting algorithm does not pay much attention to these estimation methods and what we should do is to choose a proper method from the existing methods for each parameter.

The proposed sorting algorithm is not an on-line processing method, so the complexity analysis of the implement is ignored.

30.4 Analysis for Sorting Algorithm Errors

According to the analysis, the estimation RMSEE for the initial position, the frequency deviation (normalized by the symbol rate) and the initial phase of the burst signal is constricted to the quantity of 10^0, 10^{-5} and 10^{-2}, respectively, when the ES/N0 is greater than 6 dB [5, 6, 8]. Analyzing the equality (30.3), we can find that the estimation error of the initial position and the initial phase is independent between different bursts, i.e., the error can be ignored by setting a proper threshold. But the frequency deviation is accumulated by the time t. Even though the estimation RMSEE of frequency deviation is small, the estimation error of the frequency deviation is the main influence factor for the error of proposed sorting algorithm.

Assuming the estimations of the phase and the position are equal (or close) to their theoretical value. Denote the theoretical frequency deviation as Δf, the estimation phase of the mth burst as Pha_m, the estimation frequency deviation of the base burst as $\Delta f'$, the initial position of the base burst as α_0, the initial position of the mth burst as α_m, then

$$\begin{aligned} Pha_m &= 2\pi\Delta f(\alpha_m - \alpha_0) + \phi_{i0} \\ Pha'_m &= 2\pi\Delta f'(\alpha_m - \alpha_0) + \phi_{i0} \end{aligned} \tag{30.5}$$

In (30.5), Pha'_m is the derivation from $\Delta f'$. Subtracting Pha'_m from Pha_m,

$$Pha_m - Pha'_m = 2\pi\Delta f_e(\alpha_m - \alpha_0) \tag{30.6}$$

In (30.6), $\Delta f_e = \Delta f - \Delta f'$. If $abs(Pha_m - Pha'_m) < \Delta_1$, the mth burst and the base burst both belong to the same user. Δ_1 is an experienced threshold.

In order to overcome the error of sorting different bursts, we consider two methods: one by shortening the accumulated time $(\alpha_m - \alpha_0)$ and the other by decreasing the estimation error (Δf_e) of Δf.

30.4.1 Shortening the Accumulated Time

In equality (30.3), we do not limit the accumulated time, so the sorting performance is heavily influenced by the estimation error of Δf. Now, we shorten the accumulated time by replacing the base burst if the number of bursts belongs to the same user x is larger than N_r. N_r is a fixed-integer. Then the base burst is replaced by the newest burst belongs to x. Figure 30.3 shows the result of shortening the accumulated time for Δf. The accumulated time is shortened as $\{\alpha_1, \alpha_2 - \alpha_1, \alpha_3 - \alpha_2, \cdots, \alpha_k - \alpha_{k-1}, \cdots\}$.

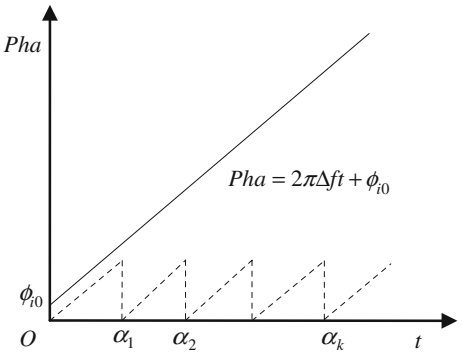


Fig. 30.3 The result of shortening the accumulated time

30.4.2 Decreasing the Estimation Error of Frequency Deviation

According to the analysis, we know that the ML estimation of Δf is an unbiased estimate [6]. Assuming the estimation result of Δf is $\Delta f'$, so

$$E\{\Delta f'\} = \Delta f \tag{30.7}$$

If we average the frequency deviation estimation value of all the bursts which belong to the same user x,

$$\Delta f'_a = \frac{\Delta f'_1 + \Delta f'_2 + \cdots + \Delta f'_M}{M} \tag{30.8}$$

where, M is the number of bursts which belong to x, $\Delta f'_i$ is the frequency deviation estimation value of the ith burst in x.When M becomes larger, $\Delta f'_a$ becomes closer to the real frequency deviation Δf. So we can use $\Delta f'_a$ as the fine frequency deviation estimation of the base burst to sort other bursts.

30.5 Experimental Results

Performance of the proposed sorting algorithm is presented by experimental evaluation. Assuming four users work at the same carrier frequency (5,000 Hz) and produce the bursts in proper order (1–2–3–4). All the burst signals are modulated by QPSK. The initial phase of each user is 0, pi/6, pi/3, pi/2. The symbol rate and the sampling rate are 2,500 and 40,000 Hz, respectively.

In order to evaluate the performance of the proposed sorting algorithm, we simulate 100 bursts (each user has 25 bursts). For each user, the ideal frequency deviations is -4e-4, -2e-4, 2e-4, 4e-4 (normalized by the symbol rate), respectively. Figure 30.4 shows that the performance of the proposed sorting algorithm gets better when ES/N0 increases. Especially when ES/N0 $>$ 6 dB, the accurate rate of the sorting algorithm is greater than 90 %.

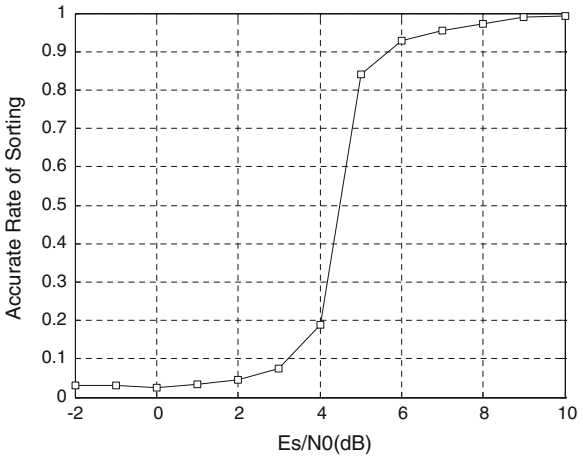


Fig. 30.4 Performance of sorting against ES/N0

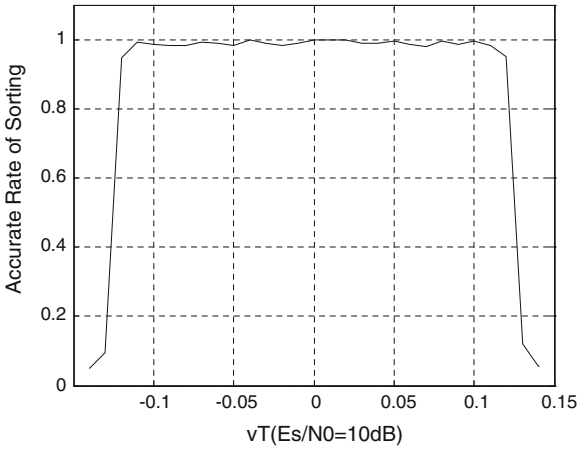


Fig. 30.5 Performance of sorting against frequency deviation (v is the frequency deviation, T is the symbol period, vT is normarlized frequency deviation by the symbol rate)

Then, we simulate 100 bursts (each user has 25 bursts) under ES/N0 = 10 dB. Figure 30.5 shows that the performance of our algorithm keeps stable when the frequency deviation changes. Therefore, we can conclude that the performance of our sorting algorithm is almost uninfluenced by the estimation error of the frequency deviation.

Finally, we simulate 100 bursts (each user has 25 bursts) under ES/N0 = 10 dB to observe the visualized sorting results as shown in Fig. 30.6. For each user, the ideal frequency deviations is -4e-4, -2e-4, 2e-4, 4e-4 (normalized by the symbol rate), respectively. Statistics shows that the accurate rate of our algorithm is 100 %. Figure 30.7 shows the original sorting result (shown in Fig. 30.2) against the improved sorting result (after the correction of sorting errors). The original sorting result is converted from the improved sorting result, because of easy observation.

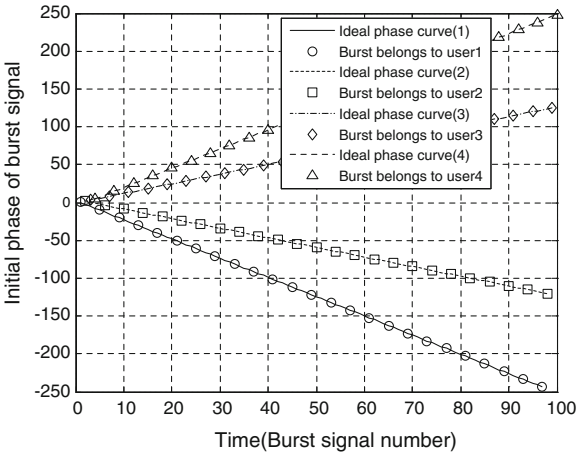


Fig. 30.6 Results of the proposed sorting algorithm (ES/N0 = 10 dB)

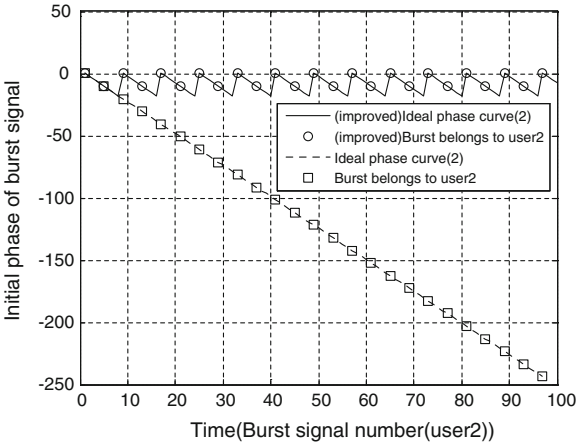


Fig. 30.7 Original sorting result against improved sorting result

30.6 Conclusion

A new algorithm for burst signal sorting was proposed in this paper. The proposed algorithm utilized the continuity of carrier phase to sort different bursts which belonged to different users. This algorithm relied on the inherent relation between the initial phase, the initial position and the frequency deviation of the burst signal, and did not depend on the information which the signal carries. Simulation results showed that the proposed sorting algorithm had an accurate rate greater than 0.9 when ES/N0 $>$ 6 dB and was robust to the frequency deviation.

However, it should be noted that the theoretical performance of the proposed algorithm was not analyzed in this paper. It is expected and will be researchers' future work.

References

1. Yu, Z., et al.: A multi-parameter synthetic signal sorting algorithm based on clustering. ICEMI '2007, vol. 2, pp. 363–366 (2007)
2. Guo, Q., et al.: A novel sorting method of radar signals based on support vector clustering and delaminating coupling. ICCI '06, vol. 2, pp. 839–844 (2006)
3. Zhang, Y., Sun, G.: Application of radial basis function neural networks in complicated radar signal measurement and sorting. ICEMI'2007, vol. 3, pp. 375–378 (2007)
4. Huang, Y., Lu, Y.: Small carrier frequency difference detection based on the relative phase entropy. ISCIT 2007, pp 1417–1422 (2007)
5. Huang, Y.-l., et al.: Robust burst detection based on the average likelihood ratio test. J. Electron. Inf. Technol. **32**(2), 345–349 (2010)
6. Viterbi, A.J.: Principles of Coherent Communication. McGraw-Hill, New York (1966)
7. Morelli, M., Andrea, A.N.D., et al.: Feedforward ML-based timing estimation with PSK signals. IEEE Commun. Lett. **1**, 80–82 (1997)
8. Noels, N., et al.: Carrier phase and frequency estimation for pilot-symbol assisted transmission: bounds and algorithms. IEEE Trans. Signal Process. **53**, 4578–4587 (2005)
9. Erup, L., et al.: Interpolation in digital modems-part II: implementation and performance. IEEE Trans. Commun. **41**, 998–1008 (1993)

Chapter 31
Fast Recapture and Positioning Algorithm Based on PMF-FFT Structure

Xinpeng Yue, Haiyang Quan and Lidong Lan

Abstract GNSS signal acquisition is the most important process in a receiver followed by tracking and extraction of navigation bits. Partial Matching Filter (PMF) and Fast Fourier Transform (FFT) algorithm has advantages in acquisition speed and hardware complexity. In general, GNSS navigation data acquisition needs a common frame synchronization algorithm, which takes one sub-frame period of time for a determination. This means the receiver will take at least 6 s to reposition after the signal lost lock and recapture. In this paper, the design of PMF-FFT based receiver is described. A fast method of solving the long-time frame synchronization problems is proposed. The method uses the α-β filter algorithm to correct local time and estimate signal sending time. Experimental results show that the proposed methods for the PMF-FFT based receiver are able to perform faster and reliable acquisition and reposition.

Keywords GNSS · PMF-FFT · Recapture · GPS · Compass-2

31.1 Introduction

Global Navigation Satellite System (GNSS) is a satellite-based radio navigation system, such as GPS, COMPASS-2, and GALILEO and so on. It has been widely used in both military and civilian community for navigation, location, timing, and other related applications. In this paper, we mainly study on the GPS and COMPASS-2 satellite system. The satellite navigation receivers capture the RF modulated signals, down convert them to an intermediate frequency (IF), digitize them,

X. Yue (✉) · H. Quan · L. Lan
Development and Application Department, Beijing Microelectronics Technology Institute, Beijing, China
e-mail: yue017@126.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_31,

and perform signal processing to extract the position information from the navigation message.

Signal acquisition is the first step of signal process and the most important process in a GNSS receiver followed by tracking and extraction of navigation bits. Performing faster and reliable acquisition and having lower hardware complexity are very important to the receiver. PMF-FFT algorithm [1, 2] is a parallel acquisition algorithm. It has higher speed than two dimensional serial acquisition algorithms [3] and lower hardware complexity than FFT based time-domain parallel acquisition algorithm [2, 4]. We designed PMF-FFT structure and build the relevant receiver. After signal reacquisition, the receiver performs tracking, bit synchronization, frame synchronization and position processes. However, among all processes frame synchronization takes longer [2]. So, how to shorten the frame synchronization time is the most important method to decrease the position time and improve receiver performance.

31.2 The GPS Navigation Message and Frame Synchronization

The GPS and COMPASS-2 system have the similar signal structure. We will mainly introduce the GPS system. The signal is made of carrier, C/A code, and navigation message data. Each GPS satellite (or transmitter) has a unique C/A code (spreading gold code) that is orthogonal to all the other satellites' codes. GPS receivers, on the other hand, must search for these C/A codes to know which satellites are available to the user. For each code, a receiver must perform a 2-D search for carrier frequency offset and code shift, or in other words acquire the C/A code. Then it should track (or lock in) the signal and extract navigation messages. Navigation messages stream be transmitted by the satellite on the L1 carrier frequency of 1575.42 MHz at a rate of 50 bps. The message structure shall utilize a basic format of a 1500 bit long frame made up of five subframes, each subframe being 300 bits long. Subframes four and five shall be subcommutated 25 times each, so that a complete data message shall require the transmission of 25 full frames. Each subframe shall consist of ten words, each 30 bits long; the MSB of all words shall be transmitted first. Each subframe and/or page of a subframe shall contain a telemetry (TLM) word and a handover word (HOW), both of which are generated by the satellite, and shall start with the TLM/HOW pair.

The TLM word shall be transmitted first, immediately followed by the HOW. The HOW shall be followed by eight data words. Each word in each frame shall contain parity. Each TLM word is 30 bits long, occurs every 6 s in the data frame, and is the first word in each subframe/page. Each TLM word shall begin with a preamble, which is used to frame synchronization. After, extracting navigation messages from the signal, the receiver don't know where the starter is. Frame synchronization process can find the preamble 10001011 to confirm the starter of a

subframe and then start to extract navigation message and the satellite signal sending time.

31.3 Analysis of PMF-FFT Algorithm and Receiver

Partial Matching Filter will divide T_{coh}(a prediction integration time, PIT) into P portion, then every segment data $T_p = T_{coh}/P$, and every segment has C/A code chips $X = 1023^* T_P$. If P = 1, the Partial Matching Filter become the Full Matching Filter, or in other words serial acquisition algorithm. Usually receiver chooses 1 ms (period of C/A code) signal to acquire. All sample points are first filled with several zeros (first zero filling) and cut into P parts, and each part has X points. Then receiver produces local C/A code and make the same partition, M is the number of T_{coh} between the local C/A code and the GPS C/A code. Then both signal and C/A codes are at the same time sent into PMF in which average correlation operation is made. The outputs Q are filled with several zeros (Q to N) to meet the requirement of 2-based FFT and then do the FFT. Figure 31.1, show the structure of PMF-FFT algorithm.

The GPS signal can be expressed as:

$$S_i = \sqrt{A} D_{(t)} C_{(t-\tau)} \cos(\omega_0 t + \omega_d t + \varphi) + n_{(t)} \tag{31.1}$$

where A is the amplitude of signal; $D_{(t)}$ is navigation date code each element of which has a breadth of 20 ms; $C_{(t-\tau)}$ is pseudo-random code (C/A code) with a period of 1 ms and 1023 code elements in each period; ω_0 is medial frequency; ω_d is Doppler frequency; Φ is carrier phase; n is gauss noise.

The PMF filters perform correlation operation with receiving signal and generating L1 C/A code. Here, $R_{(\tau)}$ is the C/A code self-correlation the output of the n-th parts in the PIT time can be expressed as [5, 6]:

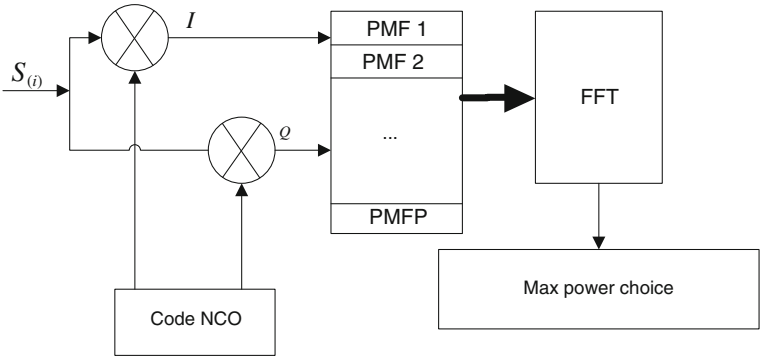


Fig. 31.1 The structure of PMF-FFT

$$Q_{(n)} = \frac{\sqrt{2A}}{2} R_{(\tau)} \frac{\sin(\pi f_d T_P)}{\sin(\pi f_d T_S)} \sin\left(n w_d T_P + \varphi\right) + N_{Q(n)} \quad \text{n} = 0 \ldots \text{p} - 1 \tag{31.2}$$

$$I_{(n)} = \frac{\sqrt{2A}}{2} R_{(\tau)} \frac{\sin(\pi f_d T_P)}{\sin(\pi f_d T_S)} \cos\left(n w_d T_P + \varphi\right) + N_{Q(n)} \quad \text{n} = 0 \ldots \text{p} - 1 \tag{31.3}$$

Combining $Q_{(n)}$ and $I_{(n)}$ as $I_{(n)} + j^* Q_{(n)}$ and then performing FFT by zero filling, the outputs can be expressed real part and image part as:

$$Q_{(k)} = \frac{\sqrt{2A}}{2} R_{(\tau)} \sin\psi \frac{\sin(\pi f_d T_P) \sin(\pi f_d T_P - k\pi P/N)}{\sin(\pi f_d T_S) \sin(\pi f_d T_P - k\pi/N)} + N_{Q(k)} \quad \text{k} = 0 \ldots \text{N} - 1 \tag{31.4}$$

$$I_{(k)} = \frac{\sqrt{2A}}{2} R_{(\tau)} \cos\psi \frac{\sin(\pi f_d T_P) \sin(\pi f_d T_P - k\pi P/N)}{\sin(\pi f_d T_S) \sin(\pi f_d T_P - k\pi/N)} + N_{Q(k)} \quad \text{k} = 0 \ldots \text{N} - 1 \tag{31.5}$$

Here, $\psi = \varphi + (\omega_d T_P / T_s - 2\pi k/N)(N-1)$.

The GNSS receiver based on PMF-FFT structure is designed to acquiring, tracking and positioning. The GNSS receiver we designed can implement the GPS, COMPASS-II and GPS + COMPASS-II navigation. There are mainly two parts of the receiver: RF part and digital process part. The RF part converts the radio frequency signal to a lower frequency and then digital process part perform acquiring, tracking and positioning process to output the results. In the designed receiver, we choose intermediate frequency of 4.092 MHZ and sampling frequency of 16.38 MHZ. Figure 31.2 shows the prototype broad of real receiver.

31.4 Faster Positioning Methods

In the GNSS receiver, the most important information is the navigation messages sending time and transforming time (or pseudo-range). The final aim of

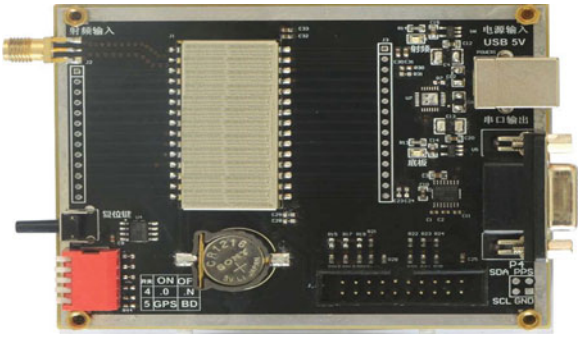


Fig. 31.2 Prototype broad of real receiver

acquisition, tracking and extracting of data is about to getting the navigation messages sending time and transforming time. The performance and capability of the receiver is mainly about the precision of time. The calculation of messages sending time can be expressed as:

$$t^{(s)} = TOW + (30^{*}w + b)^{*}0.02 + \left(c + \frac{CP}{1023}\right)^{*} 0.001 \tag{31.6}$$

where TOW is the 17 MSBs of the time-of-week count; w is the count of word which start from the prompt subframe, as we introduce in Sect. 31.2, Each subframe shall consist of ten words, each 30 bits; b is count of bits, each 20 ms width; c is cycle count of C/A code with a period of 1 ms and 1023 code elements in each period; CP is C/A code phase which mainly decide the accurate of pseudo-range. Pseudo-range is defined as below:

$$\rho = r + c * (\delta t_u - \delta t^{(s)}) + I + T + \varepsilon_p = r + c * (\delta t_u - \delta t^{(s)}) + \varepsilon \tag{31.7}$$

where t_u is receiver's time; $t^{(s)}$ is message sending time; c stands for the speed of light; r is the range between receiver and satellite; δt_u and $\delta t^{(s)}$ is the clock deviation of receiver's clock and satellite clock; I stands for ionospheric delay; T stands for troposphere delay; ε_p is measure deviation; ε is called estimation deviation, which is of the accumulation of I, T, ε_p. As we analyze the message sending time composing before, considering the ε can be neglectable, compared with cycle count of C/A code. The most important factor is how to get the accurate bits count and C/A code cycle count.

The pseudo-range is mainly decided by the distance between satellite and user. We can estimate the movement of pseudo-range. As Fig. 31.3 shows earth radius R_e is 6,368 km, the distance R_s is 26560 km. The satellite orbits the earth every 12 h. We suppose the satellite is traveling at constant velocity, so we can calculate his angular speed $\frac{d\theta}{dt} \approx \frac{2\pi}{12*3600} = 1.454 * 10^{-14}$ [rad/s] and cutting speed $v_s = R_s \frac{d\theta}{dt} \approx 3862$ [m/s]. From the geometry principle, we can get expression of v_d as:

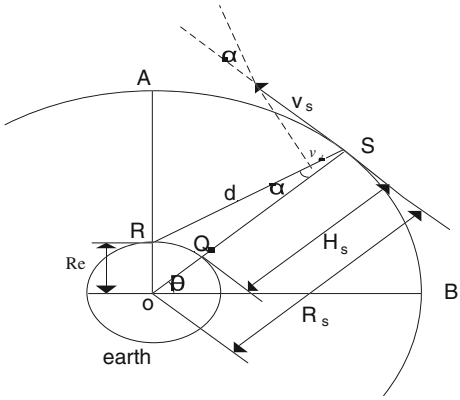


Fig. 31.3 Model of satellite orbit

$v_d = \frac{v_s R_e \cos\theta}{\sqrt{R_e^2 + R_s^2 - 2R_e R_s \sin\theta}}$. Assuming $\frac{dv_d}{d\theta}$ is zero, the maximum value $v_{dm} = v_s \frac{R_e}{R_s} =$ 925.9 [m/s]. The rate of v_d is $\frac{dv_d}{dt} = \frac{dv_d}{d\theta}\frac{d\theta}{dt}$, we can easily find that when $\theta = 90^\circ$, the $\frac{dv_d}{d\theta}$ and $\frac{dv_d}{dt}$ have the max absolute value, so we can get the max absolute value of $\frac{dv_d}{dt} = 0.177$ m/s^2.

The analysis shows us that there is a very small acceleration in the direction between receiver and satellites. In other words, the velocity rate is smooth. Therefore, we can estimate the distance between satellites and receiver using the having information with α-β filter after the signal lost and recapturing, realizing faster positioning. The faster position α-β method can be described as:

$$\begin{cases} x_k = (1-\alpha)x_a + \alpha x_b \\ v_k = (1-\alpha)v_a + \alpha v_b \\ x_a = (1-\beta)x_{k-1} + \beta v_{k-1} t \\ v_a = (1-\beta)v_{k-1} + \dfrac{\beta}{t}(x_k - x_a) \end{cases} \tag{31.8}$$

x_k is filtered distance; x_a is estimated distance; x_b is measured distance; x_{k-1} is last recorded distance. v_k is filtered velocity; v_a is estimated velocity; v_b is measured velocity; v_{k-1} is last recorded velocity .

In the receiver soft design, the receiver records the bits count and ms count before signal lost, and refreshes the satellite data sending time with the receiver count, so we can quickly obtain the massage sending time after the signal tracking again without frame synchronization. When tracking again, we import measuring sending time and estimating sending time to α-β filter, so we can get more precise signal transforming time and achieve more precise position. Compared with others, such as frame synchronization method [7, 8], this method can also achieve faster reposition after tracking again and reduce the calculations, in other words,

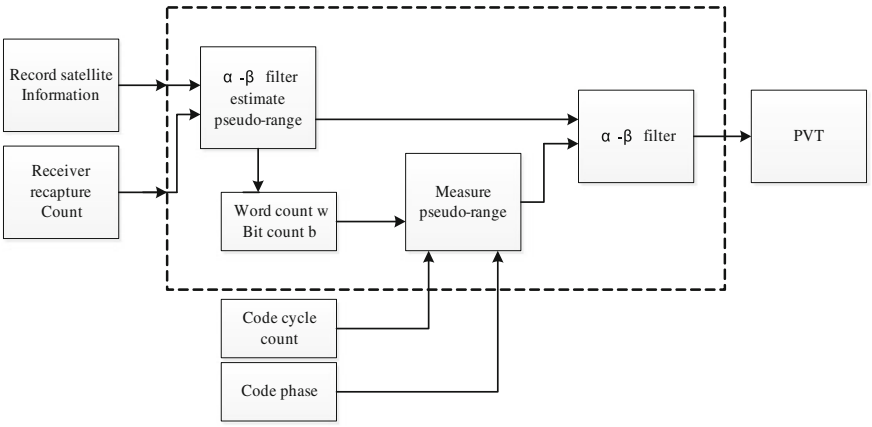


Fig. 31.4 Flow chart of fast synchronization and accurate reposition

Table 31.1 Comparision of two test channel

Number of experiments	1st		2nd		3rd		4th	
Channel	1st	2nd	1st	2nd	1st	2nd	1st	2nd
Estimated pseudo-range(s)	0.07483608	0.07483548	0.07941842	0.07941863	0.07268464	0.07268504	0.08116371	0.08116353
Pseudo-range(s)	0.07483534	0.07483534	0.07940885	0.07940885	0.07268527	0.07268527	0.08116398	0.08116398
Word count w	6	6	7	7	8	8	5	5
Bit count b	5	5	0	0	27	27	18	18
Reception time(s)	0.8		0.85		0.89		1.2	

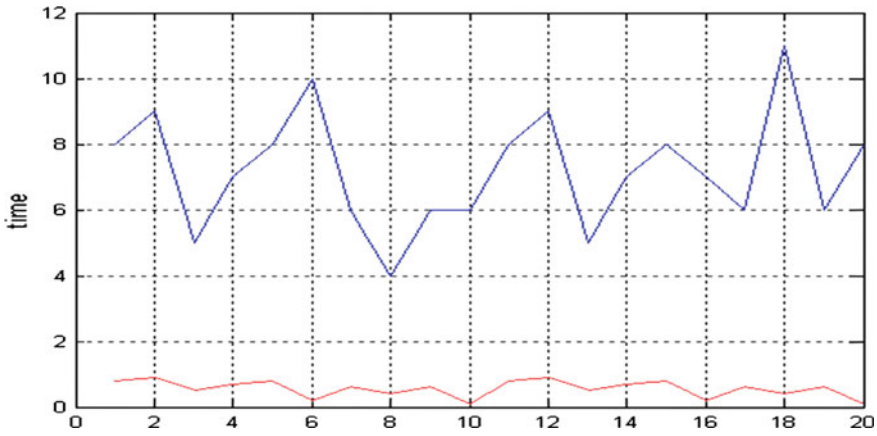


Fig. 31.5 Time consumption of reposition

the hardware time cost can be lower. Figure 31.4 is the flow chart of fast frame synchronization and accurate reposition.

31.5 Experiment Results

To test and verify the correctness, effectiveness and practicality, we mainly carried out two sets of experiments on our receiver. The experimental L1 signal is from Spirent signal hardware simulator GSS8000 and real satellite.

Fig. 31.6 Outfield experiment

31.5.1 The Correctness Verification

In the experiments, we choose two tracking channel to track the same satellite, one as the reference channel, the other as the test channel. First, let the two channels track the prn7 satellite, then channel 1 starts to recapture the signal, and we read the output data, repeating four times. The results are shown in the Table 31.1.

Estimate pseudo-range is the α-β filter estimated pseudo-range. Pseudo-range stands for the receiver measured pseudo-range. The experimental results of the Table 31.1 show that channel 1 gets the same w and b with channel 2 witch is reference channel, computed from estimated pseudo-range, and verify the correctness of proposing method.

31.5.2 The Rapidity of Reposition

Figure 31.5 shows the time consumption of reposition after tracking again. The blue is consumption of the normal position method which needs 6 s to frame synchronization [2]. The red is the consumption of proposing method. The normal position method consumes longer time than the proposed method, which takes no more than one second.

31.5.3 The Practicality Verification

The proposed method can make sure the continuous trajectory after the receiver through the overpass. This experiment was taken in South Four Ring Road, in Beijing. Figure 31.6 shows the output of outfield experiment.

31.6 Conclusion

In this paper, the PMF-FFT algorithm has been introduced and based receiver has been described. The α-β filter based algorithm has been used to reduce the time consumption of frame synchronization. The result shows that the PMF-FFT structure has a faster acquisition speed and the receiver can perform an accurate position after tracking again. In the common scene, the receiver can have a good performance and make sure the continuum of position. There are some problems in the proposed method, for example, the time of signal losing has been limited. So, in the future work, we will concentrate on this and other adding method.

References

1. Akopian, D.: Fast FFT based GPS satellite acquisition methods. IEEE Proc. Radar Sonar Navig. **152**(4), 277–286 (2005)
2. Xie, G.: Principle of GPS and Receiver Design. Publishing House of Electronics Industry, Beijing (2010)
3. Genzici, s.: Mean acquisition time analysis of fixed-step serial search algorithms. IEEE Trans. Wirel. Commun. **8**(3), 1536–1276 (2009)
4. Grant, P.M., Spangenberg, S.M., Scott, I., et al.: Doppler estimation for fast acquisition in spread spectrum communication systems. In: Proceedings of IEEE 5th International Symposium on Spread Spectrum Techniques and Applications, pp. 106–110(1998)
5. Qi, H., Shi, X., Ji, L.: PMF-FFT algorithm for PN code acquisition. J Xi'an Technol. Univ. **30**(1), 57–61 (2010)
6. Tantartans, S, Lam, A.W, Vincent, P.J.: Noncoherent sequential acquisition of PN sequence for DS/SS communications with/without channel fading. IEEE Trans. Comput. **43**(3), 1738–1745 (1995)
7. Shi, G., Xiao, L.G., Chen, Y.: A fast frame synchronization method of GPS receiver after signal lost and recapturing. Measur. Control Technol. **31**(4), 124–129 (2012)
8. Sajabi, C., Chen, C.I.H., Lin, D.M., et al.: FPGA frequency domain based GPS coarse acquisition processor using FFT. In: Proceedings Instrumentation and Measurement Technology Conference, pp. 2353–2358 (2006)

Chapter 32
An Evaluation Computing Method Based on Cloud Model with Core Space and its Application: Bridges Management Evaluation

Ling Chen, Le Ma and Zhao Liang

Abstract In a multi-factor comprehensive evaluation, the factors related to objects are always various and most of them have the characteristics of uncertainty. Taking the mapping between qualitative and quantitative knowledge of cloud model, a high dimensional cloud model with core space was built. And then based on a sample set of maintenance and management of the 55 bridges in Chongqing and an index system with six first-level indices, parameters of the high dimensional cloud model with core space and mean membership of every bridge sample were computed and gotten. Compared with the results of cloud model, experts experience and support vector machine for this sample, it indicated the cloud model with core space could be applied to a multi-attribute evaluation well. Finally, according to the evaluation, some suggestion was given.

Keywords Core space · High dimensional cloud model · Performance evaluation · Bridges management

32.1 Introduction

In practice, one object will be influenced by many factors and most factors are from language description of realistic world and are with the characteristic of qualitative. Thus an evaluation is featured with multi-attribute and qualitative. For most evaluation, every attribute or weight of indicator should be confirmed at the very beginning, such as weighted average model, fuzzy synthetic evaluation model

L. Chen (✉) · L. Ma
College of Automation, Chongqing University, Chongqing, China
e-mail: ccgg005@gmail.com

Z. Liang
Southwest University, Chongqing, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_32,

and analytic hierarchy process [1–3], which are based on experts' grade or experts' weighted experience and are of randomness and subjectivity to some degree. Artificial intelligence, which is very popular among present researchers [4–6], is to extract effective rules and knowledge from sample information of experts' evaluative experience to make evaluation. However, due to the limitation of sample amount and obtainable experts' experience, the effectiveness of evaluation which is based on rule extraction and classification is not satisfied well.

In this paper, considering the uncertainty such as randomness, subjectivity and fuzziness in evaluation and cutting down the dependence on weighted information and sample amount, it is of great significance to introduce cloud model which is based on traditional fuzzy mathematics and probability statistics. A high dimensional cloud model with core space will be established to evaluate and analyze real condition in Chongqing's bridges maintenance and management.

The remaining sections of this paper are organized as follows. Section 32.2 introduces the new high dimensional cloud model with core space. Section 32.3 describes the general problem of evaluating bridge management quality, and presents the index system for evaluations. Section 32.4 describes our experimental design for data collection and performance comparison. Section 32.5 concludes the paper and suggests directions for future work.

32.2 A High Dimensional Cloud Model with Core Space

32.2.1 The Main Principle of Cloud Model

Could model is an uncertain transformation model between one qualitative concept which is expressed by natural language value and quantitative representation. Given U as a domain expressed by exact number, U is corresponding to qualitative concept A, for every element x in domain, there is a random number with stable tendency $y = \mu_A(x)$, y, is the certainty of x to concept A, the distribution of certainty y in domain is named cloud model [7–9].

The number characteristics of cloud are represented by expectation E_x, entropy E_n, and hyper entropy H_e, which reflect the quantitative feature of qualitative concept A. E_x means the dot which can best represent this qualitative concept in number field, reflecting the position of cloud center. On one hand, E_n reflects the scope of number field space being accepted by language value, being indistinguishable measure of qualitative concept; on the other hand, reflects that the dot in number field space can represent the probability of this language value, showing cloud droplets of qualitative concept having randomness. H_e is the uncertain measure of entropy, reflecting coherency of uncertain degree of all data dots representing this language value in number field space, namely coherency of cloud droplets. The larger the hyper entropy is, the larger the dispersion of cloud droplets is, the larger the randomness of certainty degree is and the thicker the cloud is.

The three number characteristics of cloud model integrate fuzziness and randomness, making up of mapping between qualitative and quantitative, so a cloud model can be $C(E_x, E_n, H_e)$.

Generating algorithm of cloud is named cloud generator which consists of normal cloud generator [7], X condition cloud generator, Y condition cloud generator and reverse cloud generator. Normal and X condition cloud generators are usually used in model evaluation. Normal cloud generator refers to could droplets produced by number characteristics of cloud.

32.2.2 Building the New Model with Core Space

Definition 1 If domain U, $U \in R^m$, $x_i = [x_{i1}, x_{i2}, \cdots, x_{im}]$ is any element in U, and if subset H of U exists, $H \in R^m$, any elements $x_{ik}(\mathrm{k} = 1, 2, \ldots \mathrm{m})$ in H is $a_{k1} \le x_{ik} \le a_{k2}$, and the certainty degree of the element in H by high dimensional cloud model $\mu = F(x_i)$ is $\mu < 1$, and the other elements' certainty degree in U is $\mu < 1$, then H is the core of domain U. Elements' certainty degree of high dimensional cloud model shows in Eq. (32.1):

$$\mu = F(x_{i1}, x_{i2}, \ldots x_{im}) = \begin{cases} 1 & a_{k1} \le x_{ik} \le a_{k2} \quad k = 1, 2, \ldots m \\ \exp\left(-\sum_{k=1}^{m} \frac{(x_{ik}-a)^2}{2E'_{nk}}\right), & x_{ik} < a_{k1} \cup x_{ik} > a_{k2} \end{cases} \tag{32.1}$$

In which

$$a = \begin{cases} a_i, & x_{ik} < a_i \\ b_i, & x_{ik} > b_i \\ E_{xi}, & a_i = b_i \end{cases} \tag{32.2}$$

In the specific attribute consideration, H can be seen as assemblage of all attributes' most excellent chosen interval [8], and when in evaluation, H can be supposed as a set of all attributes' most excellent evaluation.

High dimensional cloud evaluation model with core space and multi attributes are established as:

1. Make sure co-domain of evaluative value of all attributes $[d_{i\min}, d_{i\max}](i = 1, 2, \ldots m)$ and the best value interval $[a_i, b_i](i = 1, 2, \cdots m)$ are the most excellent evaluative core space$H$, and establish every attribute's most excellent evaluative trapezoid cloud model.
2. According to $3E_n$ rules of normal cloud, Eq. (32.3) shows the number characteristics of every dimensional cloud model, when $a_i \neq b_i$, trapezoid cloud model will be formed, and when $a_i = b_i$, normal cloud model will be formed.

$$\begin{cases} E_{xi} = a_i,\ E_{ni} = (E_{xi} - d_{i\min})/3,\ H_{ei} = E_{ni}/6 & when \quad d_{i\min} \le x_i < a_i \\ E_{xi} = b_i,\ E_{ni} = (d_{i\max} - E_{xi})/3,\ H_{ei} = E_{ni}/6 & when \quad b_i < x_i \le d_{i\max} \\ E_{xi} = x_i,\ \mu = 1 \quad when \quad a_i \le x_i \le b_i & in\ which \quad i = 1, 2, \ldots m \\ E_{xi} = a_i = b_i,\ E_{ni} = (d_{i\max} - d_{i\min})/6,\ H_{ei} = E_{ni}/6, & when \quad a_i = b_i \end{cases} \tag{32.3}$$

3. The data assemblage of all dimensional attributes which form m dimension high dimensional cloud space.

$$X_i = C(a_i, E_{ni}, H_{ei}) \cup C(b_i, E_{ni}, H_{ei}) \cup [a_i, b_i],\ (i = 1, 2 \ldots m) \tag{32.4}$$

4. If sample $x_i = [x_{i1}, x_{i2}, \ldots, x_{im}]$ is in core space H, the certainty of core space of this sample to the most excellent evaluation should be $\mu = 1$.
5. If sample $x_i = [x_{i1}, x_{i2}, \cdots, x_{im}]$ is not in core space H, the certainty of core space of this sample to the most excellent evaluation should be calculated according to Eqs. (32.1) and (32.2).
6. On the basis of samples, order the most excellent evaluation core space, namely, sample evaluation result order.

32.3 The Problem of Bridge Management Evaluation

As China's economy has grown, bridge construction has developed rapidly in the Chongqing province. At present, China contains more than 570,000 bridges and In Chongqing province alone there are over 8000. While there is a drive to speed up construction, it is also necessary to improve maintenance management. By evaluating the status of bridge maintenance and the activities of management, we can quantify the effectiveness of current procedures and develop concrete measures for raising the managerial level.

32.3.1 An Index System for Quantifying Bridge Maintenance

Taking bridge maintenance and management of Chongqing as example, according to the experts' advice of Chongqing Bridge Association, Chinese bridge technical standard and related factors, the influencing factors of bridge maintenance and management are proposed, including six indexes which further divided into 25 sub-indexes and 113 qualitative and quantitative indexes. The six indexes are:

maintenance and management condition, maintenance and management expense, quality of technical staff, bridge's construction quality, daily average traffic flow, and service life. Due to limitation of length of this paper, the first level indexes are listed in the Table 32.1.

32.3.2 Experiment Data Set

According to bridge maintenance and management index system, designed examination chart is handed to technical staff, management staff and bridge association experts to evaluate pointed bridges. This chart is mainly designed to the description of bridge management condition, and then each bridge's maintenance and management will be evaluated by bridge experts who will combine the examination and objective condition of bridge (service life, traffic flow and maintenance expense). The evaluation result will be stored as a chart in the data base, and the table of comprehensive evaluation factors of examined 55 bridges' maintenance and management is obtained. "Daily average traffic flow" refers to the data of one year's traffic flow which is averaged to every day. Table 32.1 shows 10 bridges' data. The evaluation result is classified as 1, 2 and 3; class 1 refers to the best maintenance and management condition, 3 the worst.

32.4 Experiment Studying

Firstly, the most excellent evaluation interval of every attribute needed to be set. Table 32.2 shows the most excellent evaluation interval of every attribute, among

Table 32.1 The scores of maintenance and management factors of 20 large bridges in Chongqing

Bridges code	Rank from experts	Index1 Ages	Index 2 Quality of engineer and technician	Index 3 Mean vehicle flow per day	Index 4 Maintenance expenses	Index 5 Maintenance state	Index 6 Construction quality
1	1	4	12	60529	7	61	16
2	1	2	15	20000	8	69	11
3	1	1	15	15000	8	68	11
4	1	2	15	20000	8	76	12
5	1	4	12	64000	8	57	17
6	1	2	15	35000	5	57	8
7	1	15	10	54814	8	64	15
8	1	2	15	35000	5	61	13
9	1	3	9	3000	5	62	14
10	1	2	9	15000	5	60	10

Table 32.2 Optimal interval of every first level index

Index1	Index 2	Index 3	Index 4	Index 5	Index 6
[1, 4]	16	[0,15000]	9	76	17

which indexes 1 and 3 are interval values, the other indexes are maximum values of corresponding attributes Table 32.3.

Table 32.4 shows the certainty degree of bridge maintenance and management to core space, while the result compared with experts can be seen in Fig. 32.1. Due to limitation of length of this paper, Table 32.4 only contained 38 samples of 55 bridges.

The samples in Fig. 32.1 with (**) means that the rank evaluated by high dimensional cloud model with core space is different from experts' evaluation result. According to the certainty of high dimensional model with core space, the maintenance and management level of NO. 33 should be listed in class two, while experts' evaluation result is class three. After detail comparison of sample and inquiry of experts' advice, it is found that the difference is mainly because experts'

Table 32.3 cloud model of every first level index

Index1	Index2	Index3	Index4	Index5	Index6
C(1, 1, 0.1),$x_1 < 1$	C(16, 5, 0.8)	C(15, 5, 0.8),$x_3 < 15$	C(9, 3, 0.5)	C(76, 23, 4)	C(17, 6, 1)
C(4, 10, 1.7),$x_1 > 4$		C(15, 27, 5),$x_3 > 15$			

Table 32.4 Average membership of every bridge for core space

Rank from experts	Membership for core space	Bridges code	Rank from experts	Membership for core space	Bridges code
1	0.92095	4	2	0.40273	17
1	0.87916	2	2	0.36179	31
1	0.87708	3	2	0.36105	26
1	0.7439	8	2	0.3379	34
1	0.72532	9	2	0.32824	27
1	0.71778	7	3	0.31632	36
1	0.7011	1	3	0.2885	35
1	0.68209	16	3	0.27654	41
1	0.66562	5	2**	0.27295**	33**
1	0.64128	12	3	0.26888	40
1	0.63703	10	3	0.26863	39
1	0.5939	6	3	0.26849	37
1	0.57353	18	3	0.26795	38
1	0.57072	11	3	0.26696	45
2	0.56011	20	3	0.23027	43
2	0.51749	22	3	0.22978	42
2	0.50855	14	3	0.21656	46
2	0.50508	21	3	0.21169	48

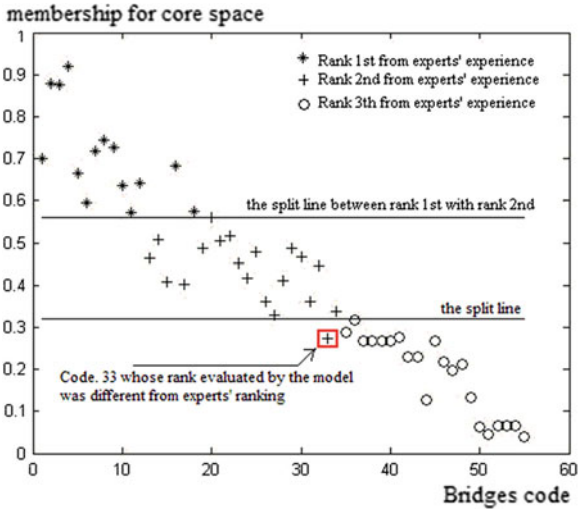


Fig. 32.1 Evaluation comparison with experts experience and cloud model with core space

evaluation on index 5 is "good". While the evaluation value of NO. 33 bridge's index 5 differs greatly from experts' evaluation value as class 2 bridge. However, from the analysis of cloud model with core space, NO. 33 bridge's overall level is close to class 2. Figure 32.1 clearly shows combined with experts' evaluation result and certainty to core space, the dividing line of certainty is very obvious between different classes, further illustrating that the evaluation method which is based on high dimensional cloud model with core space is effective.

To compare evaluation effects, SVM is adopted at the same time to normalize and classify 55 samples, among which 50 samples are training samples, five are testing samples. In this application, when the kernel was a radial basis function. The cross-validation parameter v was set to 3, the kernel function parameter C was 32,768, and g was 0.0019, classification accuracy was 0.6. When v was 7, C was 8,388,606 and g was 7.63e-6, the accuracy of the SVM was unchanged.

It shows that the certainty of each sample bridge can be obtained in high dimensional cloud model with core space evaluation method and the detail information of order are more than simple classification. Compared with experts' evaluation result, the effect is better. At the same time, the comparison of SVM to samples classification accuracy shows that sample amount and attributes uncertainty influences SVM classification.

32.5 Conclusion

The introduction of high dimensional cloud model with core space into bridge management evaluation can fully consider the existence of various uncertain errors, making the evaluation result more effective and reasonable.

The result above shows the high dimensional cloud model with core space for multi-attribute evaluation can enrich present information, and it can get not only accurate and reasonable evaluation classification, but also much delicate information of evaluation process. The method can be applied for a multi-attribute evaluation well.

With the development of Chinese economy and society, the reinforcement of bridge maintenance and management is very urgent and evaluation on the level of bridge maintenance and management is to understand and supervise the condition and process.

Though the division of core space can simplify weighted factor, the accuracy of core space is increased at the same time. So how to integrate weighted information of attribute into model still needs further study.

Acknowledgments Our work is supported by the Fundamental Research Funds for the Central Universities (Project No. CDJZR12170014).

References

1. Lin, H., Lee, H., Wang, D.: Evaluation of factors influencing knowledge sharing based on a fuzzy AHP approach. J. Inf. Sci. **35**(1), 25–44 (2009)
2. Xu, Y., Wang, L.: Fuzzy comprehensive evaluation model based on rough set theory. In: Proceedings of the 5th IEEE International Conference on Cognitive Informatics, pp. 877–880 (2006)
3. Tenenhaus, A., Giron, A., Viennet, E., et al.: Kernel logistic PLS: a tool for supervised nonlinear dimensionality reduction and binary classification. Comput. Stat. Data Anal. **51**(9), 4083–4100 (2007)
4. Dembuzynski, K., Greco, S., Slowinski, R.: Rough set approach to multiple criteria classification with imprecise evaluations and assignments. Eur. J. Oper. Res. **198**(2), 626–636 (2009)
5. Song, J., Zhang Z.: Oil refining enterprise performance evaluation based on DEA and SVM. In: Proceedings of the Second International Workshop on Knowledge Discovery and Data Mining, pp. 401–404 (2009)
6. Yu, J., Wang, Y., Chen, P., Chen, P.: Fetal weight estimation using the evolutionary fuzzy support vector regression for low-birth-weight fetuses. IEEE Trans. Inf. Technol. Biomed. 13(1), 57–66 (2009)
7. Li, D., Liu, C., Gan, W.: A new cognitive model: cloud model. Int. J. Intell. Syst. **24**(3), 357–375 (2009)
8. Jia, S., Mao, B.: Research on CFCM: car following model using cloud model theory. J. Transp. Syst. Eng. Inf. Technol. **7**(6), 67–73 (2007)
9. Wang, S.X., Zhang, L., Wang, S.A., et al.: A cloud-based trust model for evaluating quality of web services. J. Comput. Sci. Technol. **25**(6), 1130–1142 (2010)

Chapter 33
A Game Theory Based MapReduce Scheduling Algorithm

Ge Song, Lei Yu, Zide Meng and Xuelian Lin

Abstract A Hadoop MapReduce cluster is an environment where multi-users, multi-jobs and multi-tasks share the same physical resources. Because of the competitive relationship among the jobs, we need to select the most suitable job to be sent to the cluster. In this paper we consider this problem as a two-level scheduling problem based on a detailed cost model. Then we abstract these scheduling problems into two games. And we solve these games in using some methods of game theory to achieve the solution. Our strategy improves the utilization efficiency of each type of the resources. And it can also avoid the unnecessary transmission of data.

Keywords Multi-level scheduling · Task assignment · Resource utilization efficiency · Bidding model · The Hungarian method · Game theory · MapReduce · Hadoop

G. Song · L. Yu (✉)
Ecole Centrale de Pekin, Beijing University of Aeronautics and Astronautics, Beijing, China
e-mail: yulei@buaa.edu.cn

G. Song
e-mail: sophiesongge@gmail.com

Z. Meng · X. Lin
School of Computer Science and Engineering, Beijing University of Aeronautics and Astronautics, Beijing, China
e-mail: mengzide@act.buaa.edu.c

X. Lin
e-mail: linxl@act.buaa.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_33,

33.1 Introduction

MapReduce [1] is a programming model designed by Google for processing large scale data sets parallel and distributed. It provides a simple and powerful way to let the programs run in a distributed environment automatically. Apache Hadoop [2] is an open source implementation of MapReduce. It consists of a distributed file system named HDFS (Hadoop Distributed File System), and a MapReduce programming framework. Due to the simplicity of the programming model and its elastic scalability and fine-grained run-time fault tolerance [3], Hadoop is popular among the commercial enterprises, the financial institutions, the scientific laboratories and the government organizations.

Generally, a Hadoop job will be scheduled according to three steps: user level, job level and task level. And usually, in a Hadoop cluster three types of resources will be consumed, they are: CPU, Disk and Network.

This paper analyses the advantages and the shortcomings of the existed Hadoop scheduling strategies, and designs a new scheduling algorithm. This algorithm firstly meets the multi-level nature of the Hadoop scheduling environment, to minimize the average waiting time per-user. Then it finds a balance point between "transfer data" and "wait" when running map tasks, to optimize the global cost of all the map tasks.

33.2 Related Work

33.2.1 The Analysis for the Existing Hadoop Scheduling Algorithms

In order to provide convenience to the users, the scheduler is designed as a pluggable model, so that users can design their own scheduler under their needs. And at the same time, Hadoop framework also integrates three schedulers to be invoked by users.

Firstly, the scheduler by default uses the FIFO strategy. Capacity Scheduler [4] and Fair Scheduler [5] are also integrated in a Hadoop framework. Capacity Scheduler wants to prevent the resources being exclusive. Fair Scheduler's purpose is to make sure that the resources will be allocated fairly to each job.

At the same time, there also appear many schedulers proposed by the users aiming at different application scenarios:

1. The first type wants to ensure the data locality [6, 7]. It comes to a conclusion that there is a conflict between fairness and data locality [6]. In order to solve this problem, it gives an algorithm called Delay Scheduling [6]. Its mean idea is: instead of launching a job according to fairness whose tasks cannot be run locally, it chooses to let it wait for a small amount of time, and let other jobs launch their tasks locally. But this paper did not give a balance between

"waiting" and "transmission" from the point of view of the global cluster. Another method to avoid unnecessary data transmission is: to assign tasks to a node, local map tasks are always preferred over non-local map tasks, no matter which job the task belongs to [8]. This method can guarantee the Data Locality in a certain extent, but it greatly impairs the levels of Hadoop scheduling.
2. The second type is deadline scheduling [9]. Its purpose is to allocate the appropriate amount of resources to the job so that it meets the required deadline.
3. The third type is called performance-driven scheduling. It wants to dynamically predict the performance of the tasks, and uses the prediction to adjust the allocation of resources for the jobs [10].

33.2.2 Cost Model

The cost models are usually used in task assignment and resource allocation. There are three common kinds of cost models. The first one doesn't consider the steps within a task, they simply think that the complete time of a local task is 1, and that of a remote task is 3. This kind of cost model is usually used in approximation algorithms, such as [11]. The second one is called a "non-accurate" model. This kind of model usually assumes that the cluster is isomorphic, and the running times of the tasks are equal. They use the performance of an already run task or the history trace to predict the execution time of a new task. Deadline scheduling often uses this type of cost model.

The third kind of model divides the cost as Disk IO, CPU and network transfer cost and so on. Someone gives a method to quantitatively describe each kind of cost of a task [12]. We can use it to effectively and accurately predict the performance of a task. The scheduling algorithm in our paper is based on a prediction of the costs of the tasks according to [12].

33.2.3 Game Theory Used for Scheduling in Cloud

Game Theory studies the balance when the decision-making bodies interact among each other under a related constraint. It solves an optimization problem of vectors which contains the target vector and the strategy vector. It is originally used in economics, but recently its approaches are always successfully used in resource allocations in cloud environment [13].

The scheduling problems of a distributed environment involve complicated optimization requirements. In solving this kind of problems, traditional methods usually combine the individual optimizations as a solution. And it only cares about the individual rather than the entirety. For a scheduling system in a Hadoop environment, the multi-level nature as well as the diversity of the optimization objectives makes a game theory method obviously better to solve this problem.

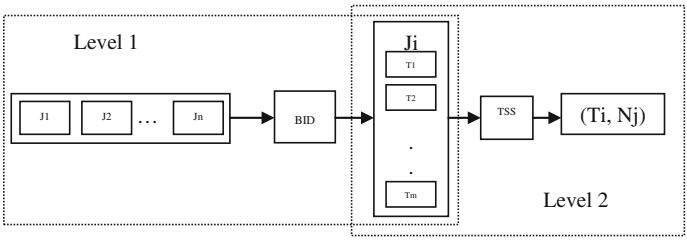


Fig. 33.1 Scheduling Schema

33.3 Problem Statement

After researching, we found the Hadoop scheduling environment has two natures below:

1. Multi-Level: We can schedule the Hadoop jobs by three levels: user level, job level and task level.
2. Consuming of multi-resource: Running a Hadoop job will consume three kinds of resources: CPU, Disk IO and Network IO. Other scheduling strategies only consider the allocation of CPU resources. But through some benchmarks we found that sometimes the random Disk IO will give a greater impact for the completion time of a job.

So we hope to design a scheduling algorithm, which can meet the following characteristics:

1. It can meet the needs of multi-level.
2. It should improve the utilization of the multi type of resources.
3. It has to find a balance point between "transfer" and "wait".

To fit the purpose above, we design a 2 level scheduling algorithm as shown in Fig. 33.1. Then we will present these 2 level scheduling algorithms respectively.

33.3.1 Level 1: Jobs Scheduling (with the Information of User)

To simplify the problem we choose to add the user information into the definition of a job, so that we do not have to consider about the user level scheduling separately.

Definition 1 A MapReduce Job Queue (Q):

A $Q(j_1, j_2, \ldots, j_q)$ is a vector which represents the set of submitted jobs. j_i means each job in Q.

Let $q = |Q|$, which means the number of jobs in the current Q.

Let j_i_user be the user who submits j_i.

Q dynamically changes when a job comes to Q or when a job leaves Q.

Definition 2 Total User Time: The sum of the time of all the jobs submitted by this user.

Wait Time: The time that the job has waited after it was submitted.

Estimate Exec Time: The execution time of the job estimated by the cost model.

According to the analysis above, we can give the definition of the "bid" given by a job to represent its priority of being executed.

Definition 3: Bid

$$\text{bid} = \frac{1}{\text{TotalUserTime}} \times \frac{\text{WaitTime}}{\text{EstimateExecTime}}$$

$\frac{1}{\text{TotalUserTime}}$ is designed for the fairness. And $\frac{\text{WaitTime}}{\text{EstimateExecTime}}$ is used to reduce the average waiting time of the jobs.

Game 1: Hadoop Job Scheduling Game

The target of this game is to choose a suitable job to be executed next.

To achieve this target we choose the bidding model, which is a dynamic non-cooperative game. There is a competitive relationship among the jobs. They compete for resources. Every job wants to be executed as soon as possible. The bid defined in definition 3 can reflect the trend of our scheduling purpose. So we let each job in Q submit a bid to JobTracker according to definition 3, then the job with the highest bid will be sent to the cluster and run.

33.3.2 Level 2: Tasks Scheduling

Tasks scheduling is a cooperative game because we want to make the execution cost of the job minimum which means we should take the tasks as an entirety. Our goal is to reduce the global complete cost of all the map tasks. To reduce this cost means to reduce the depletion of the resources of the cluster.

To achieve this goal, we give some definitions below:

Definition 4 A MapReduce Job (MR-Job)

A MR-Job is a pair (T, N) where T is a set of tasks, and N is a set of nodes.

Let $m = |T|$, and $n = |N|$, which means, m is the number of tasks in T, and n is the number of nodes in N.

Definition 5: A tasks scheduling strategy (TSS)

A TSS is a function $X: T \rightarrow N$ that assigns each task to a node n.

We define also a x_{ij}, if $X(j) = i$, then $x_{ij} = 1$, if not, $x_{ij} = 0$. Which means, if assign task j to node i, then $x_{ij} = 1$, if not, $x_{ij} = 0$.

Definition 6: A Cost Function C:

Let C(i, j) = (C, D, N) be the cost vector of running task j on node i, with C, D, N represent the cost of CPU, Disk I/O, and Network I/O separately.

Let $c_{ij} = c \times C + d \times N \times D + n$, where c, d, n are the weights of C, D, N. These weights will vary because of the type of the jobs. If the job is compute-intensive, then c will be bigger. If it is a data parsing job, then d and n will be bigger.

We define c_{ij} the cost of running task j on node i, we call the matrix c_{ij} a cost matrix. And the cost matrix c_{ij} is called the parameter matrix for Hungarian Method.

Definition 7: Total cost under a TSS:

$$Z = \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} \times x_{ij}$$

Game 2: Hadoop Tasks Scheduling Game

In this game we have to find an assignment for map tasks to minimize the execution cost of the entire map tasks. This is an assignment problem, which is a static cooperative game. In an assignment problem, we have m tasks, and they will be completed by n nodes. And we have to find a task scheduling scheme that let the total cost be the least. If we transfer this target into mathematics, it means to find a X(t), that minimize Z, where X(t) can represent an assignment.

We choose Hungarian Method [14] to solve this problem. And we translate this problem into mathematics:

$$\min z = \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} x_{ij} \tag{33.1}$$

$$s.t = \sum_{i=1}^{n} x_{ij} = 1 \quad i = 1, 2, \ldots, n \tag{33.2}$$

$$\sum_{j=1}^{m} x_{ij} = 1 \quad j = 1, 2, \ldots, n \tag{33.3}$$

$$x_{ji} = 0 \text{ or } 1 \tag{33.4}$$

The second equation means each task can only be executed by one node. And the third one means each node can only execute one task at one time. These 2 equations are the constraints for this assignment problem.

Theorem 1: According to the nature of the matrix multiplication, z won't change if we plus or minus a constant to all the elements in a row or in a column of c_{ij}.

Which means that there won't be any differences to the assignment if we plus or minus a same constant to a row or a column in c_{ij}.

Proof: Suppose $c'_{ij} = c_{ij} \pm (u_i + v_j)$, then,

$$Z' = \sum_{i=1}^{n}\sum_{j=1}^{m} c'_{ij}x_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}x_{ij} \pm \left(\sum_{i=1}^{n}\sum_{j=1}^{m} u_i x_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m} v_j x_{ij}\right)$$

x_{ij} is a vector with the value of 0 or 1, so the above equation is equal to:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}x_{ij} \pm \left(\sum_{i=1}^{n} u_i + \sum_{j=1}^{m} v_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}x_{ij} + K = Z + K$$

K is a constant. So the X_u which makes Z reach the minimum will also makes Z' minimum.■

The Hungarian Method depends on this theorem. So we can solve the problem as follows:

Step 1: Input a cost matrix c_{ij}.
Step 2: Find the smallest values in each column, for each element in this column, minus this smallest value, repeat for each row.
Step 3: Check in each column, mark the first 0 in this column, the other 0 will be deleted.
Step 4: Check whether the number of 0 is equal to that of tasks. If equal, break. If not, use the least lines to cover all the 0, and then delete the elements being covered.
Step 5: Find the minimum value in the elements non-deleted, and for all the rows which contains the elements non-deleted, minus this minimum. A new matrix will be formed after doing this. Come to step 3 for this new matrix.

33.4 Experiment Studying

We intend to use simulations to verify the effectiveness of our algorithm.

Average Waiting Time per User:

We generate a set of jobs, with the number from 30 to 200, and schedule them with both FIFO policy and BID policy proposed in our paper. Then we record the average waiting time per job as Fig. 33.2:

And the average waiting time per user is shown as Fig. 33.3:

From these 2 figures we can see that both the average time per job and per use of Bid method given in this paper are shorter than those of FIFO. And with the increase in the numbers of jobs, this gap becomes more and more obvious.

Fig. 33.2 Compare the bid method with FIFO for the average waiting time per job

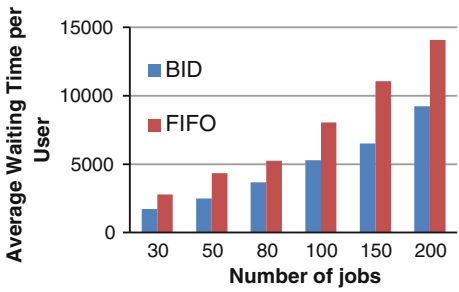


Fig. 33.3 Compare the bid method with FIFO for the average waiting time per user

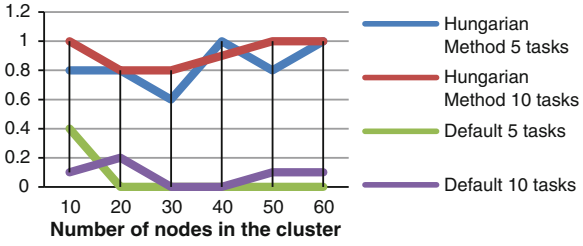


Fig. 33.4 Data Locality Rate

Data Locality Rate:

We suppose that the number of tasks run locally is l, and the number of tasks run remotely is r. Then we can define a data locality rate as: DLR $= \frac{l}{r}$. This rate describes the degree of tasks' localization. In this simulation we suppose there are 3 replicas for each input split, stored in 3 nodes separately. We execute 5 tasks and 10 tasks separately in the cluster with the number of nodes vary from 10 to 60. Compared with the scheduling strategy in Hadoop by default, we can see the data locality rate as Fig. 33.4:

We can see the method we proposed is much better than that by default in Hadoop, especially when the number of nodes is much bigger than that of tasks.

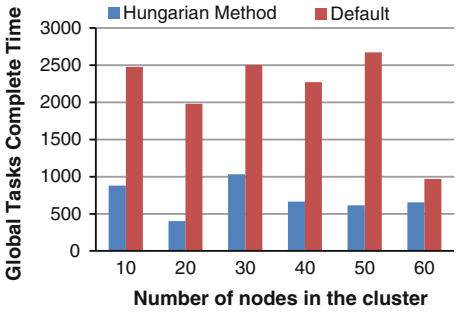


Fig. 33.5 Global complete time for 5 tasks

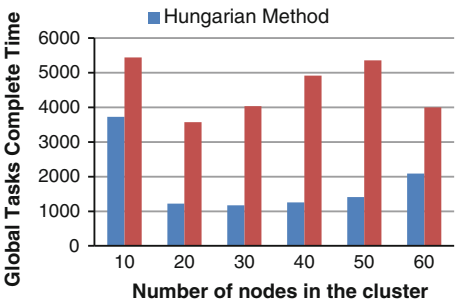


Fig. 33.6 Global complete time for 10 tasks

Global Tasks Complete Time:

In this part, we want to measure the global complete time of all the tasks for a job which contains 5 map tasks (Fig. 33.5), and another job which contains 10 map tasks (Fig. 33.6).

From these 2 pictures we can tell that our method will take much less time for complete all the tasks than the scheduling strategy by default.

33.5 Conclusion

Through analyzing the performance of the scheduler model of Hadoop, we find some problems and give a game theory based method to solve these problems. We divide a Hadoop scheduling problem into two steps—job level and task level. For the job level scheduling we choose to use a bid model, and we define this bid in order to guarantee the fairness and reduce the average waiting time. Then for tasks level, we change this problem into an assignment problem and use the Hungarian Method to optimize this problem. At last we do some simulation experiences to prove the efficiency of our algorithm.

For the further research, we want to improve the performance of our scheduling strategy in the following aspects: (1) Prove this algorithm in mathematics in using the methods of game theory. (2) Do experiences in a real cluster. (3) Automatically give the weights of the costs Etc.

Acknowledgments This work was supported by the National High Technology Research and Development Program of China (No. 2011AA010502) and the National Science and Technology Pillar Program (2012BAH07B01)

References

1. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. In: Proceedings of the Sixth Symposium on Operating System Design and Implementation, Usenix Association, San Francisco, 6–8 Dec 2004
2. Apache Hadoop. http://hadoop.apache.org
3. Jiang, D., Ooi, B.C., et al.: The performance of MapReduce: an in-depth study. Proc. VLDB Endow. **3**(1), 494–505 (2010)
4. Capacity Scheduler. http://hadoop.apache.org/common/docs/r0.20.2/capacity_scheduler.html
5. Fair Scheduler. http://hadoop.apache.org/mapreduce/docs/r0.21.0/fair_scheduler.html
6. Zaharia, M., Borthakur, D., et al.: Delay scheduling: a simple technique for achieving locality and fairness in cluster scheduling. In: Proceedings of the EuroSys' 10 (2010)
7. Zaharia, M., Borthakur, D. et al.: Job scheduling for multi-user MapReduce clusters. Technical Report, EECS Department, University of California, Berkeley (2009)
8. He, C., Lu, Y., et al.: Matchmaking: a new MapReduce scheduling technique. In: Proceedings of the CloudCom '11 (2011)
9. Verma, A., Cherkasova, L., et al.: ARIA: automatic resource inference and allocation for MapReduce Environments. In: Proceedings of the ICAC' 11 (2011)
10. Polo, J., Carera, D., et al.: Performance-driven task co-scheduling for MapReduce environments. In: Proceedings of the NOMS' (2010)
11. Fischer, M.J., Su, X., et al.: Assigning tasks for efficiency in Hadoop. In: Proceedings of the SPAA '10 (2010)
12. Lin, X., Meng, Z., et al.: A practical performance model for Hadoop MapReduce, Cluster Computing Workshops (CLUSTER WORKSHOPS), IEEE International Conference (2012)
13. Khan, S.U., Ahmad, I.: Non-cooperative, semi-cooperative, and cooperative games based grid resource allocation. In: Parallel and Distributed Processing Symposium, pp. 101 (2006)
14. Kuhn, H.W.: The Hungarian method for the assignment problem. Bryn Mawr College, Pennsylvania

Chapter 34
Dynamic USBKEY System on Multiple Verification Algorithm

Yixiang Yao, Jinghua Gao and Ying Gong

Abstract On account of the closed products and other defective products in the current market, this paper puts forward and carries out the Dynamic USBKEY System. This system is based on Multiple Verification Algorithm and is able to verify the validity of users' identity in a high-strength dynamic channel. Firstly, the security of the entire system is based on the strength of the random key. The overall design and the adopted algorithm are open. Secondly, it can solve the problems within the channel, the verification method and the program's self-preservation. Thirdly, the system provides a more secure solution under the rapid programming mode. The developers can apply the system on their own programs through the opened cross-language interface. As a result, the development cycle can be shortened and the security strength of their program can be improved.

Keywords USBKEY · One-time encryption · Network security · Software security · Rapid programming

34.1 Introduction

The developers of application software have paid close attention to the encryption-protection of commercial software. In order to protect intellectual property and avoid piracy, a variety of encryption technologies have emerged. USBKEY, one of these technologies, has taken over the market with its extraordinary superiority [1].

The current market is dominated by the third (programmable) and the fourth generation (smart card) USBKEY systems. However, the seemingly high security strength of these systems depends on encrypting the structure of the system itself so that they are insecure when the structural information are let out or cracked. The

Y. Yao (✉) · J. Gao · Y. Gong
Department of Information Security, Computer Science and Technology Institute, Civil Aviation University of China, Tianjin, China
e-mail: bigyyx@gmail.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_34,

main methods to crack the USBKEY are as follows: (1) cloning or copying, (2) using the debugging tools to trace and decrypt, (3) using blocking procedure to modify the communication between the software and USBKEY so as to acquire the communication data. All these attacking methods are great threatens to the USBKEY systems.

The USBKEY systems are usually used in protecting some specific software or system like CAD and there is no commonly used USBKEY system that can adapt to every user's application. It is necessary to design a new type of USBKEY system which is more secure, universally adaptable and easily installed.

In order to solve the problems described above, this paper comes up with a Dynamic USBKEY System based on Multiple Verification Algorithm. In this system, Multiple Verification is embodied in the diversity of authentication tokens (such as the digital certificates) and the dynamic feature manifests in one-time encryption used in the channel of transmission.

34.2 System Structure and the Principle of Module Design

The design of the USBKEY system depicted in this paper follows the principles shown below: (1) it can be quickly put into use by developers (2) the function of it can be expanded according to users' demand (3) its structure is open and the security strength is entirely based on the random key in transmission (4) the security of channel (5) the safety of its components can be verified (6) the authentication tokens have multiple dimensional patterns.

The entire USBKEY system includes four components: application programs (CreateKey Application, which is designed to create keys, VerifyKey Application, which is designed to verify the created keys, ConfMgr Application, which is designed to manage configurations), USBKEY hardware, the database of the server and PwdGuard Control (the security password control which is used as user interface).

Based on the four components, the system is divided into two states, namely Creating State and Verification State. The Creating State is designed to initialize the USBKEY and update the database. The Verification State mainly deals with cross-validation and controls users' action as is shown in Fig. 34.1. When a user tries to operate the application and meet the verification point, the Verification State will be triggered.

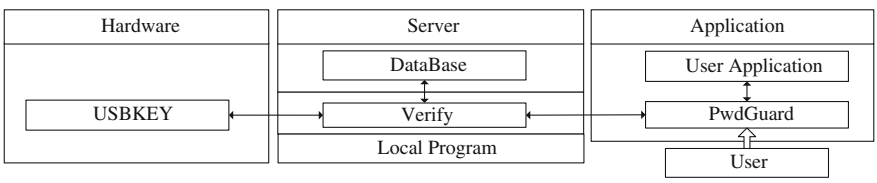


Fig. 34.1 The main structure of the Verification State

All communications of the system are based on their own communication protocols. Therefore, users can develop more plug-ins according to their own demand or adopt rapid programming by using PwdGuard.

34.2.1 Design of Dynamic Transmission Channel

It is generally believed that the security of transmission channel depends on the complexity of the protocol [2]. On contrary, cryptographers believe that any algorithm based on the complexity of protocols can be inducted and conjectured through statistical regularities. However, there is one exception that no one-time encryption can be cracked down even with infinite computing resources [3].

The one-time encryption scheme of the channel is based on the key exchange algorithm under Public-Key Encryption Infrastructure (PKI) and a symmetric encryption algorithm of higher security level. However, Man-In-The-Middle Attacks exist in some of the key exchange algorithms so that the channel needs a trust-worthy Certification Agency (CA) to issue certificates. Those CAs mentioned are beyond the protocols [3]. Since the USBKEY hardware only communicates with VerifyKey, the USBKEY cannot build a CA and is impossible to adopt a complete procedure of exchanging keys directly. Here, we put forward a scheme independent of certificates and CAs: assume that the Creating State is safe, PK and SK, a pair of public and private keys created in the Creating State, are saved in database and USBKEY respectively. Then the Creating State will use the PK and SK directly to complete the key exchange.

For the sake of safety, we set a safety time threshold t. When a time period exceeds t, the VerifyKey will generate a new pair of PKnew and SKnew which will be updated to USBKEY by the old pair of PK and SK.

The procedure is divided into four phases of communications as below:

1. VerifyKey generates message p which will be encrypted into c by using the following formulas then VerifyKey will send c to USBKEY.

$$k = Random() \tag{34.1}$$

$$c = E_k(p) \tag{34.2}$$

2. USBKEY returns ACK after receiving the message.
3. After receiving ACK, VerifyKey changes k to k', then send it to USBKEY.

$$k' = PKA(k,\ SK) \tag{34.3}$$

4. USBKEY gets k and p via decryption.

$$k = PKA(k',\ PK) \tag{34.4}$$

$$p = D_k(c) \tag{34.5}$$

Then v, the data need to be sent back, will be encrypted into v' and then send back to VerifyKey.

$$v' = E_k(v) \tag{34.6}$$

Hereto, the entire procedure of the communications completes.

Besides, we use SSL directly to guarantee the communications between the database and the VerifyKey.

In order to fight against reply attacks, we need to use time-stamps with timeout range in all data packets. Here, we leave out unnecessary details.

34.2.2 The Design of Multiple Verification Algorithm

The traditional mode of token protection is possible to be cracked down. Therefore, it is essential to design a new kind of multi-dimensional transformation method for tokens.

In Creating State, CreateKey needs to produce several units of various verification tokens T_i. For all kinds of tokens, they have their own different ways to transform. In other words, $T_i' = F(T_i)$ in which F may be a one-way function. Then T_i and T_i' will be saved respectively in the USBKEY and the database.

Assume $\Omega\{T\}$ is the sample space of the token in traditional protection mode, $\Omega\{F\}$ is the sample space for the verification algorithm. Computed by this algorithm, the sample space will become $\Omega\{T_i\}$ (i = 1, 2…,n). The sample space computed by the verification algorithm will become $\Omega\{F_j\}$ (j = 1, 2,…,m). X, Y is random variables, then

$$\sum_{i=1}^{n} P(X = T_i) = P(X = T) \tag{34.7}$$

$$\sum_{j=1}^{m} P(Y = F_j) = P(Y = F) \tag{34.8}$$

So under the protection of Multiple Algorithm, the possibility of tokens which may be cracked down is:

$$P(X = T_i,\ Y = F_j) = P(X = T_i)P(Y = F_j) = \frac{1}{nm}P(X = T,\ Y = F) \tag{34.9}$$

Besides, tokens cannot be saved continuously in the memory and they can only be stored separately though self-defined structure of data.

34.2.3 The Design of Mutual Verification and Self-Protection

Since the USBKEY components may be replaced or modified, Mutual Verification is particularly important. Before transmission, both sides which are reciprocally the subject and the object should verify each other through hash verification. The information can only be transmitted when the verifications are passed.

The system also needs to prevent itself from decompiling and tracing. Without the anti-tracing technologies, the software will be exposed by the cracker using debugger and monitor [4]. The common protective measure in the market is packing. We recommend the high strength virtual machine packer which has a more remarkable protective ability.

In addition, we can adopt self-made methods to protect the core codes. For example, the common 0xCC breakpoint can be detected by acquiring the machine code during execution and we can also use the debugging mark in the memory of Windows to judge or avoid debugging etc.

34.3 Module Design

34.3.1 The Design of Program Module

Here is an overview of the design for CreateKey and VerifyKey. ConfMgr is used to manage the configuration. The related parameters can be redefined so as to be compatible with different environment.

1. CreateKey is responsible for Creating State. The main functions include: creating, transforming and saving the verification token, generating and storing the initial PKI parameters, writing the USBKEY and updating the database.
2. As the center of entire Verification State, VerifyKey is transparent to users. When PwdGuard receives a request, the Verify will transfer the request to USBKEY which will return the ID and the relevant token T. According to the ID, a duplication of verification token T' will be granted from the database. Finally, VerifyKey will compare T' and F(T) and use the result to generate execution event which will be executed by PwdGuard.

34.3.2 The Design of Hardware Module

As the carrier of the system's hardware, USBKEY is responsible for storing the authentication tokens and assisting VerifyKey to complete verification process. As is shown in Fig. 34.2, when data arrives at the communication interface, it should be decrypted and decoded by Cryptographer and Protocol Explanation respectively. Then the data has three routes: (a) acquiring a random number (b) being created or acquiring a verification token (c) calling other functions. Particularly, the operations of verification token and function must go through internal memory and then return. Here, the operation of verification token means the creation of verification token (Creating State) and the random reading of verification token (Verification State). While the function is used in two ways: (a) part of the function (used to running in VerifyKey) is executed in the USBKEY and it will return the result (b) the function only used by USBKEY itself.

34.3.3 The Design of Interface Module

The Interface Module is used to connect the user's applications, its function includes monitor and control the user's operation. Based on the ActiveX control, we designed a special InputBox to fight against detecting of asterisk password and KeyLogger. It also has the function of verification.

1. Measures to prevent asterisk password from being detected: under the Windows system, the password InputBox only changes password into asterisk. However, the cracker can use the Handle of the InputBox to get the password. Our new control is designed to avoid this problem by caching the password indirectly and forging a fake display, as is shown in Fig. 34.3a.
2. Ways to prevent the KeyLogger: Message Hook is a mechanism provided by Windows and it can allow the system to monitor the processing of Message [5].

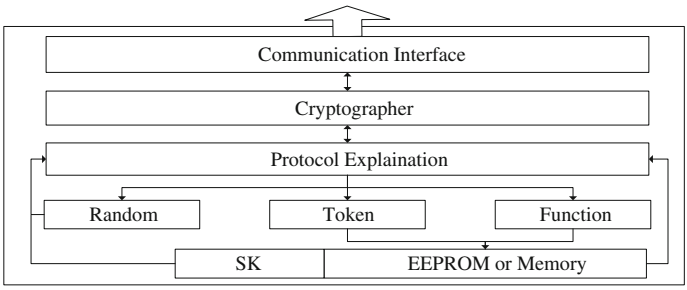


Fig. 34.2 The schematic of hardware module

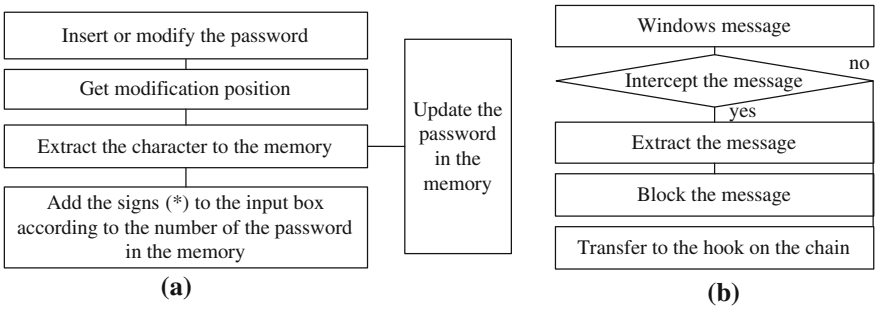


Fig. 34.3 Two processes of PwdGuard. **a** The process of anti-asterisk detect. **b** The process of anti-keylogger

When a new Hook is created, it will be placed at the top of the Hook Chain. As shown in Fig. 34.3b, we can load and unload Hooks repeatedly in small intervals to ensure that all Messages (the Messages of WH_KEYBOARD_LL and WH_DEBUG) will reach the top of the Hook Chain before being transmitted.

34.4 Experimental Tests and Comparison

We installed the database into a testing server and the other components in a common PC. Limited by the space of this paper, we only give the screenshots of the data flow and the detecting result of the asterisk viewer. In particular, Fig. 34.4 (left) shows the data flow in the transmission channel when the system is doing the same operation.

Table 34.1 presents the comparison between the USBKEY system introduced in this paper and a certain one from current market.

Fig. 34.4 Data flaw in transmission channel and Asterisk Viewer's detecting result

Table 34.1 Function comparisons between USBKEY systems

Function	Testing system	Comparison system
Channel transmission	Different data flow each time.	When execute the same function, the data flow is same.
Program modification	No disassembling information in OllyDbg1.1.	Disassembling the core code is possible.
Password detecting	The password cannot be cracked down through password viewer and KeyLogger.	The password can be cracked down.
Customizing components	Each part can be developed independently and allows self-defined interface.	Cannot be extended
System structure	Open, the security is based on random key.	Confidential, part of the security strength is based on protocols

34.5 Conclusion

In this study, researchers came up with the theory of Dynamic USBKEY System on Multiple Verification Algorithm which can reduce the threats of cracking and eavesdropping. Through the protocol's decoupled structure, self-verification and mutual verification, the system achieves custom extension while ensuring safety. Developers can deploy the system into their own applications through simple configuration. How to guarantee the safety of the system when the server is cracked down will be future top priority in researches.

Acknowledgments This work is supported by National Training Programs of Innovation and Entrepreneurship for Students No. 201210059060. We wish to thank Prof. Guo Li for his valuable instructions on earlier drafts of this paper.

References

1. Li, M., Shen, T.: Design of a USB software dog. Chin. J. Electron Devices **29**(1), 205–208 (2006)
2. Xu, M., Zhuge, Z.: Reliability analysis and design of a USB softdog. Chin. Mech. Electr. Eng. Mag. **24**(7), 47–49 (2007)
3. Schneier, B.: Applied cryptography-protocols, algorithms, and source code in C, 2nd edn. China Machine Press, Beijing (2004)
4. Duan, G.: Encryption and decryption, 3rd edn. Chinese Publishing House of Electronic Industry, Beijing (2008)
5. Richter, J.: Programming application for Microsoft Windows, 4th edn. Microsoft Press, Washington (2000)

Chapter 35
Anomaly Detection Algorithm Based on Pattern Density in Time Series

Mingwei Leng, Weiyi Yu, Shuai Wu and Hong Hu

Abstract Anomaly detection in a time series has attracted a lot of attentions in the last decade, and is still a hot topic in time series mining. In this paper, an anomaly detection algorithm based on pattern density is proposed. The proposed algorithm uses the anomaly factor to identify top k anomaly patterns. Firstly, a time series is represented based on its key points. Secondly, the represented time series is partitioned into patterns set. Thirdly, anomaly factor of each pattern is calculated, and anomaly factor is presented to measure the anomalous degree of a pattern by taking into account the characters of its neighbors. Finally, Top k anomaly patterns are identified. The effectiveness of the anomaly detection algorithm is demonstrated with standard and artificial time series, and the experimental results show that the algorithm can find out all anomaly patterns with different lengths.

Keywords Data mining · Time series · Anomaly factor · Anomaly patterns

35.1 Introduction

Recently, the increasing use of time series, has initiated various research and development attempts in the field of data mining. Anomaly pattern detection in a time series has extensive uses in a wide variety of application such as in business, industry, medicine, science, stock market or entertainment. Anomaly detection is still an important and hot topic. Most of time series are high dimensional and feature correlational, detecting anomaly patterns directly in such data is very expensive. In addition, the characters of its neighbors are the same as that of a normal pattern, and the relation of a pattern and its neighborhoods should be

M. Leng (✉) · W. Yu · S. Wu · H. Hu
School of Mathematics and Computer, Shangrao Normal University, Shangrao, China
e-mail: lengmw@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_35,

considered in the process of identifying anomaly patterns. The anomaly patterns which are identified by considering the relation of their neighbors should be more meaningful. Although, there have been many representation method for time series, we use key point method to compress time series before identifying anomaly patterns. As far as we know, the KNN rule has not been applied into anomaly pattern detection. In this paper, we present anomaly factor based on pattern density to measure anomalous degree of a pattern by taking in account its neighbors.

The rest of this paper is organized as follows. Section 35.2 gives an overview of related works on anomaly detection. Section 35.3 gives a few definitions. In Sect. 35.4, we present anomaly detection algorithm based on pattern density. Section 35.5 aims at demonstrating the effectiveness of our methods with standard and artificial time series. In the last section, we conclude this paper.

35.2 Related Works

There has been an extensive study on mining time series in the last decade. Many high level representations methods have been proposed, and we only introduce some of the most commonly used representations. The first representation method we discussed is Discrete Fourier Transform (DFT) [1], which transforms a time series from time domain into frequency domain. A few representation methods are proposed in recent years [2–5]. Yi and Faloutsos [3] uses PAA (Piecewise Aggregate Approximation) to achieve the goal of symbolizing time series. The more popular method is SAX [4]. But these two methods do not fit the time series which change fast. Fu et al. propose a representation method which suits for financial time series [5], Leng et al. propose a re-representation method based on key points for time series [2]. And Chen et al. propose a novel warping distance [6], which can measure similarity of patterns better compared with other methods, and the authors use it to detect patterns in streaming time series. More representation methods for time series are summarized in [7].

Anomaly detecting focuses on discovering the anomaly behaviors of time series, and it is still a challenging topic. Ma et al. utilize support vector regression to detect anomaly observations in a long time series [8]. Keogh et al. propose algorithm of discord detection [9], and Keogh et al. have developed a number of techniques for discord detection [10–12]. The shape information is one of most important characters in time series, and some detecting techniques are proposed based on shapes [12–14]. Wei et al. introduce the new problem of finding shape discord in large image databases [12]. Leng et al. use anomaly factor to measure the degree of anomaly patterns [13]. Liu et al. use HHM and dynamic programming to segment time series and detect anomaly in a large collection of shapes [14].

35.3 Formal Definitions

This section gives some definitions which will be used in this paper.

Definition 1 Key point. Given a series $T = t_1, t_2, \ldots, t_m$, and $t_i, t_{i+1}, \ldots, t_j$ are contiguous points, the maximum or minimum t_{i1} of $t_i, t_{i+1}, \ldots, t_j$ is called a key point of T if and only if t_{i1} satisfies the following conditions, where $1 \leq i, j \leq m$.

(i) $t_{i1}$is t_1 (the first points of time series T) or,
(ii) $t_{i1}$is t_m (the last points of time series T) or,
(iii) let t_{key1} and t_{key2} $(key1 < i, j < key2)$ are the closest key points, and t_{i1} holds that $(t_{i1} - t_{key1})^*(t_{key2} - t_{i1}) < 0$.

Definition 2 Pattern density. Given a pattern p, its density $\omega(p)$ is defined as,

$$\omega(p) = \frac{\sum_{q \in N(p)} |N(p) \cap N(q)|}{k} \tag{35.1}$$

where $N(p)$ is constructed by the k nearest neighbors of p, and symbol $|N(p)|$ denotes the number of patterns in $N(p)$.

Definition 3 $k\_dis(\bullet)$. Given a pattern set P and a pattern p in P, $k\_dis(\bullet)$ is the distance between p and its k-th nearest neighbor.

Definition 4 Anomaly factor. Given a pattern p, anomaly factor of p is defined as,

$$f(p) = \frac{k\_dis(p)}{\omega(p)} \tag{35.2}$$

35.4 Anomaly Detection Algorithm Based on Pattern Density

Most of the anomaly detection algorithms require giving the lengths of anomaly patterns before detecting, but the lengths of anomaly patterns are unknown sometimes. How to obtain the length of an anomaly pattern is a problem. In this section, we use quadratic regression model to segment the represented time series into patterns, and it can get the lengths of patterns automatically. We adopt DTW (Dynamic Time Warping) to calculate the dissimilarity between patterns, and top k anomaly patterns are identified based on the anomaly factor of each pattern. DTW is an efficient method of measuring dissimilarity between time series or patterns, and more detailed information is given in [15]. The anomaly detection algorithm is given as algorithm 1.

Algorithm 1: Anomaly pattern detection based on pattern density.

1. Input: time series T, the threshold of quadratic regression ε_1, self-similarity threshold of pattern ε_2, the value of k_1 in KNN and the number of anomaly pattern k_2.
2. Use key points to represent T, the represented time series is T_{key}, $T_{key}(i)$ denotes the position of i-th key point in T.
3. Segment T_{key} into pattern set.
   - 3.1. Let $s_1 = 1$ denote the first point of the first pattern in T_{key}, $l = 4$ denote initial length of each pattern, calculate quadratic regression function f.
   - 3.2. If $\left|\sum_{j=s_1}^{s_1+l-1}(f(j) - T(T_{key}(j)))\right| < \varepsilon_1$, let $l = l + 1$, recalculate quadratic regression function f, else let $count = count + 1$($count$ initial value is zero), let $e_1 = s_1 + l - 1, pattern(count) = (s_1, e_1)$
   - 3.3. Let $j = 1$, if $DTW(T(T_{key}(s_1,\ e_1)),\ T(T_{key}(s_1 + j,\ e_1 + j))) \le \varepsilon_2$ Let $j = j + 1$, recalculate $DTW(T(T_{key}(s_1,\ e_1)),\ T(T_{key}(s_1 + j,\ e_1 + j)))$ until it larger than ε_2 or $j \ge e_1 - s_1$.
   - 3.4. Let $s_1 = s_1 + j,\ l = 4$, goto step 3.2.
   - 3.5. Repeat above process until the end point of some pattern is the last value in T_{key}.
4. Find out k_1 nearest neighbors for each pattern, let $KNN(pattern)$ save these neighbors.
5. Use $f(p) = k\_dis(p)/\omega(p)$ and $KNN(pattern)$ to calculate anomaly factor for each pattern, let $anomaly(pattern)$ denote anomaly factor set, $anomaly(pattern(i))$ is anomaly factor of the i-th pattern.
6. Find out k_2 anomaly patterns.
   - 6.1. Find out the first anomaly pattern with max value in $anomaly(pattern)$, and reset its anomaly factor to zero.
   - 6.2. Find out the max value v in $anomaly(pattern)$, suppose v is anomaly factor of pattern p.
   - 6.3. If p overlaps with some existing anomaly pattern, then combine them into one and goto step 6.2, else add p into anomaly pattern set and goto step 6.2.
   - 6.4. Repeat step 6.2 and 6.3 until finding out k_2 anomaly patterns.
7. Output anomaly pattern set.

Algorithm 1 adopts quadratic regression method to segment the represented time series T_{key} into pattern sets. We adopt the overlapping method to partition T_{key}. If two adjacent patterns are similar, then they are called self-similar patterns, and one of them is redundant and we delete it. In the step 3, if and only if the dissimilarity of adjacent patterns is larger than ε_2, they can be appeared in pattern set.

The core of algorithm 1 is the step 4–6. Calculating the density and anomaly factor of each pattern requires finding out k nearest neighbors for all patterns. Step 5 utilizes the $KNN(pattern)$ to compute $anomaly(pattern)$. Step 6 identifies k_2 anomaly patterns based on $anomaly(pattern)$. Since some of adjacent patterns are

overlapping, then these adjacent overlapping patterns should be merged into one, and step 6 achieves this goal.

35.5 Experimental Results

In this section, we demonstrate our anomaly detection algorithm with both real life and artificial time series. The real life the time series are used in [4] and the mechanism of generating artificial time series is given in Sect. 35.2. Firstly, we use key points to represent time series. Secondly, segmenting the represented time series into patterns, and finally top k anomaly patterns are identified with anomaly factors.

35.5.1 Real Life Time Series

This section demonstrates our algorithm with two kinds of ECG time series. We set $k_1 = 6$, $\varepsilon_1 = 0.5$, and $\varepsilon_2 = 1.0$. The first ECG time series contains one anomaly pattern, and Fig. 35.1 shows the anomaly pattern in bold line.

The most anomalous pattern shown in Fig. 35.1 is the real anomaly pattern. And anomaly factor of the second most anomalous pattern is much less than that of the first most anomalous pattern.

The anomaly pattern shown in Fig. 35.1 is very simple and 'clear' example. Figure 35.2 shows an ECG that has several different types of anomaly patterns. We identify the five most anomalous patterns and show them in bold line.

Anomaly pattern is the pattern whose characters are much different with that of the rest of patterns in many literatures, but degree of difference is difficult to measure. So we identify the most five anomalous patterns, and 1st-anomaly denotes the most anomalous pattern, 2nd-anomaly is the second anomalous pattern.

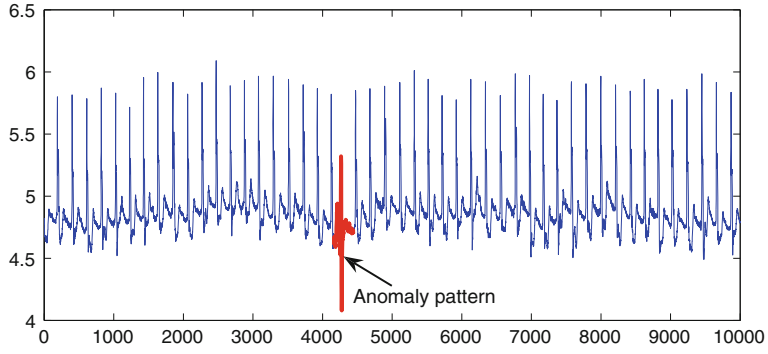


Fig. 35.1 An excerpt of an ECG that contains 1 anomaly pattern (highlighted in *bold line*)

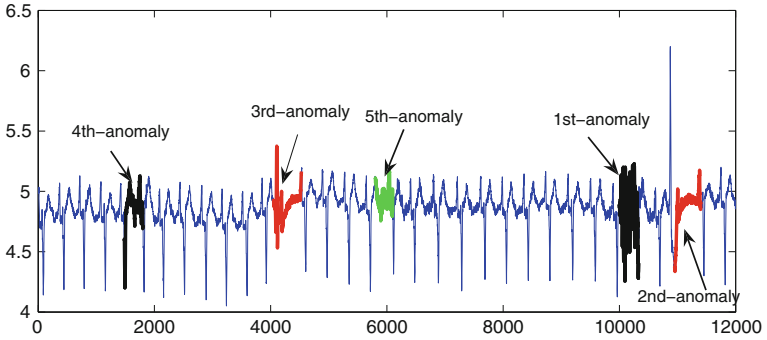


Fig. 35.2 An excerpt of an ECG that contains different types of anomaly patterns (highlighted in *bold line*)

35.5.2 Artificial Time Series

Artificial time series is generated from the following stochastic processes.

$$X(t) = \sin\left(\frac{40\pi}{N}t\right) + n(t) + e(t) \quad (1 \le t \le 3600) \tag{35.3}$$

where $t = 1, 2, \ldots, N$, $N = 1200$, and $n(t)$ is an additive Gaussian noise with zero-mean and a SDT of 0.1. $e(t)$ is defined as,

$$e(t) = \begin{cases} n_1(t) & 1001 \le t \le 1100 \\ 0 & otherwise \end{cases} \tag{35.4}$$

where $n1(t)$ follows a normal distribution of $N(0, 0.5)$.

$X(t)$ has one anomaly pattern, and it is identified with our detection algorithm. Anomaly factors of the rest of patterns are much less than that of this pattern. The anomaly pattern is plotted in bold line (Fig. 35.3).

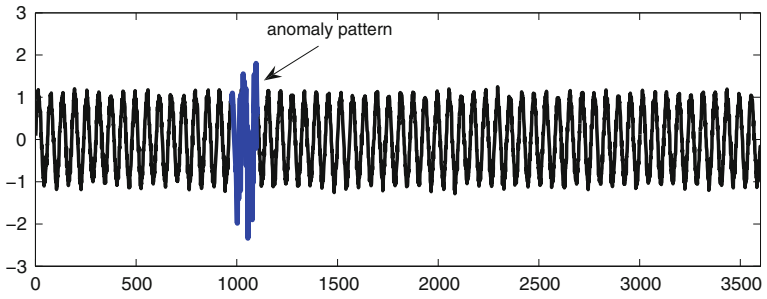


Fig. 35.3 An artificial time series that contains 1 anomaly patterns (highlighted in *bold line*)

35.6 Conclusion

This paper proposes a method of anomaly pattern detection algorithm. The most interesting of its contribution is that we introduce the KNN rule into anomaly pattern detection. Top $k2$ anomaly patterns are identified based on pattern density. Primitive experimental results demonstrate the promising performance of the proposed anomaly pattern detection algorithm. Meanwhile, many topics brought up by this paper are still open. Researchers need to give the threshold of quadratic regression ε_1. If its value is too large, then the anomaly factor of each pattern will be very large. Then detecting anomaly patterns is meaningless in the pattern set which is obtained with the value of ε_1. If its value is too small, then the algorithm also can not obtain meaningful anomaly patterns. The value of ε_1 is given before running algorithm 1, it is not self-adaptive. And this is our research works in the future.

Acknowledgments The authors would like to thank Nature Science Foundation of Jiangxi Education Department (GJJ11609), P. R. China.

References

1. Chan, K., Fu, A.W.: Efficient time series matching by wavelets. In: Proceedings of the ICDE, pp. 126–133 (1999)
2. Leng, M., Lai, X., Tan, G., Xu, X.: Time series representation for anomaly detection. In: Proceedings of the IEEE ICCSIT, pp. 628–632 (2009)
3. Yi, B.K., Faloutsos, C.: Fast time sequence indexing for arbitrary Lp norms. In: Proceedings of the VLDB, pp. 385–394 (2000)
4. Keogh, E., Lin, J., Fu, A.: Hot SAX: efficiently finding the most unusual time series subsequence. In: Proceedings of the ICDM, pp. 226–233 (2005)
5. Fu, T.C., Chung, F.L., Luk, R., Ng, C.M.: A specialized binary tree for financial time series representation. In: Proceedings of the KDD/TDM, pp. 96–103 (2004)
6. Chen, Y., Nascimento, M.A., Ooi, B., Tung, A.K.H.: SpADe: on shape-based pattern detection in streaming time series. In: Proceedings of the ICDE, pp. 786–795 (2007)
7. Fu, T.C.: A review on time series data mining. Eng. Appl. Artif. Intell. **24**(1), 164–181 (2011)
8. Ma, J., Perkins, S.: Online novelty detection on temporal sequences. In: Proceedings of the KDD, pp. 613–618 (2003)
9. Keogh, E., Lin, J., Lee, S.-H., Herle, H.V.: Finding the most unusual time series subsequence: algorithms and applications. Knowl. Inf. Syst. **11**(1), 1–27 (2006)
10. Bu, Y., Leung, T., Fu, A., Keogh, E., Pei, J., Meshkin, S.: Wat: finding top-k discords in time series database. In: Proceedings of the SDM, pp. 449–454 (2007)
11. Chuah, M., Fu, F.: ECG anomaly detection via time series analysis. In: Proceedings of the LNCS, vol. 4743, pp. 123–135 (2007)
12. Wei, L., Keogh, E., Xi, X.: Saxually explicit images: finding unusual shapes. In: Proceedings of the ICDM, pp. 711–720 (2006)
13. Leng, M., Chen, X., Li, L.: Variable length methods for detecting anomaly patterns in time series. In: Proceedings of the ISCID, vol. 2, pp. 52–56 (2008)
14. Liu, Z., Yu, J.X., Chen, L., Wu, D.: Detection of shape anomalies: a probabilistic approach using hidden markov models. In: Proceedings of the ICDE, pp. 1325–1327 (2008)
15. Keogh, E., Ratanamahatana, C.A.: Exact indexing of dynamic time warping. Knowl. Inf. Syst. **7**(3), 406–441 (2005)

Chapter 36
Integrative Optimal Design of Experiments with Multiple Sources

Hanyan Huang, Yuntao Chen, Mingshan Shao and Hua Zhang

Abstract Equivalent surrogate experiments are important information sources when the prototype experiment is limited to do. To explore the design methodology for the experiments with multiple sources, the optimal design of experiment for damage assessment is studied as an example and the following works are done. Firstly, the discrepancy between the prototype experiment and the four equivalent surrogate experiments, which are reduced scale test, test with surrogate drones, static test and simulation, is analyzed. Secondly, the parameter model about the discrepancy is constructed, and so does the fusion model. Thirdly, the integrative optimal design is developed, and then the iterative algorithm for constructing the integrative Dn-optimal design is also discussed. Lastly, an example about the integrative Dn-optimal design of the flight and static experiments about the projectile penetrating into the concrete is given, and the example shows that the proposed method is more efficient than the standard Dn-optimal design.

Keywords Damage response function · Integrative · D-optimal design · Discrepancy modeling · Fusion assessment

36.1 Introduction

Physics experiment or observation is the most reliable approach to investigate the performance of the machine or the mechanism of the nature. However, limited by cost, resource, time and some other reasons, the prototype experiment is often limited to do. In order to extend the information source, the equivalent surrogate experiment is usually done. Model experiment has always played an important role

H. Huang (✉) · Y. Chen · M. Shao · H. Zhang
Wuhan Mechanical Technology College, Wuhan, China
e-mail: hhy47822587@tom.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_36,

in the development of airplane, ship, vehicle and airship [1]. Researchers in America and Russia often use accelerated aging testing to assist the generic experiments to evaluate the reliability of some apparatus with high reliability such as spaceflight equipment and missile [1]. Wang did experiment utilizing a surrogate Radar with the same system capability and performance as the expected system [2]. The high ballistic test and the short range test are often implemented to substitute the long range ballistic to evaluate the precision of missile [3, 4]. Test on airplane is a very important stage to evaluate the reliability of the equipments on missile [5]. Static test [6], reduced-scale test [7] and experiment with surrogate drones [8] are the three common experiments to evaluate the lethality of the warhead. With the development of computer science, simulation is becoming a very important tool to assistant physics experiment.

As the environment or the target is different, the result of the equivalent surrogate experiment can not be used directly. In engineering application, before fusion and evaluation, the result of equivalent surrogate experiment should be transferred to the same condition with the prototype experiment via conversion or error compensation [3]. However, the weapon system is so complex that the conversion and compensation are not accurate. What is more, the sample size of the equivalent surrogate experiment is relatively larger, thus if the results are fused directly, the surrogate result will flood the prototype result. Define the credibility of the surrogate experiment, and then the fusion is an admissive method [9]. To increase the reliability of the evaluation, the experiment should provide as much information as possible. Thus, it is a trend that the design of experiment (DOE) be introduced to the small sample experiment [10]. However, the standard DOE aims to design only one kind of experiment [11]. In this paper, the experiment for damage assessment is taken as an example to discuss the integrative optimal design of experiments with multiple sources.

36.2 Discrepancy and Fusion Model for Damage Assessment

36.2.1 Discrepancy Model about the Equivalent Surrogate Experiments

The process of the warhead attacking the target involves not only the characters of the warhead and the target, but also the interaction between them. To simplify the analysis, Deitz, Klopcic, Starks and Walbert developed the V/L taxonomy and divided the Vulnerability space into four sub-spaces: the threat-target interaction initial conditions, the target component damage states, the target capability and the target combat utility [12]. The work of damage assessment is to get the mapping from one space to the next space.

A key problem in V/L taxonomy is to get the mapping from sub-space 1 to sub-space 2, which is called the damage response function and can be denoted as

$$y = f(\mathbf{x}) + \varepsilon = f(x_1, \cdots, x_p) + \varepsilon \tag{36.1}$$

where y express the physics damage of components, for instance, the damage area or the damage probability, etc. The vector $\mathbf{x} = \left(x_1, \cdots, x_p\right)'$ denotes the influence factors from the warhead, the environment and the target. To deduce the experience function, the flight test and the equivalent surrogate experiment such as the reduced scaled test, the static test, the test with surrogate drones and the simulation make up of the primary information sources.

Take the reduced scale test for example. Small warhead is used to simulate the real warhead based on the similitude theory. Dimension method is the most usual similitude principle. Due to it, if the system can be described by k independent dimensions, then it can also be denoted by s quantities without dimension, then

$$\Pi = g(\Pi_1, \Pi_2, \cdots \Pi_s) + \varepsilon \;\; s = p - k \tag{36.2}$$

Two systems are similar as $\Pi_i = \Pi_i^{'}(i = 1, \cdots, s)$. However, two systems which are similar in theory will not be exactly similar. Xu summed up the four reasons about the discrepancy [14]. In a whole, those factors can be separated into the random items and the items that can not be reduced. For instance, the error of measure and the uneven of the material are the random items, the gravity and some characters of the materials are the other kind. Therefore, while implementing the reduced scale test, the size of the scale model should be controlled to reduce the discrepancy, moreover, the discrepancy should be analyzed.

If the multiple of the scale is l, the item can not be reduced is Π_s, due to the Theorem of the mean, the real observation of the reduced scale test is

$$\begin{aligned} \Pi &= g(\Pi_1, \Pi_2, \cdots \Pi_s') + \varepsilon \\ &= g(\Pi_1, \Pi_2, \cdots \Pi_s) + \partial g(\Pi_1, \Pi_2, \cdots \tilde{\Pi}_s) / \partial \Pi_s \cdot \left(\Pi_s' - \Pi_s\right) + \varepsilon \end{aligned} \tag{36.3}$$

where $\tilde{\Pi}_s$ is between Π_s and Π_s'. If l is small or the response is non-sensitive to Π, the discrepancy between the reduced scale test and the prototype test will be also small. For simple, the observation of the reduced scale test with error can be denoted as

$$\Pi = g(\Pi_1, \Pi_2, \cdots \Pi_s) + \delta(\Pi_1, \cdots, \Pi_s, l) + \varepsilon' \tag{36.4}$$

Similarly, the observations about the other three kinds of test (static test, the test with surrogate drones and the simulation) can also modeled by

$$y = f(x_1, x_2, \cdots, x_p) + \delta(x_1, x_2, \cdots, x_p) + \varepsilon' \tag{36.5}$$

Formula (36.4) and (36.5) can also be denoted as

$$\Delta(\mathbf{x}) = y - \tilde{y} = \delta(x_1, x_2, \cdots, x_p) + \varepsilon' \tag{36.6}$$

As the damage process is so complex that deduction through the physics mechanism to get $\Delta(\mathbf{x})$ is hard to carry out. Thus, induction via the experiments is preferred. We can use parameter model to describe the discrepancy as

$$\Delta(\mathbf{x}) = \alpha'\delta(\mathbf{x}) + \varepsilon' \quad \varepsilon' \sim \mathrm{N}(0, \tilde{\sigma}^2) \tag{36.7}$$

In formula (36.7), the vector $\boldsymbol{\alpha} = (\alpha_1, \cdots \alpha_k)^T$ denotes $k \times 1$ parameters to be estimated. The vector $\mathbf{x} = (x_1, \cdots, x_p)^T$ denotes the controlled factors. The vector $\boldsymbol{\delta}(\mathbf{x}) = (\delta_1(\mathbf{x}), \delta_2(\mathbf{x}), \cdots, \delta_k(\mathbf{x}))^T$ denotes k independent linear regression models defined on a compact subspace Ω in $\mathbf{R}^p$, and ε' denotes the random discrepancy.

History observations and simulations are the main data sources to deduce Δ. As the observation y and $\tilde{y}$ on the same point $\mathbf{x}$ are often unavailable, we can firstly get the damage response functions on the two experiment state from the history observations, and then estimate the difference by mutual prediction.

Consider denoting the damage response function with a linear model, which is

$$y = \beta'\mathbf{f}(\mathbf{x}) + \varepsilon \tag{36.8}$$

where vector $\boldsymbol{\beta} = (\beta_1, \cdots \beta_m)^T$ denotes $m \times 1$ parameters to be estimated, the vector $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \cdots, f_m(\mathbf{x}))^T$ denotes m dependent linear regression models defined on a compact subspace Ω in $\mathbf{R}^p$. Uniting it with the discrepancy model, we can get the damage response function of the equivalent surrogate experiment as

$$\tilde{y} = \beta'\mathbf{f}(\mathbf{x}) + \alpha'\delta(\mathbf{x}) + \varepsilon' \tag{36.9}$$

With n_2 times of history observations, we can get the factual form of formula (36.8) and (36.9), and then the parameter model of the discrepancy is built. The estimator of the random error is $\tilde{\sigma}^2 = \sum_i (y_i - \hat{y}_i)^2 \Big/ n_2 - k$, where y_i and $\hat{y}_i$ are the observation and the prediction value of the equivalent experiment considering the discrepancy.

36.2.2 Fusion Model

Consider the parameter model with three kinds of experiments:

① Prototype experiment: $y = \beta'\mathbf{f}(\mathbf{x}) + \varepsilon_1 \quad \varepsilon_1 \sim \mathrm{N}(0, \sigma_1^2)$

② Surrogate experiment 1: $y - \Delta_2(\mathbf{x}) = \beta'\mathbf{f}(\mathbf{x}) + \varepsilon_2 \; \varepsilon_2 \sim \mathrm{N}(0, \sigma_2^2)$

③ Surrogate experiment 2: $y - \Delta_3(\mathbf{x}) = \beta'\mathbf{f}(\mathbf{x}) + \varepsilon_3 \; \varepsilon_3 \sim \mathrm{N}(0, \sigma_3^2)$

As for the K kinds of equivalent surrogate experiments, N_k times of observations are done respectively $\sum N_k = N$. Let $\boldsymbol{\theta}_2 = (\Delta_2(z_{21}), \Delta_2(z_{22}), \cdots, \Delta_2(z_{2N_2}))'$ be the discrepancy vector between the prototype and surrogate experiments 1, and $\boldsymbol{\theta}_3 = (\Delta_3(z_{31}), \Delta_3(z_{32}), \cdots, \Delta_3(z_{3N_3}))'$ be the discrepancy vector between the

prototype and surrogate experiments 2, the response vector is $\mathbf{Y}_k = (y_{k1}, y_{k2}, \cdots, y_{kN_k})'$, the vector that denotes the random error is $\mathbf{e}_k$, with $\mathrm{E}(\mathbf{e}_k) = 0$, $\mathrm{cov}(\mathbf{e}_k) = \sigma_k^2 \mathbf{I}_{N_k}$, Let

$$\mathbf{F}(\xi_k) = \begin{pmatrix} f_1(z_{k1}) & f_2(z_{k1}) & \cdots & f_m(z_{k1}) \\ f_1(z_{k2}) & f_2(z_{k2}) & \cdots & f_m(z_{k2}) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(z_{kN_1}) & f_2(z_{kN_1}) & \cdots & f_m(z_{kN_k}) \end{pmatrix} (k = 1, 2, 3) \tag{36.10}$$

Then the fusion model of all the observations is

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 - \boldsymbol{\theta}_2 \\ \mathbf{Y}_3 - \boldsymbol{\theta}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{F}(\xi_1) \\ \mathbf{F}(\xi_2) \\ \mathbf{F}(\xi_3) \end{bmatrix} \cdot \boldsymbol{\beta} + \begin{bmatrix} \mathbf{e_1} \\ \mathbf{e_2} \\ \mathbf{e_3} \end{bmatrix} \tag{36.11}$$

Let $\mathbf{Y} = \left(\sigma_1^{-1}\mathbf{Y}_1, \sigma_2^{-1}(\mathbf{Y}_2 - \boldsymbol{\theta}_2), \sigma_3^{-1}(\mathbf{Y}_3 - \boldsymbol{\theta}_3)\right)'$ and

$\mathbf{X} = \left[\sigma_1^{-1}\mathbf{F}(\xi_1), \sigma_2^{-1}\mathbf{F}(\xi_2), \sigma_3^{-1}\mathbf{F}(\xi_3)\right]'$, then the LSE of $\boldsymbol{\beta}$ is

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{36.12}$$

Where

$$\begin{aligned} \mathbf{X}^T\mathbf{Y} &= \begin{bmatrix} \sigma_1^{-1}\mathbf{F}(\xi_1) \\ \sigma_2^{-1}\mathbf{F}(\xi_2) \\ \sigma_3^{-1}\mathbf{F}(\xi_3) \end{bmatrix}^T \begin{bmatrix} \sigma_1^{-1}\mathbf{Y}_1 \\ \sigma_2^{-1}(\mathbf{Y}_2 - \boldsymbol{\theta}_2) \\ \sigma_3^{-1}(\mathbf{Y}_3 - \boldsymbol{\theta}_3) \end{bmatrix} \\ &= \sigma_1^{-2}\mathbf{F}^T(\xi_1)\mathbf{Y}_1 + \sigma_2^{-2}\mathbf{F}^T(\xi_2)(\mathbf{Y}_2 - \boldsymbol{\theta}_2) + \sigma_3^{-2}\mathbf{F}^T(\xi_3)(\mathbf{Y}_3 - \boldsymbol{\theta}_3) \end{aligned} \tag{36.13}$$

$$\begin{aligned} \mathbf{X}^T\mathbf{X} &= \begin{bmatrix} \sigma_1^{-1}\mathbf{F}(\xi_1) \\ \sigma_2^{-1}\mathbf{F}(\xi_2) \\ \sigma_3^{-1}\mathbf{F}(\xi_3) \end{bmatrix}^T \begin{bmatrix} \sigma_1^{-1}\mathbf{F}(\xi_1) \\ \sigma_2^{-1}\mathbf{F}(\xi_2) \\ \sigma_3^{-1}\mathbf{F}(\xi_3) \end{bmatrix} \\ &= \sigma_1^{-2}\mathbf{F}^T(\xi_1)\mathbf{F}(\xi_1) + \sigma_2^{-2}\mathbf{F}^T(\xi_2)\mathbf{F}(\xi_2) + \sigma_3^{-2}\mathbf{F}^T(\xi_3)\mathbf{F}(\xi_3) \end{aligned} \tag{36.14}$$

As for the K kinds of experiments, if the variance of the random error after conversion is σ_k^2, then the fusion estimator of $\boldsymbol{\beta}$ is

$$\hat{\beta} = \sum_k \sigma_k^{-2}\mathbf{F}^T(\xi_k)(\mathbf{Y}_k - \boldsymbol{\theta}_k) \Big/ \sum_k \sigma_k^{-2}\mathbf{F}^T(\xi_k)\mathbf{F}(\xi_k) \tag{36.15}$$

Conclusion 1 $\mathrm{E}\widehat{\beta} = \boldsymbol{\beta}$, $\mathrm{cov}\widehat{\beta} = \left(\sum_k \sigma_k^{-2}\mathbf{F}^T(\xi_k)\mathbf{F}(\xi_k)\right)^{-1}$.

Thus, once the equivalent surrogate experiment is fused, the precision of the estimator will be improved.

36.3 Integrative Optimal Design

36.3.1 Definitions

If there are K kinds of experiments, whose random error resulted by discrepancy modeling satisfies to $\varepsilon_k \sim \mathrm{N}(0, \sigma_k^2)$, the information matrix of the integrative design ξ^{c} composed by those experiments $\{\xi_k\}$ is

$$M(\xi^{\mathrm{c}}) = \sum_k \sigma_k^{-2} \mathbf{F}^T(\xi_k)\mathbf{F}(\xi_k) = \sum_k \sigma_k^{-2} \sum_{j=1}^{N_k} f(\mathbf{z}_j) f^T(\mathbf{z}_j) = \sum_{j=1}^{N} \lambda_j f(\mathbf{z}_j) f^T(\mathbf{z}_j) \tag{36.16}$$

where $\lambda_j = I(1, N_1)\sigma_1^{-2} + I(N_1 + 1, N_1 + N_2)\sigma_2^{-2} + \cdots + I(N - N_k + 1, N)\sigma_k^{-2}$. Apparently, if all σ_k^{-2} are the same, then the information matrix of the integrative design is the same as that of the standard design.

Definition 1 if the information matrix of the integrative design ξ^{c} is $M(\xi^{\mathrm{c}})$, we call the design $\xi_{\mathrm{D}}^{\mathrm{c}}$ an integrative D-optimal design, if $\xi_{\mathrm{D}}^{\mathrm{c}} = \arg\max \det M(\xi^{\mathrm{c}})$.

Definition 2 if the information matrix of the integrative design ξ^{c} and the integrative D-optimal design $\xi_{\mathrm{D}}^{\mathrm{c}}$ are $M(\xi^{\mathrm{c}})$ and $\mathbf{M}(\xi_{\mathrm{D}}^{\mathrm{c}})$ respectively, then the integrative D-efficiency of ξ^{c} is defined as

$$d^{\mathrm{c}} = \frac{|\mathbf{M}(\xi^{\mathrm{c}})|}{|\mathbf{M}(\xi_{\mathrm{D}}^{\mathrm{c}})|} = \left|\sum_k \sigma_k^{-2}\mathbf{F}^T(\xi_k)\mathbf{F}(\xi_k)\right| \Big/ \left|\sum_k \sigma_k^{-2}\mathbf{F}^T(\xi_{\mathrm{D}k})\mathbf{F}(\xi_{\mathrm{D}k})\right| \tag{36.17}$$

Besides the D-optimal design, the analyzers care more about the Dn-optimal design which can be used directly in application. As for a design with n support points $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n$($\mathbf{z}_i$ and $\mathbf{z}_j$ can be the same), each with a weight coefficient $1/n$, then it is called an exact design $\xi(n)$.

Definition 3 an integrative design $\xi_{\mathrm{D}}^{\mathrm{c}}(n)$ is Dn-optimal, if and only if $M(\xi_{\mathrm{D}}^{\mathrm{c}}(n))$ is nonsingular and $\left|M(\xi_{\mathrm{D}}^{\mathrm{c}}(n))\right| = \max_{\xi^{\mathrm{c}}(n)} |M(\xi^{\mathrm{c}}(n))|$.

36.3.2 Construction Algorithm

Consider getting the integrative Dn-optimal design via the singular point exchange method. Let $d(\mathbf{x}, \xi^{\mathrm{c}}) = \mathbf{f}^T(\mathbf{x})M^{-1}(\xi^{\mathrm{c}})\mathbf{f}(\mathbf{x})$. In a non-degenerated design $\xi(n)$, whose information matrix is $M(\xi(n))$, then replace $\mathbf{z}_i$ with $\mathbf{z}$ to get another design $\tilde{\xi}(n)$. The weight of $\mathbf{z}_i$ is denoted as $c(i)$, only related to the order i.

Conclusion 2 the determinant of $M\left(\tilde{\xi}(n)\right)$ which is the information matrix of $\tilde{\xi}(n)$ satisfies that $\left|M\left(\tilde{\xi}(n)\right)\right| = |M(\xi(n))| \cdot [1 + c(i)(d(\mathbf{z}) - d(\mathbf{z}_i)) - c^2(i)(d(\mathbf{z})$ $d(\mathbf{z}_i) - d^2(\mathbf{z}, \mathbf{z}_i))]$, where $d(\mathbf{z}, \mathbf{z}_i) = f^T(\mathbf{z}_i)M^{-1}(\xi^{\mathrm{c}}(n))f(\mathbf{z})$, $d(\mathbf{z}_i) = d(\mathbf{z}_i, \xi^{\mathrm{c}}(n))$, $d(\mathbf{z}) = d(\mathbf{z}, \xi^{\mathrm{c}}(n))$.

Proof $M\left(\tilde{\xi}(n)\right) = M(\xi(n)) - c(i)f(\mathbf{z}_i)f^T(\mathbf{z}_i) + c(i)f(\mathbf{z})f^T(\mathbf{z})$

Set $j = \sqrt{-1}$, as $\xi(n)$ is non-degenerated, $M(\xi(n))$ is a nonsingular matrix with the size $m \times m$, and $f(\mathbf{x})$ is a vector with the size $m \times 1$, such that

$$\det M\left(\tilde{\xi}(n)\right) = \det[M(\xi(n)) - c(i)f(\mathbf{z}_i)f^T(\mathbf{z}_i) + c(i)f(\mathbf{z})f^T(\mathbf{z})] =$$

$$\det\begin{bmatrix} M(\xi(n)) & c(i)f(\mathbf{z}_i) & jc(i)f(\mathbf{z}) \\ f^T(\mathbf{z}_i) & 1 & 0 \\ jf^T(\mathbf{z}) & 0 & 1 \end{bmatrix} = \det\begin{bmatrix} M(\xi(n)) & c(i)f(\mathbf{z}_i) & jc(i)f(\mathbf{z}) \\ \mathbf{0} & 1 - c(i)d(\mathbf{z}_i) & -jc(i)d(\mathbf{z}, \mathbf{z}_i) \\ \mathbf{0} & -jc(i)d(\mathbf{z}, \mathbf{z}_i) & 1 + c(i)d(\mathbf{z}) \end{bmatrix} =$$

$$\det M(\xi(n))[1 + c(i)(d(\mathbf{z}) - d(\mathbf{z}_i)) - c^2(i)\big(d(\mathbf{z})d(\mathbf{z}_i) - d^2(\mathbf{z}, \mathbf{z}_i)\big)] \tag{36.18}$$

Let

$$\Delta(\mathbf{z}, \mathbf{z}_i) = c(i)(d(\mathbf{z}) - d(\mathbf{z}_i)) - c^2(i)\big(d(\mathbf{z})d(\mathbf{z}_i) - d^2(\mathbf{z}, \mathbf{z}_i)\big) \tag{36.19}$$

Due to conclusion 2, if $\Delta(\mathbf{z}, \mathbf{z}_i) > 0$, $\left|M\left(\tilde{\xi}(n)\right)\right| > |M(\xi(n))|$, Thus, from a initial design $\xi_0(n)$, if there exist two points $\mathbf{z}$ and $\mathbf{z}_i$ such that $\Delta(\mathbf{z}, \mathbf{z}_i) > 0$, we can get a better design $\xi_1(n)$, And then a series of designs $\xi_0, \xi_1, \cdots, \xi_s, \cdots$ satisfying that $\det M(\xi_0) \le \det M(\xi_1) \le \cdots \le \det M(\xi_s) \le \cdots \le \det M(\xi_D^c)$. Thus $\lim_{s\to\infty} \det M(\xi_s)$ exists. The iterative algorithm to construct an integrative Dn-optimal design is as follows.

Step1. for K kinds of experiments, give any original non-degenerated design $\xi_0(n)$ with n points: $\xi_0 = \begin{pmatrix} \xi_{01} & \xi_{02} & \cdots & \xi_{0K} \\ c_1 & c_2 & \cdots & c_K \end{pmatrix}$, $\xi_{0j} = \left(\mathbf{z}_{N_{j-1}+1}, \cdots, \mathbf{z}_{N_{j-1}+N_j}\right)$, and the points can be the same.

Step2. calculate the information matrix $M(\xi_0(n))$ and its inverse matrix$M^{-1}(\xi_0(n))$.

Step3. get the point $\bar{\mathbf{z}}_i$ and $\mathbf{z}_s$ which satisfy $\Delta_s(\mathbf{z}_s, \bar{\mathbf{z}}_i) = \max_{\mathbf{z}_i} \max_{\mathbf{z}} \Delta_s(\mathbf{z}, \mathbf{z}_i)$, then replace $\bar{\mathbf{z}}_i$ with $\mathbf{z}_s$ to get the next design $\xi_{s+1}(n)$.

Step4. $s = s + 1$, repeat from step 2 to step 3 until $\Delta_s(\mathbf{z}_s, \bar{\mathbf{z}}_i)$ is close enough to 0.

Step5. assign points with the measure c_k as the support points of the kth kind of experiment in the design.

Note The sequence $\{\xi_s\}$ is convergent; however, similar to the algorithm to construct the standard Dn-optimal design, the solution of the above algorithm does

not always converge to the integrative Dn-optimal design. As the result is relates only to the initial value, in application, we can set the initial value for several times, and get the design with the highest D-efficiency.

36.4 Example

Consider the influence of the blast depth and gesture of the projectile to the blast domino effect when certain concrete is attacked by the kinetic energy penetration projectile. The flight test and the static test are considered. Experiments have shown that if the fall velocity of the projectile is lower than 1000 m/s, the projectile is not distorted during the penetration [16]. Thus the projectile is assumed rigid.

The damage process can be divided into the penetration process and the blast process. Given the projectile and the drone, the fall velocity v and the fall angle θ are the most important factors in flight test. We use the penetration equation [15, 16] to get the gesture and depth before the blast. As for the static test, the depth and the angle of the warhead buried in the static test are h, γ.

To design the flight test and the static test to evaluate the damage efficiency, the number of flight test is set as 3, and the number of static test is 6. If ricochet does not happen, the value of θ is among $(50^{\circ}, 90^{\circ})$, and the fall velocity is among (300, 1000) m/S. accordingly, the angle of the warhead buried in a static test γ is among $(0^{\circ}, 70^{\circ})$, the depth is among (0.4, 1.5) m. For short, we discuss the normed parameter slope as (0, 1). Assume the variances of the two kinds of experiment after discrepancy modeling are $\hat{\sigma}_{\mathrm{m}}^2 = 0.8$, $\hat{\sigma}_{\mathrm{s}}^2 = 2.4$, thus $c_1 = 0.8$. Then use the polynomial with degree 2 to approximate the damage response function. The support points of the standard Dn-optimal design are (1, 0), (0, 0), (0.50, 0), (0, 1), (1, 1), (0.50, 0.50), (1, 0.50), (0.50, 0.50) and (0, 0.50) (Fig. 36.1).

Take the standard Dn-optimal design as the initial design, then the integrative Dn-optimal design is gotten as follows: the support points of the flight test are (1, 0), (0, 0) and (0.5, 1) while the support points of the static test are (0, 1), (1, 1), (0, 1), (1, 0.65), (0.50, 0.45) and (0, 0.55), or the support points of the flight test are (0, 1), (1, 1) and (0.50, 0), the points of the static test are (1, 0), (0, 0), (0, 0), (1,

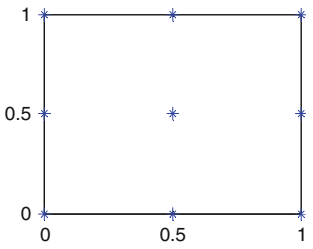


Fig. 36.1 The standard Dn-optimal design

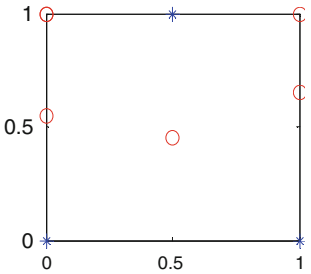


Fig. 36.2 The integrative Dn-optimal design

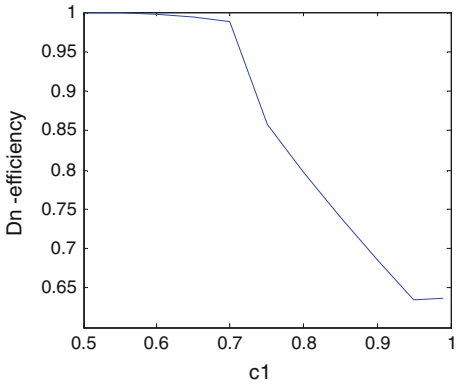


Fig. 36.3 The integrative Dn-efficiency of the standard Dn-optimal design

0.35), (0.5, 0.55) and (0, 0.45). The distributions of the design points are in Fig. 36.2.

Thus, the static test is done on condition (0^o, 1.5 m), (70^o, 1.5 m), (0^o, 1.5 m), (70^o, 0.985 m), (35^o, 0.805 m) and (0^o, 0.895 m), while the blast condition of the flight test are (70^o, 0.4 m), (0^o, 0.4 m), (35^o, 1.5 m). According to the penetration equation [15, 16], the fall velocity and fall angle are set as (50^o, 491.3 m/s), (90^o, 302.5 m/s) and (35^o, 1000 m/s).

The integrative Dn-efficiency of the standard Dn-optimal design is 0.798. If the value of c_1 changes, so does the integrative Dn-efficiency of the standard Dn-optimal design. The trend is described by Fig. 36.3. As c_1 increases, viz. the relative precision of the static test declines, the integrative Dn-efficiency of the standard Dn-optimal design declines too.

36.5 Conclusion

The design of experiments for damage assessment is taken as an example to discuss the integrative optimal design of experiment with multiple sources. The basic idea is as follows. Firstly, we should construct the parameter model about the discrepancy between the equivalent surrogate experiments and the prototype

experiment. Secondly, we should construct the fusion model about all sources of experiment. Thirdly, we should develop the integrative optimal criterion based on the fusion model. Lastly, we can get the Dn-optimal design via the iterative algorithm. The idea proposed in this paper is also fit for other kinds of experiments with multiple sources.

References

1. Loren, S.R., Ian, A.W.: Accelerated aging testing of energetic components—a current assessment of methodology. AIAA2000-3646 (2000)
2. Wang, G.Y., Wang, L.D., Yuan, X.X., et al.: Substitute equivalent reckoning theory and method for radar EM test [M]. Defense Technology Press, Peking (2002)
3. Duan, X.J., Zhou, H.Y., Yao, J.: The decomposition and integration technique of fire dispersion index and conversion of impact deviation. J. Ballistics **17**(2), 42–48 (2005)
4. Jeff, O.: Verification of simulation results using scale model flight test trajectories. AMR-AE-04-01 (2004)
5. Chen, W.C., Li, A.G.: Research on air to air missile combined environment reliability test profile. Aerosp. Shanghai **22**(4), 41–44 (2005)
6. Klopcic, J.T., Reed, H.L.: Historical perspectives on vulnerability/lethality analysis. ADA361816 (1999)
7. Chen, X.W., Zhang, F.J., Yang, S.Q., et al.: Mechanics of structural design of EPW (III): investigations on the reduced-scale test. Explos. Shock Waves **26**(2), 105–114 (2006)
8. Zhao, G.Z., Yang, Y.L.: Equivalent surrogates for armor target damage assessment by kinetic energy projectile. J. Nanjing Univ. Sci. Technol. **27**(5), 509–514 (2003)
9. Huang, H.Y., Duan, X.J., Wang, Z.M.: A novel posterior-weighted Bayesian estimation method considering the credibility of the prior information. ACTA Aeronaut. ET Astronaut. SINICA **19**(5), 1245–1251 (2008)
10. Decarl, D.: Small sample experimental design optimization and repair. DE00005982/XAB (1999)
11. Melas, V.B.: Functional approach to optimal experimental design. Springer Science + Business Media, New York (2006)
12. Klopcic, J.T., Starks, M.W., Albert, J.N.W: A taxonomy for the vulnerability/lethality analysis process. ADA250036 (1992)
13. Xu, S.L., Xiang, H.B.: Similarity theory applied in projectile penetrating into concrete target and deflection analysis. J. Projectiles Rockets Missiles Guidance **27**(3), 123–126 (2007)
14. Yang, Y.L., Zhao, G.Z., Zhang, J.X.: Study on protection level RHA equivalences of advanced armor against KE projectiles. J. Ballistics **15**(2), 64–68 (2003)
15. Lv, X.C., Xu, J.Y.: The study of the calculation about the projectile penetrating the steel-fiber concrete. J. Projectiles Rockets Missiles Guidance **26**(1), 89–92 (2006)
16. Ma, A.E., Huang, F.L.: Experimental research on oblique penetration into reinforced concret. Trans. Beijing Inst. Technol. **27**(6), 482–486 (2007)

Chapter 37
Fast Image Reconstruction Algorithm for Radio Tomographic Imaging

Zhenghuan Wang, Han Zhang, Heng Liu and Sha Zhan

Abstract Radio tomographic imaging is an emerging technology of imaging the attenuation by the objects in the area surrounded by the wireless sensor nodes to locate and track the objects. So it's significant to reconstruct the image in real-time to track the motion of the objects and also with good enough imaging quality. Tikhonov regularization can achieve the real-time requirement with acceptable imaging results by one-step multiplication. Landweber iteration can obtain better imaging quality but need many times of iteration. This paper use pre-iteration method to complete the iteration process of Landweber iteration offline and reconstruct the image online by one-step multiplication, just like Tikhonov regularization. Simulation and experiments show this method can get better imaging results than Tikhonov regularization and imaging the objects in real-time.

Keywords Radio tomographic imaging · Pre-iteration · Landweber iteration · Tikhonov regularization

37.1 Introduction

Radio tomographic imaging (RTI) is a new type of technology for location and tracking objects in the interesting area. Its basic idea is to deploy enough wireless sensor networks (WSNs) nodes surrounding the detection area [1]. If an object is located in the area, some links between the nodes which the radio signals travel

Z. Wang (✉) · H. Liu · S. Zhan
School of Information and Electronics, Bejing Institute of Technology, Beijing, China
e-mail: wangzhenghuan@bit.edu.cn

H. Zhang
Beijing University of Posts and Telecommunications, Beijing, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_37,

through will suffer great loss. Then an image which reflects the attenuation in the area is reconstructed with this technology. The bright spot is in the image where the object locates. RTI is a very appealing for security purpose because it works at radio bands which can penetrate wall and smoke with low power consumption. Other technologies either be blocked by walls or must have large transmitting power such as camera and radar. Another advantage of RTI is that it can utilize inexpensive nodes with small size, which can reduce the cost of the imaging system.

It is very significant to know the location of the objects in real-time, particularly for security areas. So the foremost aspect is that RTI must reconstruct the image fast enough to track the motion of the objects, followed by the localization precision or imaging quality. Image reconstruction of RTI is an ill-posedness problem. Some methods are used to solve this problem such as linear back projection (LBP), truncated singular value decomposition (TSVD), total variation (TV), Tikhonov regularization (TR), Landweber iteration (LI) [2–5]. LBP is very simple and can also achieve real-time imaging, but the quality of image is not quite good. TSVD and TV are too computational expensive to meet the requirement of real-time processing. TR is a good choice because this method can not only reconstruct the image real-time but also get acceptable imaging results. LI can achieve better imaging results than TR with finite iteration. This paper will use a pre-iteration method to complete the iteration process offline and reconstruct the image by one-step multiplication. So this method can not only retain the real-time characteristics as TR but also obtain better imaging results.

37.2 The Model of Radio Tomographic Imaging

If K is the number of wireless sensor nodes deployed outside the imaging area as depicted in Fig. 37.1, when there is an object in the imaging area, some links will be blocked which means the signal will suffer great attenuation, usually up to 5–10 dB. Suppose P_i is received signal strength (RSS) of link i measured when there is an object in the area while P_i^e is RSS measured when the area is vacant. Their difference y_i is the shadowing loss of the link $i$caused by object's obstruction.

$$y_i = P_i - P_i^e \tag{37.1}$$

where the unit is dBm.

In order to obtain the image of attenuation when signal travels though the imaging area, the area is divided into many square regions with the same size and each small region is called a pixel. Suppose that the total number of pixels is N. x_j is the attenuation when the signal passes through the pixel j. Then y_i can be seen as the weighted sum of x_j [1].

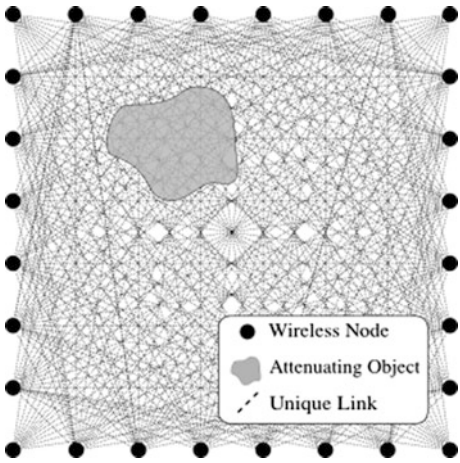


Fig. 37.1 An illustration of radio tomographic imaging with wireless sensor network

$$y_i = \sum_{j=1}^{N} w_{ij} x_j + n_i \tag{37.2}$$

where n_i is the noise and w_{ij} is the weight of pixel j for link i.

w_{ij} can be determine by the ellipse model for each link in the network, which is very effective in outdoor environment.

$$w_{ij} = \frac{1}{\sqrt{d_i}} \begin{cases} 1 & if\ d_{ij}(1) + d_{ij}(2) < d_i + \delta \\ 0 & otherwise \end{cases} \tag{37.3}$$

where δ is a tunable parameter called ellipse parameter which describes the width of the ellipse, d_i is the distance between the two nodes, $d_{ij}(1)$ and $d_{ij}(2)$ are the distances between the center of pixel j and the two nodes for link i respectively.

In order to look more compact, (37.2) can be written in matrix form as

$$\mathbf{y} = \mathbf{wx} + \mathbf{n} \tag{37.4}$$

where $\mathbf{y} = [y_1, y_2, y_3, \ldots, y_M]^T \in R^M$ is shadowing loss vector, $\mathbf{w} = [w_{ij}]_{M \times N} \in R^{N \times M}$ is weight matrix, $\mathbf{x} = [x_1, x_2, x_3, \ldots, x_N]^T \in R^N$ is pixel vector and $\mathbf{n} = [n_1, n_2, n_3, \ldots, n_M] \in R^M$ is noise vector.

37.2.1 Tikhonov Regularization

Image reconstruction of radio tomographic imaging is an ill-posed problem because the number of pixels is much more than the number of RSS measurements in the wireless sensor network. TR might be the most popular method to solve the

ill-posed problem [1–3]. The standard TR is to minimize the following objective function.

$$\min_x \|y - wx\|^2 + \lambda \|x\|^2 \tag{37.5}$$

where λ is TR parameter. The solution of (37.5) is

$$\hat{\mathbf{x}}_{TR} = (\mathbf{w}^T\mathbf{w} + \lambda\mathbf{I})^{-1}\mathbf{w}^T\mathbf{y} \tag{37.6}$$

From the (37.6) it is quite clear that once $\lambda$is determined, the $(\mathbf{w}^T\mathbf{w} + \lambda\mathbf{I})^{-1}\mathbf{w}^T$ can be calculated in advance [1]. So the procedure of TR can be divided into two steps: one is get the $(\mathbf{w}^T\mathbf{w} + \lambda\mathbf{I})^{-1}\mathbf{w}^T$ offline and the other one is online image reconstruction by one step multiplication. That means real-time processing is possible, which is a very appealing feature of TR.

37.2.2 Landweber Iteration

Landweber iteration is widely used in some other image reconstruction areas such as ECT [3, 5]. It aims to minimize the following objective function in an iterative way.

$$f(\mathbf{x}) = \frac{1}{2} \| \mathbf{y} - \mathbf{w}\mathbf{x} \|^2 = \frac{1}{2}(\mathbf{y} - \mathbf{w}\mathbf{x})^T(\mathbf{y} - \mathbf{w}\mathbf{x}) \tag{37.7}$$

The steepest gradient descent method chooses the direction in which $f(\mathbf{x})$ as new search direction for next iteration [3]. This direction is opposite to the gradient of $f(\mathbf{x})$ at current point. The iteration procedure is therefore

$$\hat{\mathbf{x}}_{k+1} = \hat{\mathbf{x}}_k - \mu\nabla f(\hat{\mathbf{x}}_k) = \hat{\mathbf{x}}_k - \mu\mathbf{w}^T(\mathbf{w}\hat{\mathbf{x}}_k - \mathbf{y}) = (\mathbf{I} - \mu\mathbf{w}^T\mathbf{w})\mathbf{x}_k + \mu\mathbf{w}^T\mathbf{y} \tag{37.8}$$

where the constant μ is known as gain factor and is used to control convergence rate. The choice of μ will be explained later.

LI method needs many times of iteration before obtaining satisfying results, which is not suitable for on-line imaging.

37.2.3 Pre-iteration Landweber Iteration

In fact, the iteration task of LI can be undertaken offline. If $\mathbf{D} = \mathbf{I} - \mu\mathbf{w}^T\mathbf{w}$, then the equation can be rewritten as

$$\hat{\mathbf{x}}_{k+1} = \mathbf{D}\hat{\mathbf{x}}_k + \mu\mathbf{w}^T\mathbf{y} \tag{37.9}$$

The matrix $\mathbf{D}$ is independent of $\mathbf{x}$ and $\mathbf{y}$ and it is an interesting feature of the LI method. Suppose that the initial value $\hat{\mathbf{x}}_0=0$, then after k iteration, the solution can calculated as follows

$$\hat{\mathbf{x}}_k= (\mathbf{I}+\mathbf{D}+\mathbf{D^2}+\mathbf{D^3}+\ldots+\mathbf{D^{k-1}})\mu\mathbf{w}^T\mathbf{y}=\mathrm{P}\mu\mathbf{w}^T\mathbf{y} \tag{37.10}$$

where $\mathrm{P}= (\mathbf{I}+\mathbf{D}+\mathbf{D^2}+\mathbf{D^3}+\ldots+\mathbf{D^{k-1}})\mu\mathbf{w}^T$. Similar to TR, the coefficient matrix $\mathbf{P}$ can be computed in advance and stored in the computer for real-time processing [6, 7]. Then the imaging process can be very simple, which just needs that the $\mathbf{P}$ multiply the observed shadowing loss vector $\mathbf{y}$. Compared to TR, this method can not only keep the real-time performance that TR holds, but also achieve better imaging results than TR.

In (37.10) the iteration number k should be appropriate chosen because LI and PLI have the drawback of semi-convergence, which means the imaging quality deteriorates when k is larger than a certain number [3, 4, 8].

There is indeed an optimal iteration number k_0 existing to make the imaging error reach the minimum, but it's very difficult to determine. In most cases, k_0 is chosen empirically and it's sufficient to meet the requirement of most cases.

37.2.4 Calculation of P

At the first glance it might be complex to compute $\mathbf{P}$ because $\mathbf{D}$ is a very large matrix with dimension $n \times n$ and the computation of $\mathbf{P}$ requires a lot of time. In fact if we use some property of $\mathbf{P}$, the computation can be simplified substantially. Suppose that the singular value decomposition (SVD) of $\mathbf{w}$ is $\mathbf{w} = \mathbf{U}\Sigma\mathbf{V}^H= \sum_{i=1}^{r} \sigma_i\mathbf{u}_i\mathbf{v}_i^H$ where $\mathbf{u}_i, \mathbf{v}_i, \sigma_i$ are the singular vector and singular value of $\mathbf{w}$ respectively [9].

Then the SVD of $\mathbf{D}$ and $\mathbf{P}$ will be

$$\mathbf{D} = \mathbf{I}-\mu\mathbf{w}^T\mathbf{w}=\mathbf{V}(\mathbf{I}-\mu\mathbf{\Sigma}^H\Sigma)\mathbf{V}^H = \mathbf{V}diag(1-\mu\sigma_1^2, 1-\mu\sigma_2^2, \ldots, \quad 1-\mu\sigma_r^2\ )\mathbf{V}^H \tag{37.11}$$

$$\mathbf{P} = (\mathbf{I}+\mathbf{D}+\mathbf{D^2}+\mathbf{D^3}+\ldots+\mathbf{D^{k_0-1}})\mu\mathbf{w}^T=\mu\mathbf{V}diag(\sum_{k=0}^{k_0-1}(1-\mu\sigma_1^2)^k, \sum_{k=0}^{k_0-1}(1-\mu\sigma_2^2)^k, \ldots, \sum_{k=0}^{k_0-1}(1-\mu\sigma_r^2)^k)\mathbf{U}^H$$
$$= \sum_{i=1}^{r}\frac{1-(1-\mu\sigma_i^2)^{k_0}}{\sigma_i}\mathbf{v}_i\mathbf{u}_i^H \tag{37.12}$$

It can be seen that once SVD of the $\mathbf{w}$ is completed and other parameters are determined, $\mathbf{P}$ can be obtained immediately through (37.12).

From the (37.12), the principle of choosing μ can also be obtained. In order to guarantee the convergence, $1-\mu\sigma_k^2$ should be less than 1. So the choice of μ should be $\mu<(1/\sigma_{\max}^2)$, where the $\sigma_{\max}$ is the largest singular value of $\mathbf{w}$.

37.3 Results

37.3.1 Compasion of TR and PLI Using Simulated Data

It is obvious that the landweber iteration and Tikhonov regularization can complete the reconstruction by one-step processing. So their computational speed is the same. The only one aspect they may be different is the imaging quality.

The relative imaging error e is defined by

$$e=\frac{\| \mathbf{x}-\widehat{\mathbf{x}} \|}{\| \mathbf{x} \|} \tag{37.13}$$

The true shadowing loss $\mathbf{x}$ is obtained by simulation. We make the following assumption that shadowing loss of the links affected by object's obstruction is about 5 dB and the shadowing loss of other links is zero. The noise vector subject to Gaussian distribution and the variance is 0.8, $\mathbf{n}\sim N(0,0.8)$.

The deployment of sensors is depicted in Fig. 37.2. The object is located at the center of the imaging area and modeled as a cylinder with radius of 10 cm. The imaging area is divided by 60*60 pixels and the size of each pixel 0.2 m*0.2 m, which means the numbers of elements in $\mathbf{x}$ is 3600. The ellipse parameter σ, TR parameter λ and gain factor μ are chosen to be 0.03 m, 100 and 0.0001 respectively.

We can see the semi-convergence phenomenon of LI or PLI from the Fig. 37.3. At the beginning the relative imaging error decreases rapidly. After about 80 iterations, the relative imaging error reaches the minimum values and the corresponding error is 0.95. While when the iteration process continues, the relative imaging error increases and it's not difficult to guess that the imaging error will

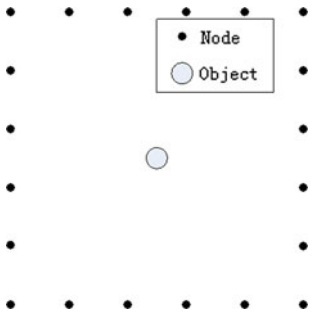


Fig. 37.2 The simulation setting

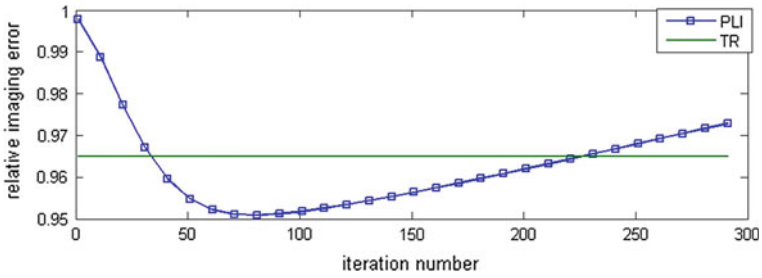


Fig. 37.3 The relative imaging error versus iteration number

converge to LS method when the iteration number reaches infinity. TR is not an iterative method, so its simulation curve is a straight line and the imaging error is 0.96 which is higher than LI or PLI. So we can conclude that PLI can achieve better imaging results than TR method.

37.3.2 Performance of PLI Using Experiments

To evaluate the performance of PLI method, the experiment was also conducted in outdoor environment. An area of 6 m × 6 m was monitored by 20 sensor nodes, as illustrated in Figs. 37.2 and 37.4. Each node was placed 1.2 m apart along the square premier of the monitored area and mounted on a tripod at a height of 1 m.

The nodes use JN5139 chip which is compatible with IEEE 802.15.4 protocol and operate at 2.4G frequency. A token ring protocol is utilized to avoid transmission collisions. During each time interval of 3 ms, one node broadcasted one packet. All the other nodes received the packet and measured the RSS of the packet. Then the token was passed to the next node in the next time interval. So the

Fig. 37.4 The experiment scenario

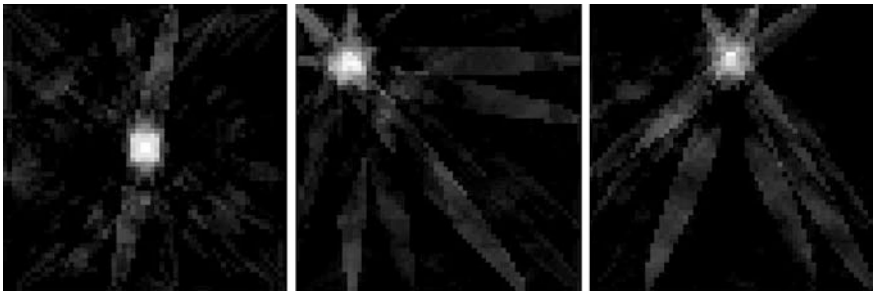


Fig. 37.5 The imaging results when one person locates the three positions

Fig. 37.6 The imaging results using PLI method when two persons locate in the monitored area

RSS on each link was updated every 60 ms. It's fast enough to track the motion of the targets in the area. The nodes sent the measured RSS data to the laptop. The laptop collected the data from the sensors and run the imaging software based on PLI in real-time

The values of parameters are the same with the simulation. Figure 37.4 shows the experiment scenario that a person was moving in the area in a norm speed. During the environment the person's location was shown real-time from the software.

Figure 37.5 shows the imaging results when person locating at three spots. From the images the person's position is clearly shown, which is the bright spot in the image.

In the model of RTI, it doesn't restrict the number of targets. In fact, RTI can also track multiple objects at the same time. Figure 37.6 shows the results when two persons moving in the monitored area using PLI method. It can be inferred when the two persons come too close, the results become a little worse and when the person stands far away, the results are much better.

37.4 Conclusion

This paper presents a pre-iteration landweber method to meet the real-time requirement of radio tomographic imaging. The PLI method can reconstruct the image by one-step multiplication, just like TR does while achieving better imaging results than TR. Simulation and experiment have demonstrated that PLI can locate and track the objects in real-time with enough precision.

Acknowledgments This work was supported in part by National Natural Science Foundation of China (No. 61101129, No. 61227001 and No. 60972017), Specialized Research Fund for the Doctoral Program of Higher Education (SRFDP) under Grant No.20091101110019 and No. 20091101120028.

References

1. Wilson, J., Patwari, N.: Radio tomographic imaging with wireless networks. IEEE Trans. Mob. Comput. **9**(5), 621–632 (2010)
2. Wilson, J., Patwari, N.: Regularization methods for radiotomographic imaging. In: Proceedings of the Virginia Tech Symposium on Wireless Personal Communications (2009)
3. Yang, W.Q., Peng, L.H.: Image reconstruction algorithms for electrical capacitance tomography. Meas. Sci. Technol. **14**, R1–14 (2003)
4. Peng, L.H., Mercus, H., Scarlett, B.: Using regularization methods for image reconstruction of electrical capacitance tomography. Meas. Sci. Technol. **17**, 96–104 (2000)
5. Yang, W.Q., Spink, D.M., York, T.A., McCann, M.: An image-reconstruction algorithm based on Landweber's iteration method for electrical-capacitance tomography. Meas. Sci. Technol. **10**, 1065–1069 (1999)
6. Liu, S., Fu, L., et al.: Prior-online iteration for image reconstruction with electrical capacitance tomography. IEEE Proc. Sci. Meas. Technol **151**(3), 195–200 (2004)
7. Wang, H., Wang, C., Yin, W.: A Pre-iteration method for the inverse problem in electrical impedance tomography. IEEE Trans. Instrum. Meas. **53**(4), 1093–1096 (2004)
8. Xiong, X.Y., Zhang, Z.T., Yang, W.Q.: A stable image reconstruction algorithm for EC. J. Zhejiang Univ. Sci. **6A**(12), 1401–1404 (2005)
9. Zhang, X.D.: Matrix Analysis and Applications. Tsinghua University Press, Beijing (2004)

Chapter 38
Reliability Monitoring Algorithm in Multi-Constellation Satellite Navigation

Lin Yang, Hao Wu, Yonggang Sun, Yongzhi Zheng and Yongxue Zhang

Abstract Global Navigation Satellite Systems (GNSS), including GPS, BeiDou, GLONASS, Galileo and other systems, are becoming more and more widely used in our today life. As a result, multi-constellation receiver that compatible with more than one system will take the place of single-constellation receiver that only uses GPS or BeiDou system. Reliability monitoring algorithm in single-constellation receiver has some limitations when applied in multi-constellation receiver. Therefore, an enhanced reliability monitoring algorithm based on weighted and statistic schemes is designed for multi-constellation receiver that compatible with GPS and BeiDou in this paper. The experiment results show that the algorithm improves the performance in both static and dynamic scenes, especially about 31 % in static scene than the BeiDou/GPS multi-constellation receiver without it.

Keywords Reliability · Navigation · Multi-constellation · Weighted · Statistic

38.1 Introduction

Global Position System (GPS), controlled by the USA, is based on a man-made constellation of 27 Earth-orbiting satellites. Using these satellites, a person or object can obtain its position, velocity and time information. While GPS can be effectively used for many navigational applications, it has limitations [1]. For example, the availability of the GPS receiver is low in urban canyons because high buildings shield line-of-sight satellite signals and receiver could not obtain an available position result with less than four satellites. As GPS is widely used in cell phones, the emergency call positioning demands (U.S. E-911 mandate, E.C. E-112

L. Yang (✉) · H. Wu · Y. Sun · Y. Zheng · Y. Zhang
Navigation Chipset Department, Beijing Microelectronics Technology Institute, Beijing, China
e-mail: bsbandwl@126.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_38,

directive) and popularity of location-based services increase year by year [2]. So navigation capability and reliability are now also required and expected in degraded signal-environments such as indoors.

With the development of BeiDou Navigation Satellite System (BNSS) in China, BeiDou is able to provide service for areas in China and its surrounding countries [3]. The simplest way to improve the reliability of navigation receiver is to acquire more satellite signals. The more satellite signals acquired, the more redundancies obtained. Therefore, multi-constellation receiver that compatible with GPS, BeiDou and/or more systems could uses 24 or more available satellites, while GPS-only receiver uses only 12 satellites. A commercial BeiDou/GPS multi-constellation receiver that based on MXT3013 baseband chip designed by Beijing Microelectronic Technology Institute (BMTI) is chosen in this paper and the algorithm and experiments are designed and done in it [4].

38.2 The Principle of Global Navigation Satellite System

Even though GPS, BeiDou, GLONASS and Galileo are all different with each other, they all belong to Global Navigation Satellite systems (GNSS) and the principles of them are similar. So GPS is chosen to introduce the principle of single-constellation receiver. Then the principle of multi-constellation receiver is discussed later.

38.2.1 The Principle of Single-Constellation Receiver

The GPS-only receiver is the most popular commercial single-constellation receiver. Therefore, take GPS as an example. The speed of signals transmitted by the GPS satellites is equal to the speed of light. So the range between satellite and receiver could be calculated from the time that signal takes from satellite to receiver. As the existence of ionospheric and tropospheric delay, the multipath delay and other bias, the range calculated above is named Pesudorange (ρ) as follows:

$$\rho = c(t_{receive} - t_{transmit}) \approx \sqrt{(x_{SV} - x_{Rcvr})^2 + (y_{SV} - y_{Rcvr})^2 + (z_{SV} - z_{Rcvr})^2} \tag{38.1}$$

c is the speed of light, $t_{transmit}$ and $t_{receive}$ are the times signal transmitted and received respectively, (x_{SV}, y_{SV}, z_{SV}) and $(x_{Rcvr}, y_{Rcvr}, z_{Rcvr})$ are the positions of satellite vehicle and receiver respectively.

As the receiver clock is not as accurate as the satellite atomic clock, the clock bias (δt_{Rcvr}) is also needed to be calculated. So at least four equations are needed to acquire an available navigation solution:

$$\begin{cases} \rho^1 = \sqrt{(x_{SV}^1 - x_{Rcvr}) + (y_{SV}^1 - y_{Rcvr}) + (z_{SV}^1 - z_{Rcvr})} + c\delta t_{Rcvr} \\ \rho^2 = \sqrt{(x_{SV}^2 - x_{Rcvr}) + (y_{SV}^2 - y_{Rcvr}) + (z_{SV}^2 - z_{Rcvr})} + c\delta t_{Rcvr} \\ \rho^3 = \sqrt{(x_{SV}^3 - x_{Rcvr}) + (y_{SV}^3 - y_{Rcvr}) + (z_{SV}^3 - z_{Rcvr})} + c\delta t_{Rcvr} \\ \rho^4 = \sqrt{(x_{SV}^4 - x_{Rcvr}) + (y_{SV}^4 - y_{Rcvr}) + (z_{SV}^4 - z_{Rcvr})} + c\delta t_{Rcvr} \end{cases} \tag{38.2}$$

Equations above could be solved by Least-Squares adjustment, and then the receiver position $(x_{Rcvr}, y_{Rcvr}, z_{Rcvr})$ and receiver clock bias δt_{Rcvr} could be concluded.

38.2.2 The Principle of Multi-constellation Receiver

Multi-constellation receiver could acquire more satellites and therefore performs better than single-constellation receiver. A BeiDou/GPS commercial receiver is chosen in this paper. The coordinate of BeiDou and GPS is almost equal and the main difference between GPS and BeiDou is the time system. GPS-only receiver uses GPST and BeiDou-only receiver uses BDT, but BeiDou/GPS receiver will both uses GPST and BDT. As a result, a BeiDou/GPS receiver at least needs five satellites from the two systems in all to acquire an available navigation solution. The result can be calculated as follows:

$$\begin{cases} \rho^1 = \sqrt{(x_{SV}^1 - x_{Rcvr}) + (y_{SV}^1 - y_{Rcvr}) + (z_{SV}^1 - z_{Rcvr})} + c\delta t_{GPS-Rcvr} \\ \quad \vdots \qquad\qquad\qquad \vdots \qquad\qquad\qquad \vdots \\ \rho^i = \sqrt{(x_{SV}^i - x_{Rcvr}) + (y_{SV}^i - y_{Rcvr}) + (z_{SV}^i - z_{Rcvr})} + c\delta t_{GPS-Rcvr} \end{cases}$$

$$\begin{cases} \rho^1 = \sqrt{(x_{SV}^1 - x_{Rcvr}) + (y_{SV}^1 - y_{Rcvr}) + (z_{SV}^1 - z_{Rcvr})} + c\delta t_{BD-Rcvr} \\ \quad \vdots \qquad\qquad\qquad \vdots \qquad\qquad\qquad \vdots \\ \rho^k = \sqrt{(x_{SV}^k - x_{Rcvr}) + (y_{SV}^k - y_{Rcvr}) + (z_{SV}^k - z_{Rcvr})} + c\delta t_{BD-Rcvr} \end{cases}$$

where $\delta t_{GPS-Rcvr}$ and $\delta t_{BD-Rcvr}$ are the clock biases between receiver clock and GPS or BeiDou time system respectively, the other variables are the same in (38.1) and (38.2).

If the sum of GPS and BeiDou satellites is five or more $(i + k \geq 5)$, receiver can conclude a solution includes receiver position $(x_{Rcvr}, y_{Rcvr}, z_{Rcvr})$ and the two clock biases, $\delta t_{GPS-Rcvr}$ and $\delta t_{BD-Rcvr}$.

38.3 Reliability Monitoring Algorithm for Multi-constellation Receiver

There are many different Receiver Autonomous Integrity Monitoring (RAIM) algorithms designed for single-constellation in related papers: Range and position comparison method [5], Least-square residual method [6], Parity vector method [7], reliability testing introduced by Heidi Kuusniemi [2], etc. The first three methods have been proved to be equal by Brown [8] and reliability testing method has been proved effective and reliable in GPS receiver.

The navigation procedure of multi-constellation receiver is more complex than single-constellation receiver, so these single-constellation RAIM methods would be modified when applied to multi-constellation receiver. In this paper, a weighted scheme is introduced to combine BeiDou and GPS satellite systems, the fault detection and exclusion of inner system is accomplished by a statistic scheme.

38.3.1 The Statistic Scheme for Inner-Constellation Faults

It is assumed that the faults in the measurements are normally distributed and that they are uncorrelated. Additionally, Least-Squares adjustment is applied in the procedure of navigation processing and statistic test is applied to detect the faults in the inner-constellation after Least-Squares adjustment. Apply linearization, iteration and Least-Squares adjustment to (38.2):

$$X_k = X_{k-1} + \Delta X = \begin{bmatrix} x_{k-1} \\ y_{k-1} \\ z_{k-1} \\ \delta t_{k-1} \end{bmatrix} + \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \\ \Delta \delta t_{Rcvr} \end{bmatrix} = \begin{bmatrix} x_{k-1} \\ y_{k-1} \\ z_{k-1} \\ \delta t_{k-1} \end{bmatrix} + (G^T G)^{-1} G^T b \quad (38.3)$$

$$G = \begin{bmatrix} -l^1(X_{k-1}) & 1 \\ \vdots & \\ -l^i(X_{k-1}) & 1 \end{bmatrix} \quad (38.4)$$

$$l^i(X_{k-1}) = \begin{bmatrix} \frac{x^i - x_{k-1}}{r^i(X_{k-1})} & \frac{y^i - y_{k-1}}{r^i(X_{k-1})} & \frac{z^i - z_{k-1}}{r^i(X_{k-1})} \end{bmatrix} \quad (38.5)$$

$$b = \begin{bmatrix} \rho^1 - \delta t_{k-1} - r^1(X_{k-1}) \\ \vdots \\ \rho^i - \delta t_{k-1} - r^i(X_{k-1}) \end{bmatrix} \quad (38.6)$$

$$r^i(X_{k-1}) = \sqrt{(x^i - x_{k-1})^2 + (y^i - y_{k-1})^2 + (z^i - z_{k-1})^2} \quad (38.7)$$

where $X = [x \quad y \quad z \quad \delta t_{Rcvr}]^T$ is the state of the receiver that contains position and clock bias, the initial of the iteration (X_0) is a rough guess of the receiver state even set zeros to it. After six or seven iterations the result X_k will be convergent. Matrix G is the design matrix, vector I is the line-of-sight unit vector between receiver and satellite, vector b is the observation vector and scalar r is the true or geometric range between receiver and satellite. Formula superscript i stands for the ith available satellite and formula subscript k stands for the number of iterations.

Statistic test uses the sum of the squares of the range residual errors (SSE) to judge whether fault exists [8]. SSE could be calculated as below:

$$SSE = v^T v \tag{38.8}$$

$$v = b - G\Delta X = b - G(G^T G)^{-1} G^T b \tag{38.9}$$

In which v represents the residual of the error between the Pesudorange and the range calculated after the navigation solution is acquired. SSE is a scalar quantity and represents the consistency between the measurements. So SSE could be used to compare with a precalculated threshold. As is assumed before, the faults are normally distributed. Therefore, the statistical distribution of SSE is Chi square distribution (χ^2) whose dimension of freedom (DOF) is $i - 4$, i is the number of available satellites. The threshold can be obtained by DOF and the false alarm probability α. The DOF determines probability density function (PDF) and α finally determines threshold. As shown in Fig. 38.1, the threshold is 11.3448 when DOF is 3 and α is 0.01. If SSE exceeds 11.3448, some faults might exist in the measurements and result is unreliable.

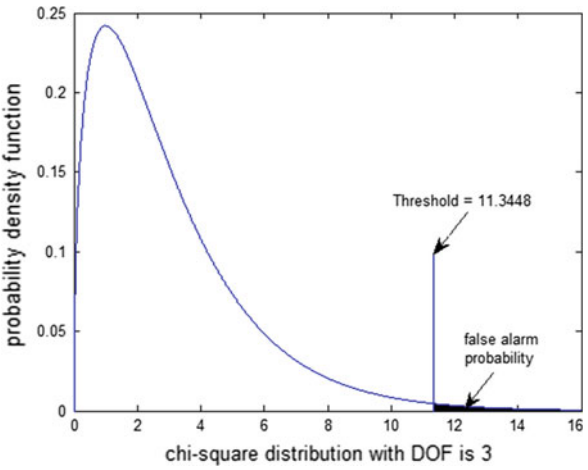


Fig. 38.1 The threshold of SSE with DOF is 3 and false alarm probability is 0.01

38.3.2 The Weighted Scheme for Joint-Constellation Solution

After statistic tests applied to every constellation respectively, the navigation result by every constellation and its SSE can be concluded. These results would be used to calculate the multi-constellation navigation solution. For a BeiDou/GPS receiver, if GPS and BeiDou satellites are all enough to conclude an available solution respectively, then joint-constellation solution can be concluded with the weights as below:

$$X_{BD/GPS} = \frac{SSE_{GPS}}{SSE_{BD} + SSE_{GPS}} X_{BD} + \frac{SSE_{BD}}{SSE_{BD} + SSE_{GPS}} X_{GPS}$$

where formula subscripts GPS, BD, GPS/BD stand for the satellite systems and the other variables are the same in (38.3) and (38.8).

If one or two constellations are not enough for navigation $(i \geq 4)$ but the sum of the two constellations satellites is five or more $(i + k \geq 5)$, an available solution can still be calculated which is introduced in Sect. 2.2.

38.4 Experiments Studying

Two experiments are designed to verify the enhanced reliability monitoring algorithm, one is static scene experiment and the other is dynamic scene.

38.4.1 Static Scene

In static scene, a 2400 s signals record was used and six faults were introduced during the period. If no reliability monitoring algorithm was applied, the performance would be poor and unreliable as shown in the left part of Fig. 38.2. The receiver with reliability monitoring algorithm performed more reliable and all the six faults were detected and excluded, so no outliers can be found in the right part of Fig. 38.2. Table 38.1 shows the statistic result and the performance improved about 31.21 %.

38.4.2 Dynamic Scene

The dynamic experiment was taken in Zhongguancun, an urban canyon in Beijing. As shown in Fig. 38.3, the blue trajectory was logged by the receiver without the

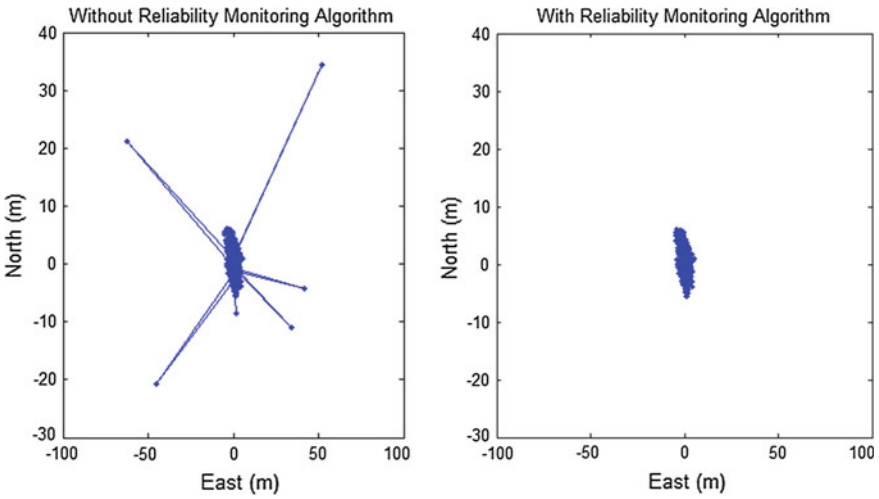


Fig. 38.2 The performance of the receiver with or without reliability monitoring algorithm in static scene

Table 38.1 Algorithm performance in Static Scene

With or Without Algorithm	North		East	
	Mean (m)	Std. (m)	Mean (m)	Std. (m)
Without	-0.32453	1.95511	0.23075	2.57130
With	-0.33241	1.71442	0.22293	1.41332

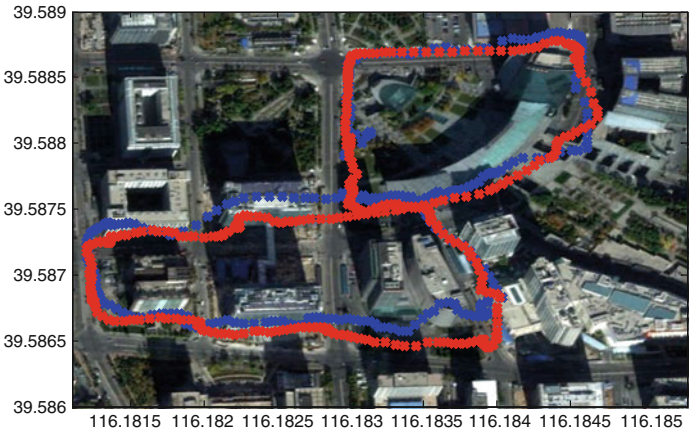


Fig. 38.3 The performance of the receiver with or without the algorithm in dynamic scene

reliability monitoring algorithm and the red trajectory was logged by the receiver with the algorithm. The red trajectory is closer to the actual road, and the reliability is improved very much.

38.5 Conclusion

This paper describes an enhanced reliability monitoring algorithm in multi-constellation satellite navigation receiver. It bases on the statistic and weighted schemes to detect faults in measurements and obtain a more reliable solution. The BeiDou/GPS multi-constellation receiver with this algorithm will perform more reliable than the receiver without it both in static and dynamic scenes. The static experiment shows that the performance is improved about 31 %.

References

1. Zekavat, S., Buehrer, M.: Handbook of Position Location: Theory, Practice and Advances, pp. 9–10. Wiley, Hoboken (2012)
2. Kuusniemi, H., Wieser, A., et al.: User-level reliability monitoring in urban personal satellite-navigation. IEEE Trans. Aerosp. Electron. Syst. **43**(4), 1305–1317 (2007)
3. China Satellite Navigation Office.: BeiDou navigation satellite system signal in space interface control document (2011)
4. Yang, L., Sun, Y., Zhang, Y., et al.: Implementation of a multi-constellation and multi-mode navigation terminal equipment. In: Proceedings of the 2012 IEEE/ION Position Location and Navigation Symposium, pp. 86–91 (2012)
5. Ober, P.B.: Integrity Prediction and Monitoring of Navigation Systems, pp. 7–9. Integricom Publishers, Leiden (2003)
6. Xie, G.: Principles of GPS and Receiver Design, pp. 118–120. Publishing House of Electronics Industry, Beijing (2009)
7. Kaplan, E.D., Hegarty, C.J.: Understanding GPS: Principles and Applications, 2nd edn, pp. 258–265. Publishing House of Electronics Industry, Beijing (2008)
8. Grover Brown, R.: A Baseline GPS RAIM scheme and a note on the equivalence of three RAIM methods. Navigation **39**(3), 301–316 (1992)

Chapter 39
CTL Model Checking Algorithm Using MapReduce

Feng Guo, Guang Wei, Mengmeng Deng and Wanlin Shi

Abstract Model checking is a promising automated verification technique. The state space explosion is the major difficulty of model checking. To deal with this problem, researchers present a new model checking algorithm for the temporal logic CTL based on MapReduce framework. And the algorithm's data structure is defined for the Kripke structure. This MapReduce algorithm outputs the set of states of the model that satisfies the formula by giving a model and a CTL formula. Researchers justify its correctness by an example with the EU formula. Finally, an example illustrates the validity of this algorithm, and the result shows this method is feasible.

Keywords Model checking · CTL · MapReduce

39.1 Introduction

In software and hardware design of complex systems, the formal verification spends more time and effort than construction. Many methods are proposed to reduce and ease this problem. As a verification technique, model checking [2, 4] is an automatic, model-based, property-verification approach. It checks exhaustively and automatically whether the given system's model satisfies a given specification. And if it is not satisfied, it will usually produce a trace of system behavior which causes this failure.

The set of all states with a system, state space can be very large or even infinite. This is called the state space explosion problem, and is one of the most serious problems with model checking.

F. Guo · G. Wei (✉) · M. Deng · W. Shi
North China University of Technology, Beijing, China
e-mail: weiguang0314@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_39,

Model checking is based on temporal logics [4], such as Linear Temporal Logic (LTL) and Computation Tree Logic (CTL). CTL is the most popular temporal logic. It's a branching-time logic, and its model of time is a tree-like structure in which the future is not sure.

CTL formula constitutes of the path operator and state operator. Each CTL temporal connective is a pair of symbols. The first is a path operator of A and E which means 'along All paths' and 'Exist one path', the second symbol is a state operator of X, F, G, or U, meaning 'neXt state', 'some Future state', 'Globally' and 'Until'. For example, the pair of symbols in E [p U q] is EU. And pairs of symbols like EU are indivisible.

To solve the state explosion problem in model checking, the paper proposes a new parallel algorithm. It is designed based on the MapReduce framework to compute the set of states of the model that satisfy the given CTL formula.

Hadoop [1, 7, 8] is an open-source framework for reliable, scalable, distributed computing. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. It enables programmers easily to write and run distributed applications that process large amounts of data. The Hadoop platform consists of the Hadoop kernel, MapReduce and HDFS, as well as a number of related projects. MapReduce is a simplified parallel programming model for large-scale datasets processing.

The processing of MapReduce works has two phases: the map phase and the reduce phase. The input and output of each phase are key-value pairs which is given in the general form

map: $< k1, v1 > \rightarrow list(< k2, v2 >)$
reduce: $< k2, list(v2) > \rightarrow list(< k3, v3 >)$

Related work: most of the distributed model checking techniques shares a common idea that each machine in the network explores the partial state space. Some related studies are presented in other papers [3] in the context of distributed model checking.

The rest of the paper is organized as follows: The next section gives the algorithm's data structure definitions. In Sect. 39.3, this paper presents the design and realization of the CTL model checking algorithm based on MapReduce framework, with an example using the EU formula. The Experiment illustrates the validity and feasible of this algorithm in Sect. 39.4. In the end, the conclusion and future work are given.

39.2 Data Structure Definitions

The various algorithms for model checking, such as LTL and CTL, are all based on a system description in terms of the Kripke structure [6]. A Kripke structure is defined over AP, a set of atomic propositions, as a 4-tuple $M = (S, I, R, L)$,

consisting of a finite set of states S, a set of initial states $I \subseteq S$, a transition relation $R \subseteq S \times S$ and a labeling function $L: S \rightarrow 2^{AP}$.

For example, consider the CTL model in Fig. 39.1. It is basically a graph whose nodes represent the reachable states of the CTL model, and whose edges represent state transitions. A labeling function maps each node to a set of properties that hold in the corresponding state. Temporal logics are traditionally interpreted in terms of Kripke structures.

As an example of Kripke structure, the CTL model in Fig. 39.1 is given by:

$S = \{s0, s1, s3\}$
$I = \{s0\}$
$R = \{(s0, s1), (s1, s2), (s2, s0), (s2, s2)\}$
$L(s0) = \{p1\}, L(s1) = \{p2\}, L(s2) = \Phi$

In MapReduce, the data structure is described as follows: Consider the key-value pair. The key represents the state ID, and the value represents the state's information, such as its status flag, pre-successors, labels and successors' information. Each row of data represents a state, and the data structure of each state is given below:

ID	Color	Pre-successors' info.	My labels	Successors' info.

ID	state identification
Color	status flag
Pre-successors' info.	all pre-successors of this state
My Labels	the labels of this state
Successors' info.	the successors' information (the successor ID and labels)

The data structure of s1 in Fig. 39.1, for example, is: s1 WHITE|s0|p2|s2, null

If the state adds a new label, then it should inform all its pre-successors to update, letting the pre-successors update their successor's information. The notification data structure is as follows:

ID of the pre-successor to inform	Color	The state's ID	The state's new labels

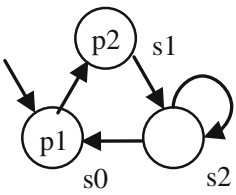


Fig. 39.1 A simple CTL model

39.3 Algorithm Design and Realization

The CTL model-checking algorithm uses the labeling algorithm [5]. The labeling algorithm is an algorithm which gives a model and a CTL formula, outputs the set of states of the model that satisfy the given formula.

The following will be given the thoughts of the labeling of the CTL formula EU.

E [p1 U p2], suppose p1 and p2 are the sub formulas of the CTL formula satisfying all the immediate sub formulas of p1 and p2 have already been labeled.

1) If any state is labeled with p2, then label it with E [p1 U p2].

   Label any state with E [p1 U p2] if it is labeled with p1.

2) Repeat: If any state is labeled with p1 and at least one of its successors is labeled with E [p1 U p2], label it with E [p1 U p2]. Repeat this until there is no change. Figure 39.2 illustrates this step.

The CTL model checking algorithm based on MapReduce needs to realize the map and reduce functions.

Map Function

Map function is designed to deal with all the states of the CTL model and output the relative result after the received key/value pair's treatment. If any state is labeled, then the map function must produce outputs to update the state's pre-successors' relative value (Fig. 39.3).

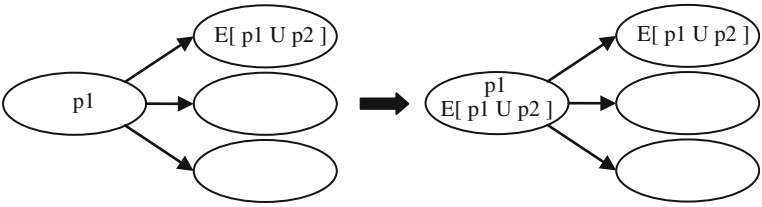


Fig. 39.2 The iteration step of the procedure for labeling states with the sub formula E [p1 U p2]

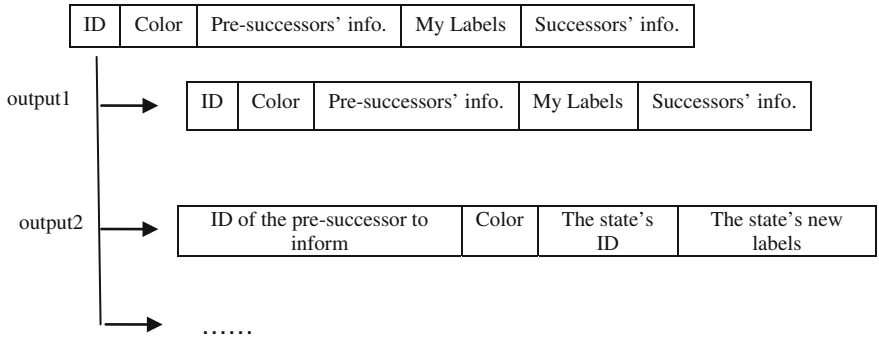


Fig. 39.3 Map function' input and outputs

The map function for the CTL formula EU is given below.

Algorithm 1. The Map Function for the CTL formula EU

```
/*Vi is the processing state*/
Input: <Vi, v>
Output: the changed state Vi information; the update information
If the state is labeled with EpUq then
  Output the state
  Return
End if
If the state is labeled with q then
  Label it with EpUq
  Output the state
  Output the update information which Inform all pre-successors of the state to update
  MapReduce counter adds 1
  Return
End if
If the state is labeled with p then
  If one of its successors is labeled with EpUq then
    Label the state with EpUq
    Output the state
        Inform all pre-successors of the state to update
        MapReduce counter adds 1
        Return
  End if
End if
Output the state
Return
```

Reduce Function

After the map phase is over, the Reduce function is designed to process the result set of the map function. If there is no update information to this state, then output the state. Otherwise update the value of the state's successors' information.

The reduce function is given in Algorithm 2.

The iteration procedure of MapReduce

This algorithm adopts MapReduce jobs chain for iteration. MapReduce jobs are chained to run sequentially, with the output of previous MapReduce job being the input to the next. The process is controlled by a MapReduce counter and a non-MapReduce driver program that checks for termination. And the iterative process should be terminated when the MapReduce counter is zero.

39.4 Experiment Studying

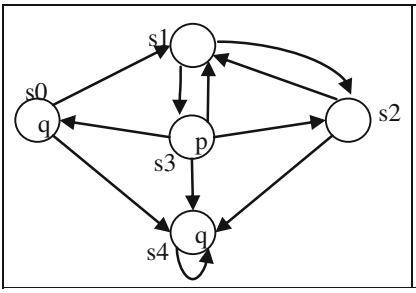
Algorithm 2. The Reduce Function for the CTL formula EU

```
/*Vi is the processing state*/
Input: <Vi, list<v>>
Output: the state Vi information
For each value of Vi do
  Judge the color and store the values
End for
If all colors are white then
  Output the state itself
Else
  Update the information of the state: modify the changed state's labels according to t
information of the red color values
  Output the state
End if
Return
```

For example, consider the following diagram of the system in Fig. 39.4, and the data are defined as follows:

According to the small amount data of the model, Hadoop runs in the pseudo-distributed mode. And the algorithm is realized by java language. Consider the system in Fig. 39.4 to compute the set [[E (T U p)]]. Here T is defined as any state of the system. From Fig. 39.4, it's easily obtained that [[p]] = {s3} and the state s3 is one of the state s1's successors. And s0 and s2 can reach s3 via s1. This can be obtained that [[E (TU p)]] = {s0, s1, s2, s3}.

The following gives the first iterative procedure of the CTL model-checking algorithm. In the Map phase, the state s3 is labeled with p, then label it with E (T U p) and output it. After that, output the information to inform the state s1, one of the pre-successors of s3, to update. Any other state is labeled without p, and none of their successor is labeled with E (T U p), so we just output these states.



The corresponding data of Fig. 39.4 for MapReduce

s0	WHITE|s3|q|s1.null;s4.q;
s1	WHITE|s0,s2|null|s2.null;s3.p;
s2	WHITE|s1,s3|null|s1.null;s4.q;
s3	WHITE|s1|p|s0.q;s2.null;s4.q;
s4	WHITE|s0,s2,s3,s4|q|s4.q;

Fig. 39.4 A system, compute the states satisfying E(T U p)

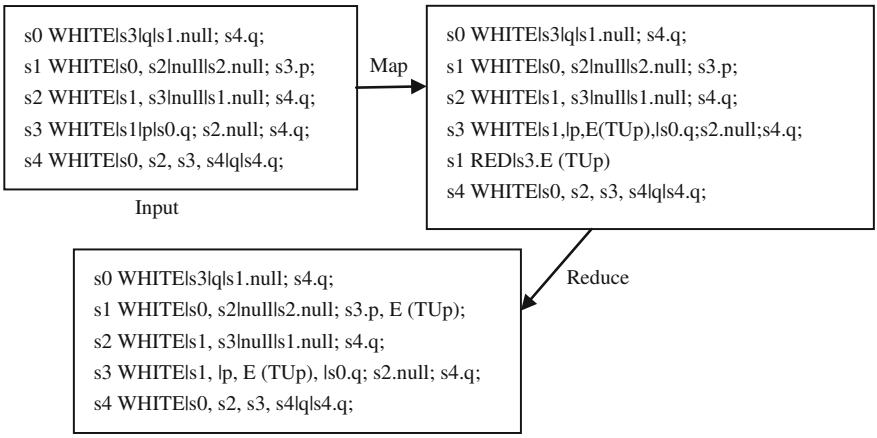


Fig. 39.5 The first iterative procedure

Then in the Reduce phase, the key-value list is handled. If any status flag of the value list is RED, then update the information of this state. In the first iteration, the state s1's information is processed and its successor, s3, adds new label E (T U p). The first iterative procedure is given in Fig. 39.5.

For the model of Fig. 39.4, to compute the states satisfying E (T U p), the results in each iteration are given below.

By analyzing experiment result and relative theory, it's obtained that the states s0, s1, s2, s3 are labeled with E (TU p). For the system of Fig. 39.4, the state set {s0, s1, s2, s3} satisfies the CTL formulas [[E (TU p)]]. The results are consistent with the previous analysis and shows that the CTL model-checking algorithm based on MapReduce is feasible.

	The result in each iteration

The 2nd iterative procedure	s0 WHITE|s3,|q,|s1.null,E(TUp);s4.q;
	s1 WHITE|s0,s2,|null,E(TUp),|s2.null;s3.p,E(TUp);
	s2 WHITE|s1,s3,|null,|s1.null,E(TUp);s4.q;
	s3 WHITE|s1,|p,E(TUp),|s0.q;s2.null;s4.q;
	s4 WHITE|s0,s2,s3,s4|q|s4.q;

The 3th iterative procedure	s0 WHITE|s3,|q,E(TUp),|s1.null,E(TUp);s4.q;
	s1 WHITE|s0,s2,|null,E(TUp),|s2.null,E(TUp);s3.p,E(TUp);
	s2 WHITE|s1,s3,|null,E(TUp),|s1.null,E(TUp);s4.q;
	s3 WHITE|s1,|p,E(TUp),|s0.q,E(TUp);s2.null,E(TUp);s4.q;
	s4 WHITE|s0,s2,s3,s4|q|s4.q;

The 4th iterative procedure	s0 WHITE|s3,|q,E(TUp),|s1.null,E(TUp);s4.q;
	s1 WHITE|s0,s2,|null,E(TUp),|s2.null,E(TUp);s3.p,E(TUp);
	s2 WHITE|s1,s3,|null,E(TUp),|s1.null,E(TUp);s4.q;
	s3 WHITE|s1,|p,E(TUp),|s0.q,E(TUp);s2.null, E(TUp);s4.q;
	s4 WHITE|s0,s2,s3,s4|q|s4.q;

39.5 Conclusion and Future Work

The paper proposes a new solution to the state explosion problem in model checking. This solution is based on MapReduce programming framework. For the future work of the approach described in this report, the performance of the presented algorithm under the large data should be further studied and be compared and evaluated against other benchmarks in distributed model checking techniques. Researchers will continue to study and improve the algorithm, or modify the Hadoop frame to solve the state space explosion problem.

Acknowledgments This work was supported by the National Natural Science Foundation of China (No.61070030, No.61111130121); Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality numbered PHR201107107; Plan of Beijing College Students' research and entrepreneurial action.

References

1. Apache.: Welcome to Apache Hadoop[EB/OL]. http://hadoop.apache.org/ (2012)
2. Baier, C., Katoen, J.P.: Principles of Model Checking (Vol. 26202649). MIT Press, New York (2008)
3. Bourahla, M.: Distributed CTL model checking. Software, IEEE Proceedings-IET. **152**(6): 297–308 (2005)
4. Clarke, E.M., Grumberg, O., Peled, D A.: Model Checking. MIT press, Cambridge (2000)
5. Huth, M., Ryan, M.: Logic in Computer Science: Modelling and Reasoning About Systems, 2nd edn. Cambridge University Press, Cambridge (2004)
6. Kripke Structure (model checking)-Wikipedia, the Free Encyclopedia [EB/OL]. http://en.wikipedia.org/wiki/Kripke_structure_(model_checking) (2012)
7. Lam, C.: Hadoop in Action. Manning Publications Co, Greenwich (2010)
8. White, T.: Hadoop: The definitive guide, O'Reilly Media, Sebastopol (2012)

Chapter 40
LBSG: A Load Balancing Scenario Based on Genetic Algorithm

Shan Jin and Wei Zhou

Abstract Resource load balancing problem of the large-scale and heterogeneous network was studied. The problem was modeled and analyzed theoretically at first, and an objective function which satisfied the host and network constraints was designed. Then, a multi-objective minimum spanning tree problem was researched, and then a multi-objective genetic algorithm was devised. At last, a dynamic load balancing scenario was proposed. The simulation results show that, LBSG can balance the load effectively between the light-load nodes and the overload ones. Besides, the performance of the scenario is obviously better in a larger scale network.

Keywords Load balancing · Multi-objective · Genetic algorithm · Distributed

40.1 Introduction

With the rapid development of the computer and network technology, the parallel and distributed computing becomes more and more popular. Meanwhile, the reduced cost of computer hardware enables network-based cluster and grid computing attract more and more researchers. However, the actual utilization rate of the potential performance of these systems is usually less than 10 % [1] which makes the system operate inefficiently. In the face of the growing system visits and the increasing complex computer needs, and combined with the technical and economic aspects of the two facts of the constraints, it appears more and more low prices that depend on improving the performance of a network servers to solve efficiency problem. Distributed system, which uses the effective loading balance, is a key technology to improve the system resource utilization and the performance of parallel computing [2].

S. Jin (✉) · W. Zhou
Center of Information Technology, China Nuclear Power Technology Research Institute, China Guangdong Nuclear Power Holding Co., Ltd, Shenzhen, China
e-mail: jinshan@cgnpc.com.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_40,

In large-scale distributed system model, the heterogeneity, opening and sharing of the network make it difficult to compute the load of calculation and communication in advance. Therefore, how to realize the logical data resources effective load balancing research becomes one of the difficult problems [3–7]. In view of the traditional resource load balancing scheme is insufficient, a distributed dynamic resource load balancing scheme named LBSG (Load Balancing Scenario based on Genetic algorithm) is designed.

40.2 Model and Description

40.2.1 Mathematical Model

LBSG constructs a network topology using the multi-criteria minimum spanning tree [8] as a basic model. The two-dimensional plane can be depicted as a complete graph, $G = (V, E)$, where $V = \{v_1, v_2,\ldots,v_n\}$ is the set of vertices represented the network node of the system and $E = \{e_1, e_2, \ldots, e_n\}$ is the set of edges represented link between the network nodes. Each edge e_i is associated with q weights, $\Omega_i = \{\omega_{i1}, \omega_{i2}, \ldots, \omega_{iq}\}$ $(i = 1, 2,\ldots, m)$.

Considering the practical application of feasibility and the mathematical modeling of convenience, we set single hop routing delay threshold of node communication for t_{max}. Take node v_i for example, we define the local coverage for a two-dimensional circular surface in graph G_i, $G_i = (V_i, E_i)$, where $\mathrm{v_i}$ is the center of a circle and t_{max} is the radius of it. G_i within any spanning tree is denoted as $\mathrm{T_i}$. First, the whole network is divided into several local coverage of G_i by LBSG. Second, the local topology will be constructed in accordance with the multi-objective minimum spanning tree for independent in every G_i. Finally, the global topological structure of the network achieves approximate optimization.

40.2.2 Objective Function

For the purpose of constructing a reliable, efficient, scalable distributed system, we need to comprehensive consider node resources, link delay, and network communication interference conditions. Not only the multi-objective optimization has the peak value, but also a variety of constraints obtain a reasonable compromise. For any local coverage domain G_i, where $V_i = \{v_{i1}, v_{i2},\ldots, v_{in}\}$, $E_i = \{e_{i1}, e_{i2},\ldots, e_{im}\}$,$X_i = (x_{i1}, x_{i2},\ldots, x_{im})$, the objective function is defined in formula (40.1).

To link e_j, the used bandwidth is depicted as $\omega_{j1} = deg(x_j)$, the communication delay of such link is depicted as $\omega_{j2} = time(x_j)$, the delay jitter value of such link network is depicted as $\omega_{j3} = jet(x_j)$, and the loss rate of data transmission of such link is depicted as $\omega_{j4} = loss(x_j)$.

$$\begin{cases} \min z_1(X_i) = \sum_{j=i1}^{im} \omega_{j1} \cdot x_j \\ \min z_2(X_i) = \sum_{j=i1}^{im} \omega_{j2} \cdot x_j \\ \min z_3(X_i) = \sum_{j=i1}^{im} \omega_{j3} \cdot x_j \\ \min z_4(X_i) = \sum_{j=i1}^{im} \omega_{j4} \cdot x_j \\ s.t. \quad X_i \in S(T_i) \end{cases} \tag{40.1}$$

40.3 Design and Analysis of LBSG

40.3.1 Spanning Tree Coding

According to Cayley's theory, there are n^{n-2} trees in a graph which contains n nodes, therefore it can be used n-2 number arranged to only express on a tree and each number is an integer between 1 to n, this arrangement is called Prufer number [8]. This style of coding can be uniquely represented in a graph of all possible spanning tree and still represents a tree in any crossover or mutation operation. Therefore, LBSG using Prufer is expressed as the number of problems of all feasible solutions.

40.3.2 Fitness Function

A compromise method which is not directly and positively to describe how to weight the most ideal, but to identify with the ideal solution of recent solution according to a distance measure is used in LBSG. The mathematical description of the corresponding called regret function. In accordance with the L_p, the expression forms of such function is defined in formula (40.2).

$$r(Z,p) = \|Z - Z^*\|_p = \left[\sum_{j=1}^{q} \left|z_j - z_j^*\right|^p\right]^{1/p} \tag{40.2}$$

Among them, q is the target number, p is norm parameter, z_j is the j target value, and z_j* represent the ideal value of the j target. The expression is $z_j^* = z_j^{\min}$.

The parameter p on behalf of great regret value concern, the larger value the more care of great regret value. For each target regret value can be considered [9], every regret value must be firstly changed to one order of magnitude through the normalization method, making its value are all between [0, 1], and then select the appropriate parameter p. Accordingly, the regret function of type (40.1) which typed as the form defined of (40.2) is as follows:

$$r(Z) = \left\| \frac{Z - Z^*}{Z_{\max} - Z_{\min}} \right\|_p = \left[\sum_{j=1}^{4} \left| \frac{z_j - z_j^* + \gamma}{z_j^{\max} - z_j^{\min} + \gamma} \right|^p \right]^{1/p} \tag{40.3}$$

Among them, parameter γ is the positive real numbers which between (0, 1), this function not only can avoid producing the mistake of except by 0 of formula (40.3), but also can be adjusted to a purely random selection. For many complex problems, it is difficult to find the ideal point. Therefore, we change Z^* into agents of ideal point to replace the ideal point. The agent ideal point is not given problems of ideal point, but the current corresponding to the ideal point. If we let P indicates the kind of cluster, the agent ideal point is calculated as follows:

$$\begin{cases} Z^* = (z_1^{\min}, z_2^{\min}, z_3^{\min}, z_4^{\min}) \\ z_1^{\min} = \min\{z_1(X) | X \in P\} \\ z_2^{\min} = \min\{z_2(X) | X \in P\} \\ z_3^{\min} = \min\{z_3(X) | X \in P\} \\ z_4^{\min} = \min\{z_4(X) | X \in P\} \end{cases} \tag{40.4}$$

Obviously, we get the value, $z_j^{\min} = \min \{z_j(X) \mid X \in P, j = 1, 2, 3, 4\}$, $z_j^{\max} = \max \{z_j(X) \mid X \in P, j = 1, 2, 3, 4\}$.

As for the minimize the problem, the smaller regret value the better, so we need to change the regret function into the fitness function to ensure excellent individuals with greater fitness value. Define the fitness function is as follows:

$$eval(Z(X)) = \exp\left[\frac{r_{\max} - r(Z(X))}{r(Z(X)) - r_{\min}} \right] \tag{40.5}$$

Among them, r_{max} and r_{min} respectively are representing the current generation of maximum value and the minimum value of regret.

40.3.3 Genetic Operator Designing

In solving the minimum spanning tree problem, searchers often use roulette (proportional) method [10] to select individual. The process of crossover was randomly paired as parents for all individuals within groups and exchanged both subsets of chromosomes with a certain probability P_c, which resulting in the

formation of two new generations. LBSG use single point crossover which product a random crossover point to exchange the whole intersection gene of the right end of the parents and finally form two individual. Prufer number coding is still a tree after this single point crossover. The mutation is that the mutation probability P_m alters one or several loci genetic values for the other allele in each individual of the groups. LBSG uses the exchange variation and the parent code string to randomly select two position changes after the formation of a new generation.

40.3.4 Topology Control and Load Balancing

The pseudo code of every topology control in LBSG is shown in Fig. 40.1. LBSG uses the method which based on event and message to perform load balancing operations [11].

40.4 Simulation and Evaluation

40.4.1 Simulation Setup

In order to verify the performance and efficiency of the algorithm, we carry out the simulation experiments. The underlying network uses random topological structure and the network delay between nodes is proportional to their geometry distance. Node load is the consumption of resources during its work and it can be calculated capacity, bandwidth, I/O or storage space, and other indicators or these indexes. This paper abstracts the node load to the concept of value. The maximum degree of each node is randomly distributed between 15 and 25. The initial degree is randomly distributed between 2 and 6, and other experimental parameters as follows: $t_{max} = 20$, $\deg(x_j) = [100, 200]$, $\text{time}(x_j) = [5, 50]$, $jet(x_j) = [10, 40]$, $loss(x_j) = [5\ \%, 10\ \%]$, $p = 2$, $\gamma = 0.4$, $P_c = 0.6$, $P_m = 0.003$, size $= 4$, Maxgen $= 100$.

Fig. 40.1 Pseudo code of topology control

```
1   WHILE (v_i ∈ V)
2     init ();
3     for (k = 1; k <= Maxgen; k = k +1)
4    evaluate();
5    for ( j = 2; j <= size; j = j + 2)
6       f1 = selection(); f2 = selection();
7       crossover(f1, f2);
8       mutation();
9     if (Z_min(X) < Z_min(X_0))
10          X_0 = X;
```

- Experiment 1

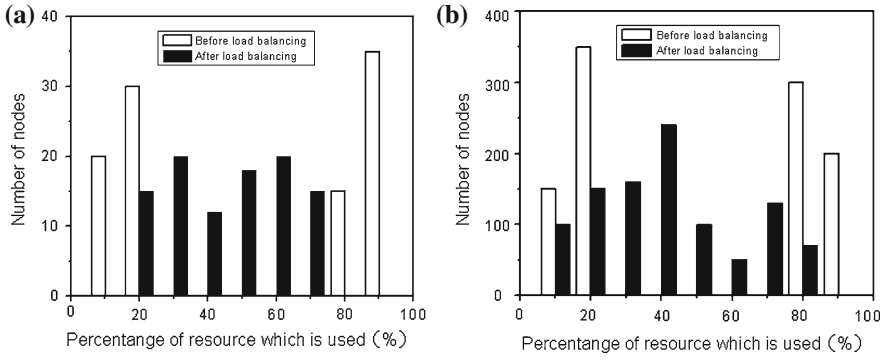


Fig. 40.2 Contrast of occupancy rate of node resource. **a** $l_v = 50$, $h_v = 50$, **b** $l_v = 500$, $h_v = 500$

40.4.2 Simulation Results and Analyses

- Experiment 1

As to all the initialized nodes in the network, we use a random algorithm to increase the resource load first, and then produce the specified number of light load and heavy load node. Based on the operation of LBSG load balancing, we finally inspect the similarities and differences on the distribution of the occupancy rate of node resources before and after equilibrium. Among them, the light load is defined as the resource occupancy rate less than 20 %, the heavy load is defined as the resource occupancy rate higher than 80 %. The experiment was divided into two groups, the first group generate a 50 light loaded nodes and 50 overloaded node ($l_v = 50$, $h_v = 50$), the second group generate 500 light loaded nodes and 500 overloaded node ($l_v = 500$, $h_v = 500$). As in Fig. 40.2, no matter the node size is 100 or 1000, and most of the occupancy rate of node resource can adjust from 20 to 80 % which proved the validity of algorithm.

- Experiment 2

Firstly, aiming at running the random network topology and using the control algorithm to optimize the topology structure of operation. Secondly, making sure that all nodes have been using degrees were increased into any integer by 0–10, and judges its resource usage. If the occupancy rate greater than or equal to 80 %, we need to move half of the degree of use (to be rounding) to $Q(v_i)$ of node mobility. In third, re-perform the topology control algorithm to form a new network topology after the completion of migration movement. It should be repeated 3 times about the two and three step as a complete experiment. Here, the $Q(v_i)$ is the nodes set which resource occupancy rate is below 60 %, λ and γ in DLBM is set to 0.7 and 0.4.

In Fig. 40.3, with the constant enlargement of the scale of the LBSG node, the performance gap increases. The reason is that density of plane node distribution

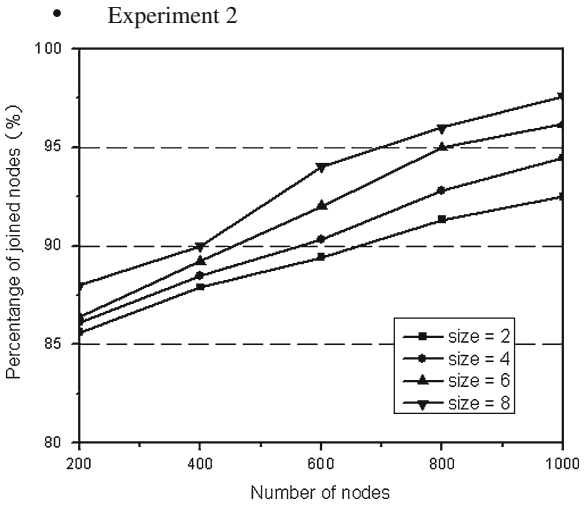


Fig. 40.3 Nodes participation rate versus group scale in different scale

becoming greater brings the increase of the spanning tree corresponding to the feasible solution. However, the expansion of the scale of population will inevitably bring the increase of time complexity, so it is not be better. In the practical application, we need to be weighed against the various factors to determine the appropriate population scale.

40.5 Conclusion

LBSG comprehensively assesses the object in the calculation of the objective function such as the link resources, the network response speed, the delay jitter and the packet loss rate of transmission. From the simulation results, the LBSG has a significant load balancing effect and other equilibrium strategy compared with the obvious advantages. At the same time, LBSG also showed good adaptability along with the expansion of the scale of network node. It is shown that LBSG is more suitable for deployment in large and complex network system. In future work, the optimal parameters will be discussed by some theoretical proof and experiment methods.

References

1. Yang, X.J., Dou, Y., Hu, Q.F.: Progress and challenges in high performance computer technology. J. Comput. Sci. Technol. **21**(5), 674–681 (2006)
2. Zheng, G B.: Achieving High Performance on Extremely Large Parallel Machines: Performance Prediction and Load Balancing. UIUC, Urbana (2005)

3. Yasar, O.: Trends in parallel computing. Parallel Comput. **33**(2), 81–82 (2007)
4. Devine, K.D., Boman, E.G., Heaphy, R.T., Hendrickson, B.A.: New challenges in dynamic load balancing. Appl. Numer. Math. **52**(2–3), 133–152 (2005)
5. Willebeek-LeMair, M.H., Reeves, A.P.: Strategies for dynamic load balancing on highly parallel computers. IEEE Trans. Parallel Distrib. Syst. **4**(9), 979–993 (1993)
6. Stankovic, J.A., Sidhu, I.S.: An adaptive bidding algorithm for process, clusters and distributed groups. In: Proceedings of International Conference on Distributed Computing Systems, New York, USA: IEEE, pp. 49–59 (1984)
7. Bryant, R.M., Finkel, R.A.: A stable distributed scheduling algorithm. In: Proceedings of International Wire and Cable Symposium, pp. 314–323. Comput Soc Press, Los Alamitos, California, USA (1981)
8. Zhou, G., Gen, M.: Genetic algorithms approach to the multi-criteria minimum spanning tree problem. Eur. J. Oper. Res. **114**, 141–152 (1999)
9. Ishibuchi H, Nakaskima T. Three-objective optimization in linguistic function approximation. In: Proceedings of the 2001 Congress on Evolutionary Computation, pp. 340–347. Seoul, South Korea: IEEE (2001)
10. Goldberg, D.E.: Genetic Algorithm in Search, Optimization and Machine Learning. Addison-Wesley, New York (1989)
11. Jin, S., Ren, B.: A novel distributed dynamic load balancing mechanism. In: Proceedings of the 2011 International Conference of Information Technology, Computer Engineering and Management Sciences, vol. 2, pp. 133–137. Nanjing, China: IEEE, (2011)

Chapter 41
Improved Ant Colony Algorithm for the Constrained Vehicle Routing

Guiqing Liu and Dengxu He

Abstract Using the basic ant colony algorithm to solve the constrained vehicle routing problem (CVRP) has some drawbacks such as slow convergence speed and easily getting into local optimum. To effectively solve the CVRP, this paper has proposed a new ant colony algorithm (ACA-CVRP) based on the dynamic update of local and global pheromone and improved transfer rule. In order to shorten the process, the authors introduced the candidate list and 2-opt searching strategy. The experiment result shows that ACA-CVRP achieves better performance in optimum solution compared with other five main meta-heuristic algorithms.

Keywords Ant colony algorithm · Pheromone update · 2-opt · Candidate list · CVRP

41.1 Introduction

Ant colony algorithm (ACA), which is a new swarm intelligence algorithm characterized with positive feedback system, easy to calculate and integrate with other algorithm. It has grabbed attention among scholars domestic and overseas since its appearance. The ACA has succeeded in solving optimization problems such as TSP, scheduling, and VRP and so on [1]. However, as a classic ACA, ant colony system has encountered troubles like slow convergence speed, easily getting into local optimum or stagnation behaviour when solving realistic problems. At present, the main methods to solve VRP are tabu search [2, 3], ACA, simulated

G. Liu (✉)
College of ASEAN, Guangxi University for Nationalities, Nanning, China
e-mail: lgqlucker@163.com

D. He
College of Science, Guangxi University for Nationalities, Nanning, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_41,

annealing algorithm [4] and genetic algorithm. Up to now, the Osman's TS algorithm is the most prominent one to solve the VRP considered vehicle capacity, routing distance and rich in relevant literature information. Though less relevant literature than TS, the ACA [5–7] has shown considerable potential in settling the VRP. Through the weakness analysis of ant colony system (ACS) [8], this paper has proposed an improved ACA to solve the CVRP basing on the dynamic update of local and global pheromone and improved transfer rule.

41.2 Mathematical Model of CVRP

$$Min \sum_{i=0}^{N} \sum_{j=0}^{N} \sum_{k=1}^{N} c_{ij} x_{ij}^{k} \tag{41.1}$$

CVRP (vehicle routing problem under the constraints of vehicle capacity and distance) means to meet requirements of all customers around the single depot aiming at the shortest distance [9]. c_{ij} means the cost from station i to station j. t_{ij} represents the known distance from station i to station j. s_i represents the given service time of station i. Q^k represents the fixed capacity of vehicle k. T^k represents the longest running distance of vehicle k. d_j represents given the quantity demand of station j. N and K respectively represent the amount of station and vehicle.

41.2.1 Constrained Conditions

$$\sum_{i=1} \sum_{j=1} x_{ij}^{k} d_i \le Q^k \quad 1 \le k \le K \tag{41.2}$$

$$\sum_{i=0}^{N} \sum_{j=0}^{N} x_{ij}^{k} (t_{ij} + s_i) \le T^k \quad 1 \le k \le K \tag{41.3}$$

$$\sum_{i=0}^{N} x_{ij}^{k} - \sum_{i=0}^{N} x_{ji}^{k} = 0 \quad 1 \le k \le K,\ 0 \le j \le N \tag{41.4}$$

$$\sum_{i=0}^{N} \sum_{k=1}^{K} x_{ij}^{k} = \begin{cases} 1 & 1 \le j \le N \\ K & j = 0 \end{cases} \tag{41.5}$$

$$x_{ij}^{k} \in \{0, 1\}\ \ 1 \le k \le K,\ \ 0 \le i,\ j \le N \tag{41.6}$$

Formula (41.2) means the capacity of every vehicle.

Formula (41.3) defines the longest distance each vehicle can run. The total length concludes distance between all the stations and the equivalent distance when one vehicle staying at the station.

Formula (41.4) represents that the vehicle arrives at station i, it must leave i.

Formula (41.5) defines the vehicle must stop at each station (one vehicle at a time). All the vehicles must return to the depot.

Formula (41.6) defines x_{ij}^k is a quadratic function, when vehicle k travels from station i to station j, $x_{ij}^k = 1$, otherwise $x_{ij}^k = 0$.

41.3 Basic Principle of ACS

Let n mean the amount of cities. m means the amount of ants.d_{ij} means the distance between city i and city j. $\eta_{ij}(t) = 1/d_{ij}$ means the ant's degree of expectancy transfer from city i to city j at time t. α, β means the importance parameter of pheromone strength τ and heuristic information when ants selecting routing. At the initial moment of the algorithm, put ants (amount m) to cities (amount n) randomly. Set pheromone amount of every routing $\tau_{ij}(0) = c$ ($c$is constant). Each ant select the transfer city according to pheromone amount along the routing. Ant k selects city j according formula (41.7) (41.8) at time t.

$$j = \begin{cases} \arg \max\limits_{s \in allowed_k} \left\{ [\tau_{is}(t)]^{\alpha} [\eta_{is}(t)]^{\beta} \right\} & if\ q \le q_0 \\ J & otherwise \end{cases} \tag{41.7}$$

$$P_{ij}^k = \begin{cases} \dfrac{[\tau_{ij}(t)]^{\alpha}[\eta_{ij}(t)]^{\beta}}{\sum\limits_{s \in allowed_k} [\tau_{is}(t)]^{\alpha}[\eta_{is}(t)]^{\beta}}, & if\ j \in allowed_k \\ 0 & otherwise \end{cases} \tag{41.8}$$

$q_0 \in (0, 1)$ a constant, $q \in (0, 1)$ a random number, $allowed_k$ means a set which ant k can select city j from. If $q > q_0$, ant k will select city j randomly according to formula (41.8).

When one ant transfers from city i to city j, local pheromone will be updated according to formula (41.9).

$$\tau_{ij}(t+1) = (1-\varepsilon)\ \tau_{ij}(t) + \varepsilon\tau_0 \quad \varepsilon \in (0,\ 1) \quad \tau_0 \text{ is Constant} \tag{41.9}$$

When every ant goes through all these cites, global pheromone will be updated for the best solution according to formula (41.10).

$$\tau_{ij}(t+n) = (1-\rho)\tau_{ij}(t) + \rho\Delta\tau_{ij}^{gb} \quad \rho \in (0,\ \ 1\ ) \tag{41.10}$$

$$\Delta\tau_{ij}^{gb} = \frac{1}{L_{gb}}$$

L_{gb} is the global best solution.

41.4 Improvement of ACA

41.4.1 New Transfer Rule

Just like greedy algorithm, ants are led by Pheromone concentration and distance between cities in the process of transferring. At the beginning, distance between cities plays the main role. With the application of positive feedback system, pheromone is getting increasingly important. As a result, possibility of local optimum and stagnation behavior is getting greater. In order to balance the factor role between τ_{ij} and η_{ij} during the process of optimizing, we replace τ_{ij} with τ_{ij}/τ_0 (marked τ^*_{ij}). The formula (41.7) and (41.8) were respectively turned into formula (41.11) and (41.12),

$$j = \begin{cases} \arg \max\limits_{s \in allowed_k} \{[\tau^*_{is}(t)]^{\alpha}[\eta_{is}(t)]^{\beta}\} & if \quad q \le q_0 \\ J & otherwise \end{cases} \tag{41.11}$$

$$P^k_{ij} = \begin{cases} \dfrac{[\tau^*_{ij}(t)]^{\alpha}[\eta_{ij}(t)]^{\beta}}{\sum\limits_{s \in allowed_k} [\tau^*_{is}(t)]^{\alpha}[\eta_{is}(t)]^{\beta}}, & if\ j \in allowed_k \\ 0 & otherwise \end{cases} \tag{41.12}$$

41.4.2 Dynamic Update of Local Pheromone

In the ACS, the variant of local pheromone updating is a constant τ_0. By analyzing the solution of the ant paths we find that few cities are passed by ants at the beginning when the possibility of building optimum solution is relatively large because of large range of free choice. In the end the number of passed cities is larger and larger. Due to the restriction of taboo lists, the optional range of next city is becoming smaller and smaller. Based on the analysis above, local pheromone was updated according to formula (41.13):

$$\tau^{new}_{ij} = (1 - \varepsilon)\tau^{old}_{ij} + \varepsilon \Delta\tau_{ij} \tag{41.13}$$

$$\Delta\tau_{ij} = \max(c) - S\frac{\max(c) - \min(c)}{n}$$

where $\varepsilon \in (0, 1)$ is a local evaporation coefficient of phermone, $\max(c)$ and $\min(c)$ are maximum and minimum amount of pheromone updating respectively, S is the city number of taboo list, n is the city total, $\Delta\tau_{ij}$ is a variant which is decreasing linearly from the maximum.

41.4.3 Dynamic Update of Global Pheromone

Only the best path is updated globally for ACS compared with AS when all the ants complete their cycles. Since the difference of pheromone amount on the ant paths is increasing, it is easy to fall into local optimum and produce stagnation in ACS. Based on the analysis above, strengthening positive feedback mechanism can speed up the convergence of the current optimal solution, but easily lead to premature phenomenon. To balance this contradiction, we firstly compared the L_{ib} with L_{gb}^{*} in this paper and classify secondly according to strengths and weaknesses of L_{ib}. Gobal pheromone was dynamically updated according to formula (41.14)

$$\tau_{ij}(t+n) = (1-\rho)\tau_{ij}(t) + \rho\Delta\tau_{ij} \tag{41.14}$$

$$\Delta\tau_{ij} = \sum_{k=1}^{x} \omega_k Q / L_k \tag{41.15}$$

$$\omega_k = \begin{cases} 2 - (L_k/L_{gb}^{*})^{p} & if\ (L_{ib} < L_{gb}^{*})\ and\ (L_k < L_{gb}^{*}) \\ (\overline{L} - L_k)/\overline{L} - 2 & if\ (L_k = L_{ib} = L_{gb}^{*})\ and\ (\mathrm{T}_{gb}^{same} \geq num) \\ 1 - (L_k/\overline{L}\,) & if\ (L_k = L_{ib} = L_{gb}^{*})\ and\ (\mathrm{T}_{gb}^{same} < num) \\ 0 & else \end{cases} \tag{41.16}$$

where x is the path number satisfying global updating in this cycle, L_k is the path length of ant-k, $\overline{L}$ is the average path length of this cycle, L_{ib} is the iteration-best, L_{gb}^{*} is the global-best until the last iteration, T_{gb}^{same} is the number of the same global-best, *num* and p are two parameters which are determined respectively by the scale of the problem and the total iteration number, ω_k is the global pheromone updating weight of ant-k which is explained from the following two aspects.

If L_{ib} is smaller than L_{gb}^{*}, this searching effect is better than the last. In order to make full use of current information and search the next solution in the vicinity of L_{ib}, the positive feedback mechanism of better paths are treated differently in this cycle. Pheromone is adaptively globally updated by weight if the path satisfies the condition of $L_k < L_{gb}^{*}$. The smaller L_k is, the larger ω_k is, the smaller on the contrary, which can make further search for a better solution around the current optimum solution. Because updated paths are not only the optimal path, this method can effectively prevent the phenomenon of local optimum.

If L_{ib} and L_{gb}^{*} are equal, this iteration doesn't change the global-best and this searching effect is bad. If the number of same global-best reach a certain number *num* the algorithm is likely to run into local optimum. We judge the polymerization of the ant paths by calculating the difference between L_{ib} and $\overline{L}$ (the overall similarity between ant paths and the Iterative-best, the greater the similarity is, the greater the degree of polymerization is, conversely the smaller). If $\mathrm{T}_{gb}^{same} \geq num$, pheromone of path L_k will be cut by different degrees, according to the degree of

polymerization. The closer the difference is between $\overline{L}$ and L_{ib}, the greater the degree of polymerization of the entire ant colony is, conversely the smaller. If $T_{gb}^{same} < num$, pheromone of path L_k is strengthen by different degrees according to the degree of polymerization.

41.5 Main Idea of ACA-CVRP

In order to save calculating time and exclude bad routings, we introduce the candidate list when ants select stations. The candidate list is established according to the neighboring method and ascending ranking of d_{ij}. The length of candidate list influences the solution and calculating time a lot. In the process of routing construction, every ant represents for a complete routing. If $q \leq q_0$, we will pick the unselected stations in the list. If the value of $(\tau_{is}^*)^{\alpha}\eta_{ij}^{\beta}$ is maximum, we will select the routing. Otherwise, if $q > q_0$, we select the random city j according to probability distribution of formula (41.8). Considering vehicle capacity and the limited distance, if the solution violates it, we pick the second best station to replace. If all the stations are in the routing, ants will return to the depot and then restart.

In order to increases chances of finding other feasible solution, local pheromone will be updated on line according to formula (41.13). After all the ants have visited all the stations, we will calculate the routing distance of each ant and use 2-opt to locally search the shortest routing until the solution remains stable. Global pheromone will be dynamically updated according to formula (41.14) and the pheromone concentration of every routing is limited in the range of $(\tau_{\min}, \tau_{\max})$ to avoid the pheromone too dense or thin.

41.6 Analysis of Simulation

14 examples are selected from literature [2–4] for the simulation. The gaven parameters are shown in Table 41.1).

According to the description of Table 41.2 among the 14 standard experiments, SA has 2 of them succeeded getting the optimum solution while Osman'TS gets 4, tauroute gets 5, GTS gets 4 and the hybrid ACA [10] gets 6. The ACA-CVRP gets 8 optimum solutions among 14 standard experiments. If we compare the results of constrained by routing distance (C2, C5, C9, C13), ACA-CVRP has the best effect.

Table 41.1 Gaven parameters for simulation

Ant number	Pheromone amount	Length of candidate list	(α, β)	q_0	ρ
10	$\tau_0 = 10^{-4}$	$n/3$	(4, 1)	0.7	0.08

Table 41.2 Result comparison between ACA-CVRP and the other five meta heuristic algorithm

Examples	The smallest routing cost						
	SA	Taburoute	Osman's TS	GTS	Hybrid ACA	ACA-CVRP	Best published solution
C1	528	524.61	524.61	524.61	524.61	524.61	524.61
C2	838.62	835.77	844	838.6	837.55	835.26	835.26
C3	829.18	829.45	835	828.56	828.38	827.73	826.14
C4	1058	1036.16	1052	1033.21	1031.62	1030.83	1028.42
C5	1378	1322.65	1354	1318.25	1299.42	1291.45	1291.45
C6	555.43	555.43	555.43	555.43	555.43	555.43	555.43
C7	909.68	913.23	913	920.72	912.46	913.23	909.68
C8	866.75	865.94	866.75	869.48	865.94	865.94	865.94
C9	1164.12	1177.76	1188	1173.12	1162.55	1162.55	1162.55
C10	1417.85	1418.51	1422	1435.74	1409.29	1406.17	1395.85
C11	1176	1073.47	1042.11	1042.87	1044.30	1045.24	1042.11
C12	826	819.56	819.58	819.56	819.56	819.56	819.56
C13	1545.98	1573.81	1547	1545.51	1541.14	1541.14	1541.14
C14	890	866.37	866.37	866.37	867.57	867.42	866.37

Hybrid ACA is the second best; taburoute, Osman's TS, GTS and SA algorithm have bad solutions relatively.

41.7 Conclusion

This paper has improved the transfer rule, local and global pheromone update in allusion to the ACS's weakness of slow convergence speed and easily getting into local optimum in solving the constrained vehicle routing problem. Compared with taburoute, Osman's TS, GTS, TS and hybrid ACA, the ACA-CVRP has achieved better performance in optimum solution.

Acknowledgments The first author is supported by the Youth Fund Project of Guangxi University for Nationalities (No. 2011MDQN038) and the open project of China-ASEAN Studies Center of Guangxi University for Nationalities (No. 2012012).

References

1. Meng, Y., Jianshe, S., Jiping, C.: The overview of application research of ant colony optimization. Comput. Simul. **26**(6), 200–203 (2009)
2. Gendreau, M., Hertz, A., Laporte, G.: A tabu search heuristic for the vehicle routing problem. Manage. Sci. **40**, 1276–1290 (1994)

3. Toth, P., Vigo, D.: The granular tabu search and its application to the vehicle routing problem. Working Paper, DEIS, University of Bologna (1998)
4. Osman, H.: Meta strategy simulated annealing and tabu search algorithms for the vehicle routing problem. Ann. Oper. Res. **41**, 421–451 (1993)
5. Qingbao, Z., Zhijun, Y.: An ant colony algorithm based on variation and dynamic pheromone updating. J. Softw. **15**(2), 185–192 (2004)
6. Peng, Z.: An ant colony algorithm based on path similarity. Comput. Eng. Appl. **43**(32), 29–33 (2007)
7. Jie, Y., Sheng, Y.: An ant colony algorithm based on pheromone intensity. Comput. Appl. **29**(3), 865–867 (2009)
8. Dorigo, M., Gambardella, L.M.: Ant colony system: A cooperative learning approach to the traveling salesman problem. IEEE Trans. Evol. Comput. **1**(1), 53–66 (1997)
9. Doerner, K., Gronalt, M., Hartl, R.F., Reimann, M., Strauss, C., Stummer, M.: Savings ants for the vehicle routing problem. POM Working Paper 02/2002, Department of Production and Operations Management, University of Vienna (2002)
10. Xiao, Z.: Applications in the vehicle routing problem of hybrid ant colony algorithm. Comput. Eng. **37**(24), 190–192 (2011)

Chapter 42
Active Queue Management Mechanism Based on DiffServ in MPLS Networks

Yang Jiao and Li Du

Abstract Active Queue Management is the key to congestion control and enhancing IP QoS. MRED can support MPLS networks and take advantage of classifying mechanism of DiffServ and mark different businesses with different drop precedence. However, it has several limitations. In this paper, I-AMRED is proposed in order to reduce packet loss ratio and raise throughput in MPLS networks based on DiffServ and MRED. Thus, it can also adaptively control average queue length, and diminish sensitiveness to control parameters and improve stability with the service flow bursts. Two experimental schemes are designed and implemented on NS-2. Experimental results show that I-AMRED algorithm increases throughput of networks and largely decreases packet loss ratio of AF PHB traffic. When data traffic is very large or sharply increased, the performance of throughput and packet loss ratio is improved greatly, and has better stability.

Keywords DiffServ · MPLS · QoS · AQM · MRED

42.1 Introduction

In recent years, the continuous increase of traffic and service types has caused great demand in IP QoS and MPLS (Multi-Protocol Label Switching) DiffServ (Differentiated Service) model became dominant technology in solving IP QoS problems. The technological combination and interoperability [1] of MPLS and DiffServ is to take advantage of explicit routing and fast forwarding of MPLS and

Y. Jiao (✉) · L. Du (✉)
College of Information Science and Engineering, Northeastern University, Shenyang, China
e-mail: wyobj619@163.com

L. Du
e-mail: duli26@126.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_42,

the scalability of DiffServ, so choosing MPLS DiffServ environment has more practical meaning. Active Queue Management (AQM) is a mechanism which can drop packets by some strategy before router buffer is full and avoid global synchronization. RFC 2597 recommends that AQM mechanism be used to realize the multiple levels of drop precedence required in the AF PHB. MRED (Multi-level RED) [2] supporting DiffServ and MPLS is an AQM mechanism that can execute different RED policy and calculate drop probability independently for different drop precedence. In this paper, the limitations of MRED are researched and analyzed. Improved schemes are proposed and I-AMRED algorithm is raised. NS-2 platform is used to verify the effectiveness of I-AMRED algorithm. Comparisons of performance parameters such as throughput, packet loss ratio are presented between I-AMRED and MRED in two experiment schemes using NS-2.

42.2 The Limitations of MRED

MRED maintains multiple sets of RED thresholds [3]. With the continuous development of IP networks some limitations appear.

(1) Sensitiveness to RED Parameters Configuration: Configured RED parameters include the maximum threshold (max_{th}), the minimum threshold (min_{th}) and the maximum drop probability (max_p). MRED uses these configuration parameters un-self-adaptively. If RED parameters are configured improperly, the fluctuation of average queue length will become violent [4].
(2) Instability When Traffic Is Large or Sharply Increased: Essentially, linear relation between drop probability (P_b) and average queue length (avg_q) for each drop precedence leads to the instability of throughput value. When avg_q ranges from min_{th} to max_{th}, P_b will increase slowly with the increase of avg_q. When traffic sharply increases, queue buffer would have been filled up, accordingly, throughput will drop and link utilization will reduce [5].
(3) Insufficiency to Reflect Congestion Condition: MRED uses avg_q as unique evaluation parameter to reflect congestion change. However, avg_q can't always reflect congestion condition correctly. When weight value is low and the value of avg_q is high, the condition of congestion may be relieved. At this time, if deciding drop policy only according to avg_q, packet loss ratio will raise.

42.3 I-AMRED Algorithm

42.3.1 Basic Ideas of I-AMRED Algorithm

The purpose of I-AMRED is to remove the limitations of RED, and to accordingly reduce packet loss ratio.

(1) Import ARED Mechanism: In order to reduce sensitiveness to RED parameters configuration of MRED, import adaptive mechanism of ARED and make avg_q range from min_{th} to max_{th} as far as possible. Two judgment principles is introduced, that is reducing max_p value by dividing a factor α when congestion condition of networks is in a low level, and raising max_p value by multiplying a factor β when congestion condition of networks is in a high level. Factor α and β are both more than 1, and control RED policy should be more positive or more conservative by monitoring the change of avg_q every certain period.
(2) Change Linear Relation Between P_b and avg_q: To raise stability of MRED when traffic is large or sharply increased, linear relation between P_b and avg_q should be changed to exponential relation as in Fig. 42.1. The purpose of this change is to reduce P_b when avg_q approaches min_{th} (queue buffer of router is relatively idle) and to raise P_b when avg_q approaches max_{th} (queue buffer of router is almost filled up), accordingly relieving shortage of queue buffer of router before congestion and improving the stability of network throughput.
(3) Import Real-time Queue Length as Another Evaluation of Congestion Condition: To reflect congestion condition more sufficiently and reduce packet loss ratio, import real-time queue length ($qlen$) as another evaluation criterion of congestion condition of queue buffer of router. Three judgments are added:
    (1) $avg\_q \geq max_{th}$: If $qlen$ is less than max_{th}, use $qlen$ instead of avg_q to calculate drop probability, else, drop all arriving packets.
    (2) $qlen \geq max_{th}$: *No matter whether* avg_q *is more than* max_{th}, *drop all arriving packets.*
    (3) Use avg_q to calculate drop probability except the two conditions of the two judgments above.

When weight value of a certain Behavior Aggregate (BA) is very low and at this time congestion is relieved just now and avg_q is more than max_{th} but $qlen$ has

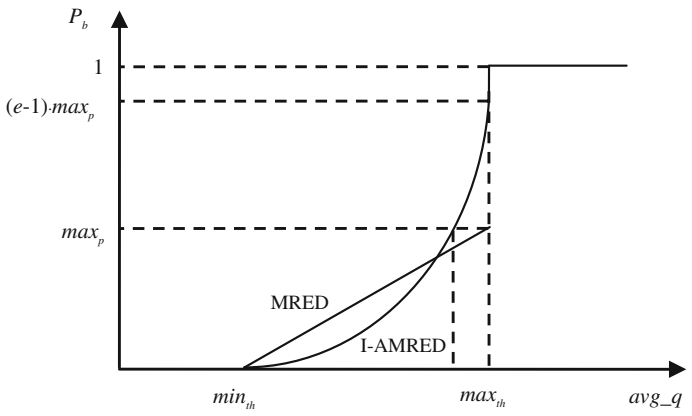


Fig. 42.1 Exponential relation of drop probability and average queue length in I-AMRED algorithm

been already less than max_{th}, it is not essential to drop arriving packets and *qlen* should be used to implement drop policy instead of drop all arriving packet. When *qlen* is more than max_{th}, queue buffer of router has been filled up, if *avg_q* is less than max_{th} and calculate P_b and let packets access, causing higher packet loss ratio, at this time, all arriving packets should be dropped. Because the three judgments above are added, congestion will be relieved earlier when *qlen* is more than max_{th} but *avg_q* is still less than max_{th}. Therefore, packet drop ratio of network will decrease totally.

42.3.2 Concrete Design of I-AMRED Algorithm

42.3.2.1 Calculation Equations

I-AMRED is an algorithm which can differentiates calculation from many BAs. In order to look convenient, variables of I-AMRED as below all aim at a certain BA.

$$avg\_q = (1 - w_q) \times avg\_q\prime + q \times w_q \tag{42.1}$$

$$P_b = max_p \times \exp\left(\frac{avg\_q - \min_{th}}{\max_{th} - \min_{th}}\right) - max_p \tag{42.2}$$

In Eq. (42.1), w_q is set as weight value of certain level drop precedence, *avg_q* is average queue length of last time and initial value of *avg_q* is zero. *q* is instantaneous queue length of sampling time. When *avg_q* ranges from min_{th} to max_{th}, P_b is calculated as in Eq. (42.2). The relation between P_b and *avg_q* in Eq. (42.2) is exponential as in Fig. 42.1.

42.3.2.2 Implementation in NS-2

NS-2 simulation platform contain DiffServ module, where MRED is programmed in the files of "dsredq.cc" and "dsredq.h" [6], so I-AMRED is implemented by modifying program code according to Eq. (42.2).

In the file"dsredq.h", add enumeration variable *status* and integer variable *qlen*, and declare function *updateIAMREDMaxP()*.

In the file"dsredq.cc", define the function *updateIAMREDMaxP()* in order to update max_p parameter periodically according to I-AMRED judgments. Modify the calculation formulae to exponential form and add judgments about *qlen* into the function *enque()*. Add the function *updateIAMREDMaxP()* into the function *calcAvg()* in order to update max_p parameter periodically for every BA.

42.4 Experiments on NS-2 Platform

42.4.1 Experiment 1

The network topology of experiment 1 is set as a simple network. Four source nodes send data to four destination nodes through a bottleneck link which is from a core router to an edge router. Three TCP agents are set on three source nodes which are marked with different drop precedence. An UDP agent is set on one source node, in which traffic is the sum of the three TCP sources. Links between edge routers and core router adopt TSW3CM. All nodes are configured as MPLS nodes.

Two schemes are set in experiment 1. First, let four source nodes send data in a high constant rate at the beginning of simulation time. Second, let two source nodes send data at the beginning of simulation time, and then at the 3rd second let other two source nodes send data. The purpose is to verify stability when traffic sharply increases.

42.4.2 Experiment 2

The network topology of experiment 2 is set as Fig. 42.2, and the purpose is to examine whether I-AMRED will be stable when the network scale is large. C is core router, E1, E2, …, E11 are source edge routers, E0 is destination edge router, S1, S2, …, S11 are source nodes, D is destination node. Every source node has a TCP agent marked by different drop precedence. These TCP agents use random number generator and adopt exponential distribution generating a number as a time interval of TCP transmission. Pareto model is used to generate a random number to assign a size of file needed to be transmitted. The arrival of packet obeys Poisson distribution. The link between C and E0 is bottleneck link. Accordingly, uncertainties of actual network distribution and data transmission can be simulated and stability of I-AMRED can be verified more practical.

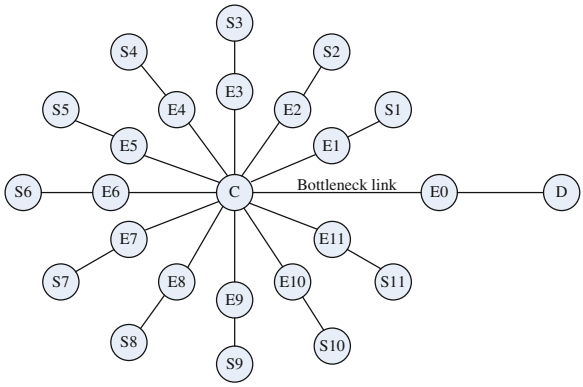


Fig. 42.2 Network topology of experiment 2

42.5 Results and Discussions

42.5.1 Experiment 1

When four source nodes send data in constant rate at the same time, set simulation time as 10 s, and throughput comparison figure between MRED and I-AMRED is shown in Fig. 42.3. So I-AMRED improves throughput on the whole, and is more stable than MRED at the beginning of simulation time. Packet loss ratio comparison between MRED and I-AMRED is shown in Table 42.1.

When letting two TCP sources send data at the beginning of simulation time, and then at the 3rd second let another TCP source and UDP source send data at the second scheme in experiment 1, throughput comparison figure between MRED and I-AMRED is shown in Fig. 42.4. So it is clear that when traffic is sharply increased, throughput using I-AMRED increases as simulation time goes on.

Packet loss ratio comparison between the two algorithms is as Table 42.2, so I-AMRED algorithm make packet loss ratio of TCP (AF PHB) reduce much, but UDP packet loss ratio raises, for drop precedence of UDP packets is high, and traffic of UDP is the sum of three TCP sources. As a whole, I-AMRED make average packet loss ratio decrease by over 10 %.

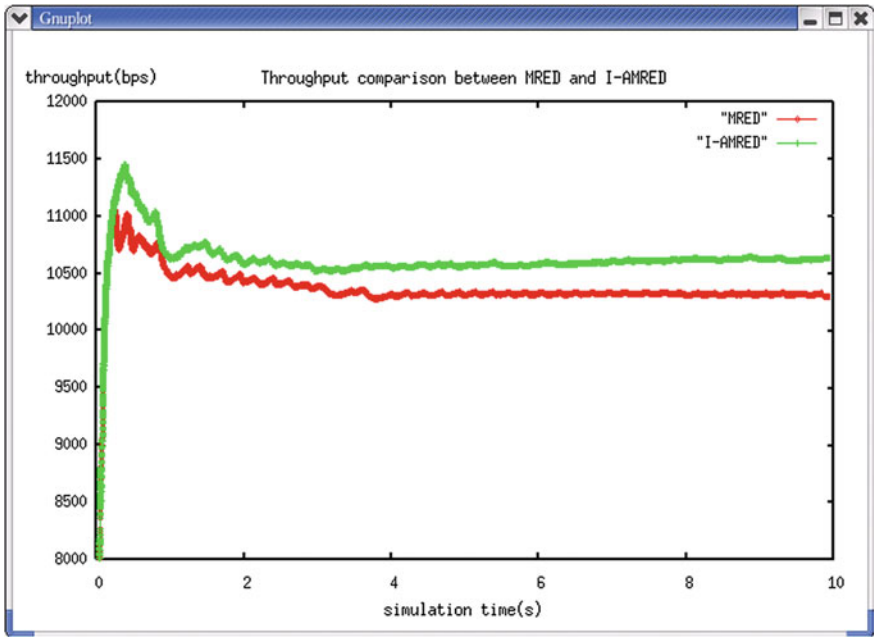


Fig. 42.3 Throughput comparison when sending rate is constant in experiment 1

Table 42.1 Drop packet ratio comparison when sending rate is constant

	TCP1 (%)	TCP2 (%)	TCP3 (%)	UDP (%)
MRED	5.699177	5.215420	6.076389	6.785214
I-AMRED	2.628697	3.158488	2.604699	8.526743

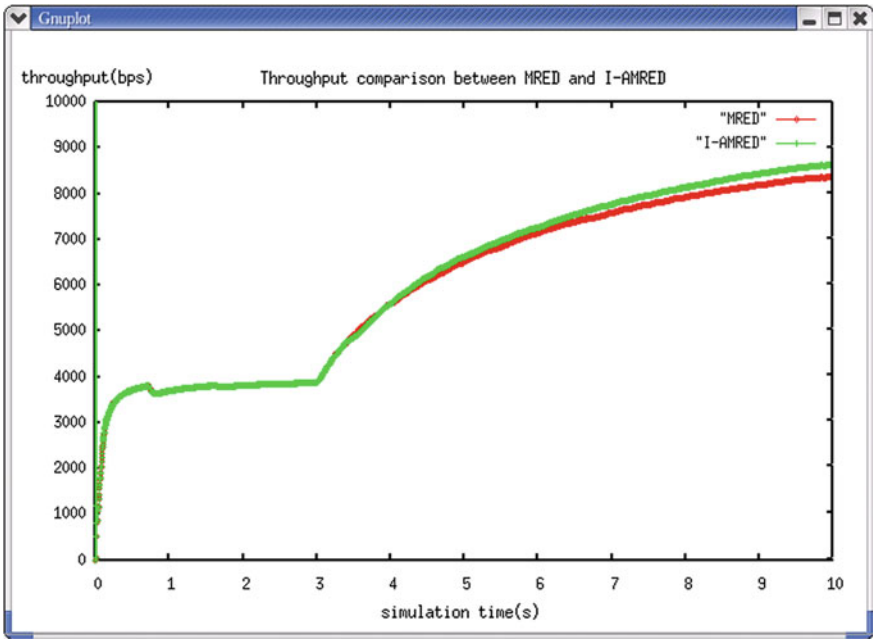


Fig. 42.4 Throughput comparison when traffic increases sharply in experiment 1

Table 42.2 Drop packet ratio comparison when traffic sharply increases at 3rd second

	TCP1 (%)	TCP2 (%)	TCP3 (%)	UDP (%)
MRED	6.156234	6.538661	5.343511	5.996960
I-AMRED	4.559118	4.552129	2.422407	10.309650

42.5.2 Experiment 2

In experiment 2, throughput comparison figure between MRED and I-AMRED is shown in Fig. 42.5, so I-AMRED make throughput even more stable and higher when network topology is more complex. Packet loss ratio comparison between MRED and I-AMRED is shown in Table 42.3.

Four TCP agents selected are examined. It is clear that packet loss ratios of four TCP sources all decrease after using I-AMRED, but the decrease of low drop

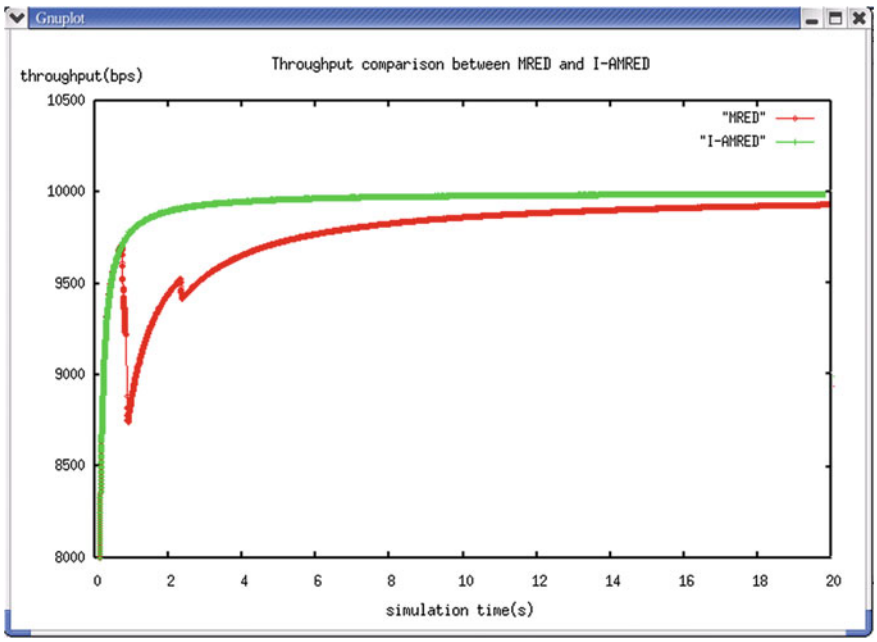


Fig. 42.5 Throughput comparison of MRED and I-AMRED in experiment 2

Table 42.3 Drop packet ratio comparison in experiment 2

	TCP1 (%)	TCP2 (%)	TCP3 (%)	TCP4 (%)
MRED	1.800000	2.824859	8.387097	9.409190
I-AMRED	0.908998	1.821934	6.895765	8.134652

precedence is more obvious, the decrease of high drop precedence is less. As a whole, average packet loss ratio decreases notably.

42.6 Conclusion

The results of experiment 1 and 2 indicate that I-AMRED raises throughput especially when traffic is large. And it sharply increases compared with MRED, and notably reduces packet loss ratio of AF PHB. Because of the improvement of throughput and packet loss ratio, sensitiveness to RED control parameter decreases, while stability of network throughput increases. These advantages make I-AMRED algorithm very suitable for current IP networks in which the amount and type of user and traffic expand rapidly. In summary, I-AMRED algorithm is ideally suitable for MPLS networks based on DiffServ.

Acknowledgments This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant Nos. N100204003.

References

1. Rahimi, M., Hashim, H., Rahman, R.A.: Implementation of quality of service (QoS) in multi protocol label switching (MPLS) networks. In: IEEE Conference Publications, pp. 98–103 (2009)
2. Qadeer, M.A., Sharma, V., Agarwal, A., Husain, S.S.: Differentiated services with multiple random early detection algorithm using ns2 simulator. In: Computer Science and Information Technology ICCSIT 2nd IEEE International Conference, pp. 144–148 (2009)
3. Makkar, R., Lambadaris, I., Salim, J.H., Seddigh, N., Nandy, B., Babiarz, J.: Empirical study of buffer management scheme for DiffServ assured forwarding PHB. In: IEEE Conference Publications, pp. 632–637 (2000)
4. Yang, W.: Research on Congestion Control Mechanism Based on Red algorithm. Nanjing Information Engineering University, Nanjing (2010)
5. Hollot, C.V., Misra, V., Towsley, D.: A control theoretic analysis of RED. Proc. IEEE INFOCOM **33**, 1013–1015 (2004)
6. Cao, Z.B.: Optimization and simulation of RED algorithm. Comput. Technol. Dev. **20**, 188–191 (2010)

Chapter 43
The Analysis and Implementation of Universal Workflow with Partition Algorithm on Finite Field

Wenxin Hu, Yaqian Yang and Guoyue Chen

Abstract To guarantee the running of system without collapse when deal with the complex changes of personnel and departments and implement the compatibility of the workflow system within different kinds of enterprises under certain regulations, Workflow with Partition Algorithm on Finite Field denoted in this paper provides an efficient solution. According to the performance of the eventually practice, this strategy completely solve the problem of addressing the issues of personnel position and department architecture changes in an enterprise as well as the requirement of enforcing the compatibility of the workflow system when deploying on different enterprises and also reduce the cost of system designing and new user importing time.

Keywords Workflow · Partition on finite field · Organizational hierarchy

43.1 Introduction

The technology of workflow starts from researches on office automation in mid-1970s [1]. Many related works have been reported in the research of self-adaptive workflow system. Two strategies of selection-adaption, through defining some phases of the workflow system during the running period instead of the creation time have been denoted [2, 3]. An interaction framework in which Manual intervening when the running system encounters some certain situations has been

W. Hu (✉) · Y. Yang
Computer Center, East China Normal University, Shanghai, China
e-mail: wxhu@cc.ecnu.edu.cn

G. Chen
Course of Electronics and Information Systems, Akita Prefectural University, Aza Ebinokuchi Tsuchiya, Yurihonjo 015-0055, Japan

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_43,

provided [4]. An architecture named Tri GSflow intends to use object oriented technology to integrate the object oriented model, role model and rule model in the workflow system [5]. Then resolution strategy and method of flexible workflow were discussed in particular from theoretical and implemental sides [6]. Some improvements on the dynamic modeling method have been made [7]. And a method was proposed to construct workflow including loop dynamically.

However, there're still some challenges within the state of the art technology. For example, it's hard to guarantee the running of system without collapse when deal with the complex changes of personnel and departments. For another, current researches of the workflow are mainly limited within the scope of s single enterprise, that means it can only deal with the flow changes took place in one certain company.

The algorithm denotes in this paper can greatly relieve the problems described above and can thus significantly improves the performance of the current workflow system. It provides an effective solution for both situations mentioned above by firstly extracting valid keywords according to organization information the enterprise possessed, and then partitioning these keywords into finite fields which is necessary for dealing with the problems made by personnel change and department restructuring.

The description of finite field partition algorithm is presented in Sect. 43.2. Section 43.3 provides the implementation and Sect. 43.4 presents a workflow implementation of the algorithm. Then Sect. 43.5 presents some concluding remarks.

43.2 Finite Field Partition Algorithm

The elements of this algorithm are described below:

(1) Pattern of organization information constitution: That means the regulations followed by denominating of each department, institution and organization in an enterprise.

Definition 1 Denote the number of enterprises with universal workflow service as m, the information constitution pattern of every enterprise as P_i, $0 < i \leq m$. So the universal workflow information constitution pattern is as follow.

$$P = \{P_{\mathrm{i}} | 0 < i \leq m\} \tag{43.1}$$

(2) Integrity Attribute of organization Information: the integrity designation of internal departments and organization according to constitution pattern.

Definition 2 Denote the integrity attribute of an enterprise as D_i, so the integrity attribute of universal workflow information is represented below.

$$D = \{D_i | 0 < i \leq m, P_i(D) \in P\} \tag{43.2}$$

(3) Hierarchy: The hierarchical relationship among the organizations according to compartmentalization from superior to inferior.

Definition 3 L_i describes hierarchy number of each enterprise which participates in workflow. The hierarchical relationship of universal workflow organization is as follow.

$$L = \{L_i | 0 < i \leq m\} \tag{43.3}$$

(4) Keyword: A group of distinguishing strings or string set for dividing information integrity attributes into separated string subset in reference to hierarchical relationship.

Definition 4 If the hierarchy number of an enterprise is L_i, the number of keyword subsets should be $L_i - 1$since $L_i - 1$subsets will separate the integrate information into L_i finite fields. K_i means all keywords of every enterprise for integrity attribute partition. The total of the number of keyword subsets in every hierarchy through the partition algorithm C equals $L_i - 1$.

$$K_i = \left\{ k | k \in K_i, \sum (C(k)) = L_i - 1 \right\} \tag{43.4}$$

The universal workflow keywords are as follow.

$$K = \{K_i | 0 < i \leq m\} \tag{43.5}$$

(5) Finite fields: describe the several fields from integrity attribute divided by keywords.

Definition 5 Assort the set of keyword from each enterprise ($L_i - 1$ in all) from high to low and match them with the integrity attribute of this enterprise. If a matched keyword (or an item of a keyword subset) is included, then partition this content as a finite field. Again, this attribute begins the next matching process with the next keyword after this partition, until all the keywords or keyword subsets in the set are processed. This procedure can be described as below:

```
1   counting variable I = 1
2   remaining properties after partition SD = Di
3   the set of partitioned properties G = Φ
4  While I <= Li-1 {
5        k = Ki (I)
6       IF SD ∩ k ≠ Φ
7       {
8          g = The content of SD partitioned by k
10         G = G ∪ g
11        SD = SD – G
12         I = I + 1
13      }
14   }
15   loop
```

According to the principle elements listed above, the hierarchy model is illustrated as follow (Shown in Table 43.1).

Following the parameter settings in the table above, the hierarchical partition of partition algorithm on finite field is demonstrated as follow. All of the internal departments in an enterprise are partitioned hierarchically. Hence, each node on the workflow can be defined by names of these departments and personnel (Fig. 43.1).

43.3 Implementation of Finite Field Partition Algorithm

43.3.1 Preparation of Basic Data for Workflow

(1) Draw keywords and number of ranks, which is also the maxim numbers of possible nodes in the workflow route, for partition from collection of sorted designations of all departments and organizations.

Table 43.1 An example of the principle elements

Elements	Examples
Constitution pattern of enterprise information	Including all complete designations from top to bottom
Integrity attributes of enterprise information	*D*: String 1 Keyword 1 String 2 Keyword 2 String 3 Keyword 3 String 4
Hierarchy	*L*: 4
Keyword	$K1 : \{\text{Keyword1}\}$,
	$K2 : \{\text{Keyword2}\}$,
	$K3 : \{\text{Keyword3}\}$
Finite fields generated after partition	$G1 : \{\text{String1 Keyword1}\}$,
	$G2 : \{\text{String2 Keyword2}\}$,
	$G3 : \{\text{String3 Keyword3}\}$,
	$G4 : \{\text{String4}\}$

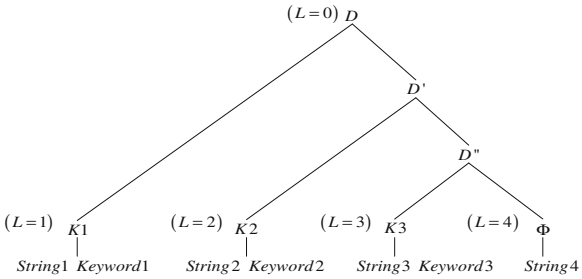


Fig. 43.1 The hierarchical partition by partition algorithm on finite field

(2) Configure keywords and the number of ranks.
(3) Divide the integral department designations into distinct rank components according to the keywords.
(4) Input all position designations which may possibly appear on workflow approval routings.
(5) Set respective position designations for every employee in the staff table.

According to the steps listed above (shown in Fig. 43.2), the basic data can be prepared for establishing workflow approval routings as well as for flexibly implementing different workflow paths. Special implementation is shown in Fig. 43.3.

43.3.2 Path Establishing in Universal Workflow System for Basic Workflow

A stabilized, regulated and limited workflow should be established in any enterprise, whatever its scale. Or, workflow would lose its significance. Thus, the subsequence step is to build up the potential workflow route in the intern enterprise. This includes four steps listed below:

(1) Set work positions of every approval node on each workflow route.
(2) Set hierarchical mark number of those positions.

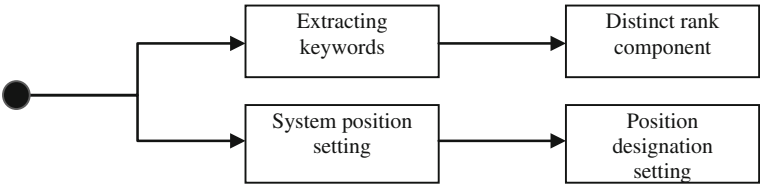


Fig. 43.2 The flow sheet of parallel processing of extracting keywords and system position setting

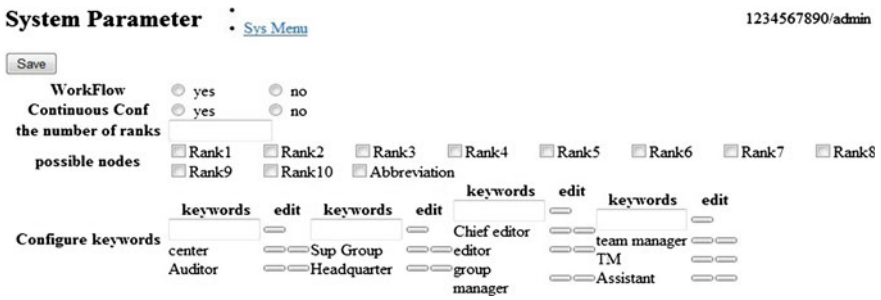


Fig. 43.3 Extraction of keywords for diving distinct rank components according to data analysis

(3) Iterate (1) and (2) until a workflow path is established.
(4) Iterate (1), (2) and (3) to create multiple paths for an enterprise.

Then, basic workflow routes can be established.

43.3.3 Workflow Setting for Special Field

In practical situation, besides the top-down approval structure in the regular enterprises, some special requirements are needed in certain ones. For instance, some enterprises require specialized quality control departments to participant in workflow approval. Otherwise, some need supervision departments activate throughout the process. Moreover, some allow employees in some certain departments that can join the procedure of approval, etc. Denote these spetial departments as special field and with those possible situations under consideration, we integrate those requirements in the universal workflow. The main steps of integration are listed below (shown in Fig. 43.4):

(1) When dealing with the scenario of department, such as quality control departments, special function departments and supervise department need to participate into the procedure of approval, we load the data of designations of those departments into the personnel position table of workflow system databases when setting the basic data. These special departments are named as special field.
(2) Set employees in relevant special departments. Normally, there are many in one.
(3) By analogy of work position, map those special departments to the nodes in the workflow.
(4) According to the different ways employees in those special departments participate into the workflow, divide the special field into collateral filed and exclusive field. All members in compatible field should join the tasks of the workflow node, the moment for participating in workflow. One and only one of those members in exclusive field should join the tasks of the workflow node.

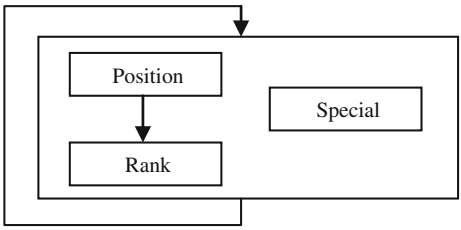


Fig. 43.4 Flow sheet of route setting of basic workflow and workflow containing special field

(5) According to the special filed nodes which participate into the action of workflow, Special field can also be divided into two different types: partial field and global field. Partial field has its unique working sequence, which means workflow members in partial field will join workflow tasks in special setting time while following the workflow order. Members of Global field in this field can join workflow tasks any time. In another word, no specific sequence is existent. In universal department, the approval timing of a member follows the time-order, like other members. The specific implementation is shown in Fig. 43.5.]

43.4 Workflow Implementation

After completing three main steps illustrated above, the basic workflow route should be established. Then, the final implementation is available. The critical point in this procedure is to allocate preset roles to single employee. Otherwise, department designations which partitioned by low, can bring us convenience

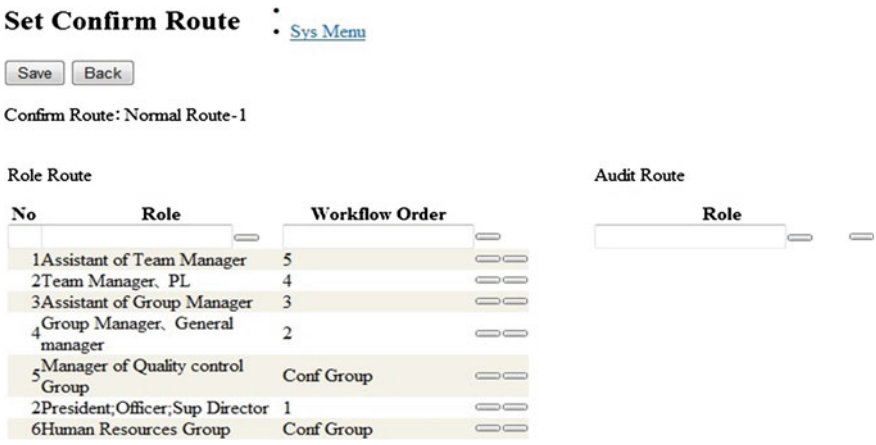
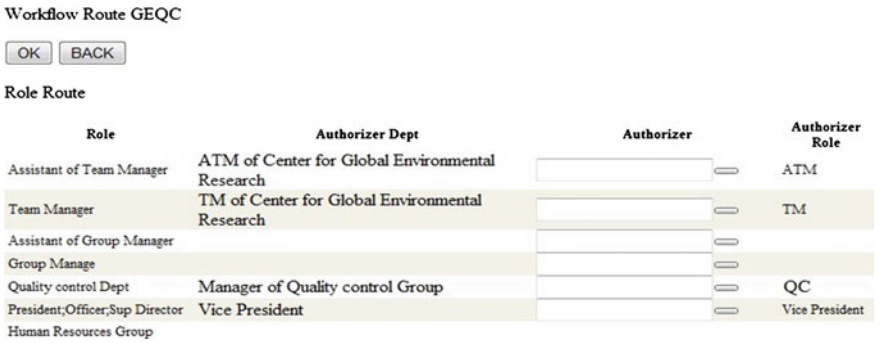


Fig. 43.5 Setting approval nodes of every rank on the route of workflow containing special field


Workflow Route GEQC

OK BACK

Role Route

Role	Authorizer Dept	Authorizer	Authorizer Role
Assistant of Team Manager	ATM of Center for Global Environmental Research		ATM
Team Manager	TM of Center for Global Environmental Research		TM
Assistant of Group Manager			
Group Manage			
Quality control Dept	Manager of Quality control Group		QC
President;Officer;Sup Director	Vice President		Vice President
Human Resources Group			

Fig. 43.6 Ascertaining an individual route of workflow

to easily confirm each roles on the approval route instead of searching rank relationships among complex staff tables. The main steps are listed below:

(1) Select a workflow route or automatically generate one depending on the different approval content.
(2) Gain all the finite fields of the applicant's department.
(3) Determine the next node which should be approved, and gain the position and the position rank number of the approval manager.
(4) Gain the desired finite fields of confirmer and its department finite fields by matching designations gained in step (2) with rank number.
(5) In this department, determine a approval manager in the members who own the position mentioned in step (3).
(6) Submit the approval request to the selected approval manager.
(7) Let the approval manager be the requirement submitter, and iterate steps from (2) to (6) until the whole workflow route is established.

The special implementation can be seen in Fig. 43.6. After selecting a certain route of workflow in the basis of user information, ascertain an individual route of workflow according to the relevant information of position and role of system settings.

43.5 Conclusion

Since it works on the enterprise information basis, the special-field workflow system based on partition algorithm on finite field can perfectly meet the requirements of organization structure changing and personnel switching. Moreover, because organization information is internal resources of an enterprise, there is no need to do a lot of system-oriented design. Therefore, it can significantly reduce the cost of system designing and new user importing time. Currently, almost all the workflow systems adopt the security mechanism based on the Task-

Role access control. However, with the system become more and more complex, other approaches are necessary to guarantee the security of the system. For example, the mixture pattern of Task-Role access control and self-adaptive mechanism can improve the system security in distributed environment and this will be our future work.

References

1. Shi, M., Yang, G., Xiang, Y., et al.: WFMS: workflow management system. Chin. J. Comput. **22**(3), 326–332 (1999)
2. Heinl, P., Horn, S., Jablonskis, S., et al.: A comprehensive approach to flexibility in workflow management systems. In: Proceedings of the International Joint Conference on Work Activities Coordination and Collaboration , pp. 79–88. AXM Press, New York (1999)
3. Horn, S., Jablonskis, S.: An approach to dynamic instance adaptation in workflow management applications. ACM Conference on Computer Supported Cooperative Work. ACM Press, Seattle (1988)
4. Jorgensenh, H.D.: Interaction as a framework for flexible workflow modeling. In: Proceedings of international ACM SIGGROUP Conference on Supporting Group Work, pp. 32–41. ACM Press, New York (2001)
5. Kappel, G., Rausch-Schott, S., Retschitzegger, W.: A framework for workflow management systems based on objects, rules and roles. ACM Comput. Surv. Sigplan Not. **35**(7), 32–40 (2000)
6. Zhou, J., Shi, M., Ye, X.: State of arts and trends on flexible workflow technology. Comput. Integr. Manuf. Syst. **11**(11), 1501–1507 (2005)
7. Cao, J.: Research and improvement on the dynamicity and adaptivity of flexible. Shang Hai Jiao Tong University, ShangHai (2009)

Chapter 44
Optimization for the Logistics Network of Electric Power Enterprise Based on a Mixed MCPSO and Simulated Annealing Algorithm

Lin Yuan, Dong Wang and Canquan Li

Abstract Recently the Electric Power Enterprise have many problems showing up. Combining the characters of Electric power enterprise supplies, researchers build an optimized network model to solve the irrationality of warehouse network, the simplex distribution and the high storage cost. This new model uses a method different from gravity and radius method which usually used by the general models. Researchers also discuss the penalty costs, construction costs and the operation costs in this paper. They use simulated annealing algorithm based on a mixed MCPSO (Multi-swarm Cooperative Particle Swarm Optimizer) in the new model. Fortunately, in an actual project of Power Supply Company, this optimized scheme is verified to be rational and effective.

Keywords Electric power enterprise · Logistics network · Mixed MCPSO and simulated annealing algorithm

44.1 Introduction

With the opening of electric power market, the power materials and goods management become one of the main factors to affect the core competence and cost, which faces a huge drastic market competition and challenge [1].

Although the power material management share some similarities to the other industries' material management, a higher demand of the selection and matching of goods is made by the characteristics of the electric power industry [2]. The storage is an important link of modern logistic management. The effective warehouse would simly integrating production, reducing cost and optimizing logistics network. The warehouse cost, which has direct influence on the cost of power

L. Yuan (✉) · D. Wang · C. Li
School of Software, Shanghai Jiao Tong University, Shanghai, China
e-mail: lynn.l.yuan@gmail.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_44,

production and construction, is related to the economy efficiency of electric power enterprise [3]. Therefore, this paper focus on how to optimize the logistics network. It provides reference and theoretical support to enhance resource integration capability, reduce cost and maintain the security of the power grid.

44.2 Logistics Network Optimization Model of Electric Power Enterprise

44.2.1 The Model of Logistics Network

It proposes a network topology of logistics network with three layers as shown in Fig. 44.1. The first layer stands a hypothesized central warehouse. On the second layer, a small amount of district warehouses are controlled by central warehouse. Those district warehouses delivery materials to the third layer which presents the turnover warehouses. The function of turnover warehouses is to support the operation of power companies.

Based on the network topology, an optimization model is established.

44.2.1.1 The Logistics Demand of District Warehouses and Turnover Warehouses

Because the material demand prediction of power industry cannot be accurately obtained by traditional demand prediction, the electricity consumption is combined [4]. The expression is as follow:

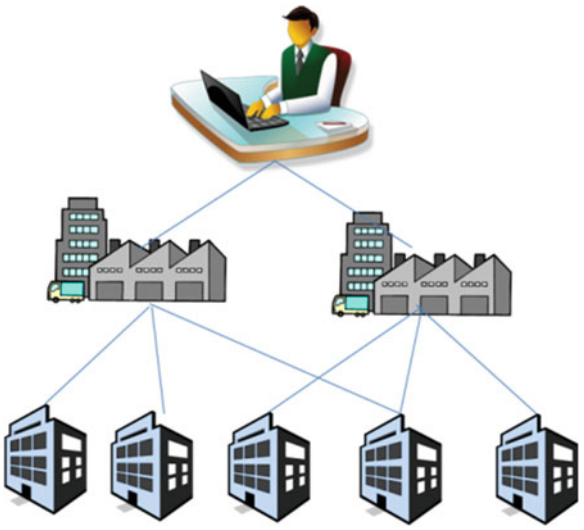


Fig. 44.1 A network topology of logistics network

$$R = Ec \times Q \tag{44.1}$$

In this function, Q means the amount of inventory used by the unit electricity consumption. R stands for the amount of annual inventory. Ec is annual electricity consumption.

We hypothesized that there are i kinds of materials. R_m represents the delivery quantity of district warehouse m. R_{mi}, as the delivery quantity of the ith kind materials from district warehouse m, transports the ith kind materials to several turnover warehouses. The ith kind of materials in a turnover warehouse can be transported by only one district warehouse. We can get the formula:

- The Sum of Storage Material

$$R = Ec \sum_{i=1}^{I} Q_i \tag{44.2}$$

- The Logistics Demand of District Warehouses

$$R_m = \sum_{i=1}^{I} \mu_{mi} R_{mi} \tag{44.3}$$

$$R_{mi} = \sum_{n=1}^{N} \omega_{mni} R_{ni}, \mu_{mi} = 1 \tag{44.4}$$

$$\sum_{m=1}^{M} \omega_{mni} = 1 \tag{44.5}$$

ω_{mni} 0–1 variable, the ith kind materials of Turnover Warehouse n is distributed from District Warehouse m or not;
R_{ni} delivery quantity of the ith kind materials from Turnover Warehouse n;

- The Logistics Demand of Turnover Warehouses

$$R_n = \sum_{i=1}^{I} R_{ni} \tag{44.6}$$

$$R_{ni} = \sum_{k=1}^{K} \sigma_{nk} R_{ki} + r_{ni} \tag{44.7}$$

$$\sum\nolimits_{n=1}^{N} \sigma_{nk} = 1, \forall k \tag{44.8}$$

K the sets of cancelled warehouses;
σ_{nk} 0–1 variable, if the materials of cancelled warehouse k is distributed from turnover warehouse n, $\sigma_{nk} = 1$;

R_{ki} delivery quantity of the ith kind materials from cancelled warehouse k;
r_{ni} output of the ith kind materials from turnover warehouse n;

44.2.1.2 The Reconstruction Cost of District Warehouses and Turnover Warehouses

According to the logistics demand, it is possible that warehouses need reconstruction. Eqs. (44.9)–(44.12) reflect the reconstruction cost.

- The Reconstruction Cost of District Warehouses

$$W = \sum_{m=1}^{M} \delta_m \left(A_1 S_{m1} + A_2 S_{m2} + A_3 \gamma_m \left(\frac{R_m}{TB} - S_{m1} - S_{m2} \right) + d_m \right) \tag{44.9}$$

$$\gamma_m = \begin{cases} 0, S_{m1} + S_{m2} \geq \frac{R_m}{TB} \\ 1, S_{m1} + S_{m2} < \frac{R_m}{TB} \end{cases} \tag{44.10}$$

- The Reconstruction Cost of Turnover Warehouses

$$U = \sum_{n=1}^{N} \delta_n \left(A_1 S_{n1} + A_2 S_{n2} + A_3 \gamma_n \left(\frac{R_n}{TB} - S_{n1} - S_{n2} \right) + d_n \right) \tag{44.11}$$

$$\gamma_n = \begin{cases} 0, S_{n1} + S_{n2} \geq \frac{R_n}{TB} \\ 1, S_{n1} + S_{n2} < \frac{R_n}{TB} \end{cases} \tag{44.12}$$

T	Turnover rate;
B	The average stock value per square meter;
A_1	Inside reconstruction cost;
A_2	Outside reconstruction cost;
A_3	The average of construction cost
S_{m1}, S_{n1}	The indoor storage area of Warehouse;
S_{m2}, S_{n2}	The outdoor storage area of Warehouse;
S_m, S_n	The area of Warehouse;
γ_m, γ_n	0–1 variable, Warehouse needs to be newly built or not;
d_m, d_n	The fixed cost of the construction of Warehouse;
δ_m	Warehouse m is district warehouse or not;
δ_n	Warehouse n is turnover warehouse or not.

44.2.1.3 The Operation Cost of District Warehouses and Turnover Warehouses

$$V = \sum_{m=1}^{M} \delta_m(m_1 + m_2)\left[S_{m1} + S_{m2} + \gamma_m\left(\frac{R_m}{TB} - S_{m1} - S_{m2}\right)\right] + \sum_{n=1}^{N} \delta_n(n_1 + n_2)\left[S_{n1} + S_{n2} + \gamma_n\left(\frac{R_n}{TB} - S_{n1} - S_{n2}\right)\right] \tag{44.13}$$

The operation cost consists of two parts: the labor cost and maintenance cost. In (44.13), means the labor cost per square meter, standing for the maintenance cost per square meter.

44.2.1.4 The Penalty Cost

In electric power company, they delivery materials from district warehouses to turnover warehouses by their own trucks. Combining the actual situation, we use penalty cost instead of traffic cost.

$$\mathrm{Y} = \sum_{\mathrm{m=1}}^{\mathrm{M}} \sum_{\mathrm{n=1}}^{\mathrm{N}} \sum_{\mathrm{i=1}}^{\mathrm{I}} \boldsymbol{\varphi}_{\mathrm{mn}}(\mathrm{d_{mn}} - \mathrm{D_1})\theta_1 \boldsymbol{\omega}_{\mathrm{mni}} \mathrm{R_{ni}} + \sum_{\mathrm{n=1}}^{\mathrm{N}} \sum_{\mathrm{k=1}}^{\mathrm{K}} \sum_{\mathrm{i=1}}^{\mathrm{I}} \boldsymbol{\sigma}_{\mathrm{nk}}(\mathrm{d_{nk}} - \mathrm{D_2})\theta_2 \mathrm{R_{ki}} \tag{44.14}$$

$\boldsymbol{\varphi}_{\mathrm{mn}}$ Turnover Warehouse n is covered by the service radius of District Warehouse m;
$\boldsymbol{\sigma}_{\mathrm{nk}}$ Cancelled Warehouse k is covered by the service radius of Turnover Warehouse;
$\mathrm{d_{nk}}$ road distance from Turnover Warehouse n to Cancelled Warehouse k;
$\mathrm{D_1}$ The service radius of District Warehouse to Turnover Warehouse;
$\mathrm{D_2}$ The service radius of Turnover Warehouse;
θ_1 Service penalty rate of District Warehouse;
θ_2 Service penalty rate of Turnover Warehouse.

44.2.1.5 The Total Storage Cost

Based on the above, the whole network model is as follow:

$$\textit{Minimize} \quad \mathrm{U} + \mathrm{W} + \mathrm{V} + \mathrm{Y} \tag{44.15}$$

$$\begin{cases} \sum_{m=1}^{M} \delta_m R_m = \sum_{n=1}^{N} \delta_n R_n \\ \sum_{m=1}^{M} \omega_{mni} = 1 \\ \sum_{n=1}^{N} \sigma_{nk} = 1, \forall k \\ S_m \geq \frac{R_m}{TB} \\ S_n \geq \frac{R_n}{TB} \end{cases} \tag{44.16}$$

44.3 A Mixed MCPSO and Simulated Annealing Algorithm

44.3.1 The Framework of Algorithm

According to Ref [5], a mixed MCPSO and simulated annealing algorithm is proposed in this paper as shown in Fig. 44.2.

44.3.2 Optimal Design Steps of Algorithm

- Initialize the Particle Swarm.
  S groups with N particles are generated. Annealing temperature is T.
- Operate PSO Algorithm by Every Subgroup.
  Before updating the status of subgroups, a main group M is selected by tournament selection to compare with the global best position of each subgroup. Every subgroup Q sends information of the personal best position to M. The positions of ith particle in d-dimensional space are expressed into$x_i = (x_{i1}, x_{i2}, \ldots, x_{id})$. The speed is represented in $v_i = (v_{i1}, v_{i2}, \ldots v_{id})$. Particle swarm is updated according to the velocity update rule of MCPSO.

$$v_{id}^{M} = v_{id}^{M} + c_1 r_1 \left(p_{id}^{M} - x_{id}^{M}\right) + c_2 r_2 \left(p_g^{M} - x_{id}^{M}\right) + \phi c_3 r_3 (p_g^{Q} - x_{id}^{M}) \tag{44.17}$$

$$x_i(t+1) = x_i(t) + v_i(t) \tag{44.18}$$

c_3 Learning factors;
r_3 Random numbers obeying the (0, 1) uniform distribution;
ϕ 0–1 variable, removal factor;

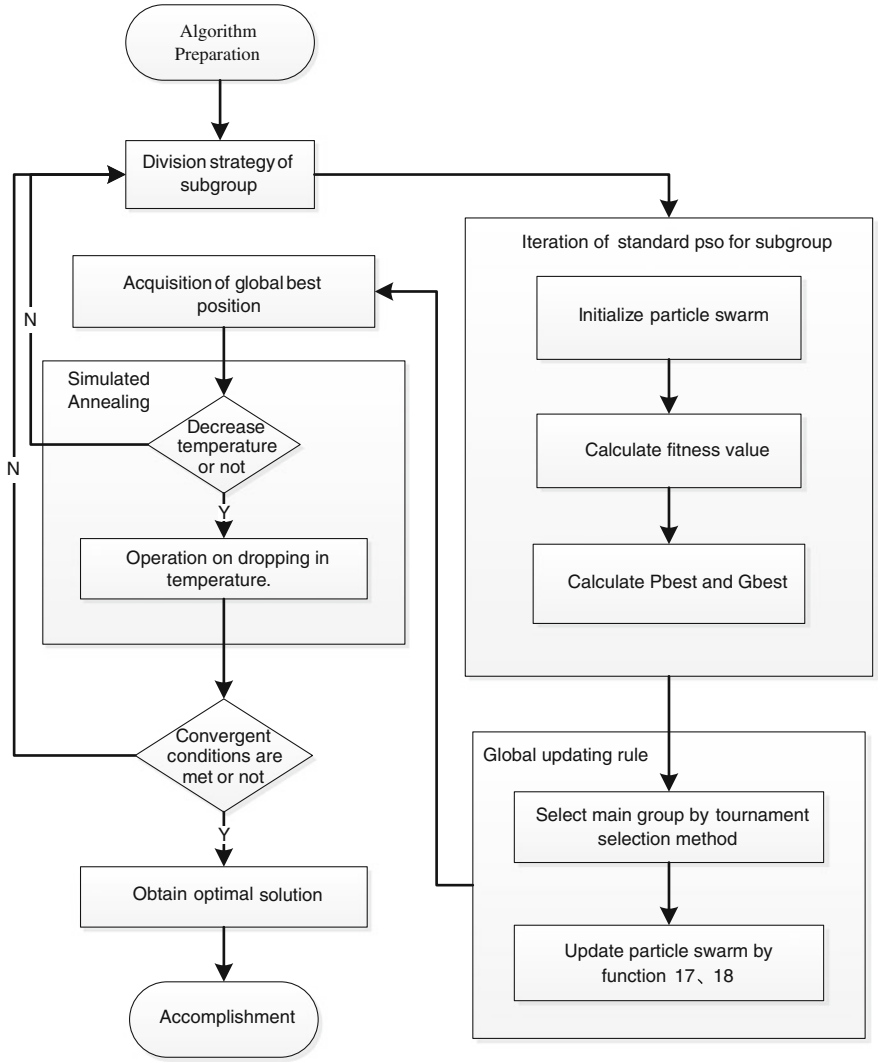


Fig. 44.2 The framework of algorithm

- Simulated Annealing
  SA is used to deal with each individual generated from the step 2. If the evolving times are less than the maximum times, return step 2.
- Accomplishment
  When a given maximum number of iteration has been performed, or the algorithm has satisfactory results, the algorithm is accomplishment.

Table 44.1 The relationship of district and turnover warehouse

District warehouse	Turnover warehouse
1	6,7,8
2	9,15,22
3	23,24
4	25,26
5	27,28

Table 44.2 The relationship of turnover and cancelled warehouse

The service radius of turnover warehouse	The service radius of cancelled warehouse
9	10,11,12,13,14
15	16,17,18,19,20,21

The service radius of No.9 warehouse covers the service radius of No.10–14
The service radius of No.15 warehouse covers the service radius of No.16–21

Table 44.3 Comparison diagram of optimized results

	District warehouse amount	Turnover warehouse amount	Cost
Before optimizing	–	28	250015167.3
After optimizing	5	12	228652211.3

44.4 Examples and Results Analysis

A power supply company somewhere in China has 28 warehouses. Compared with current situation, the results are presented. We build a network of warehouses in Table 44.1 (Tables 44.2, 44.3).

Obviously, the number of warehouse is reduced. The cost of transportation, storage and service is also reduce by 14 %.

44.5 Conclusion

Logistics network of electric power enterprise is a multiplicity and complexity network. Combining with the features of enterprise, a mathematical optimization model is put forward in this paper. In order to solve this model, a mixed MCPSO and simulated annealing algorithm is presented. In the end, the validity of the method is verified with an actual project.

References

1. Zhou, X.: The situation and development of the material management of electric power company. Business Research, No. 10, 19–20 May (2010)
2. Li, Z., Liu, J.: The situation of the material management of electric power company and the optimized scheme. China Storage Transp. Mag. **12**, 101–102 (2011)
3. Yang, Y.: Research on intensive management of power materials. North China Electric Power University, Beijing (2011)
4. Li, T.: Power grid materials warehouse scale forecast and location planning. North China Electric Power University, Beijing (2011)
5. Ben, N., Li, L., Chu, X.: Novel multi-swarm cooperative particle swarm optimization. Comput. Eng. Appl. **45**(3), 28–29 (2009)

Chapter 45
Bi-Level Programming Model and Taboo Search Algorithm in Industrial Location Under the Condition of Random Price

Yuan Qu and Zhong-ping Jiang

Abstract In this paper, the Bi-level programming model with random prize is presented to solve the industrial location problem. In terms of the binary coding, and controlling the location numbers on the operation of neighborhood, three kinds of neighbor domain are proposed by introducing the taboo search algorithm. In order to improve the taboo search algorithm for the efficiency and effectiveness of the addressed algorithm, the research utilizes a penalty function to deal with the constraints of the total amount of investment as well as introducing the principal of the model and TS. The representative example is used to illustrate the correctness of the optimal algorithm.

Keywords Industry location · Bi-level programming · Taboo search algorithm

45.1 Introduction

Looking back past researches on location problem, location-allocation problem (LA) becomes the most discussed topics in this area since it was proposed. LA mainly think about the relationship between facilities location and goods allocation, with the purpose to determine the optimal number, location and size of transit centers, hence to keep operating expenses and vehicles cost in lowest level. So far, many scholars committed to LA study and has made great remarkable efforts, some representative researchers are William [1], HodgBon [2], Keith [3], Hsieh [4] etc.

This article focuses on the industry location problem based in LA, not only the number and location, but also the output of each factory is the essential factors.

Y. Qu (✉) · Z. Jiang
School of Management, Jinan University, Guangzhou, China
e-mail: tquyuan@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_45,

This paper introduces the price as a random variable, then establish bi-level optimization model with the tabu search algorithm, and multiple sets of data will be tested; finally, a representative result and the analysis are presented.

45.2 Problem Description and Model

45.2.1 Problem Description

LA can be described as follows: leader factory will establish one or more factories whose total capacity is Q, with limitation of factory number, total investment, management, and technical force in M candidate locations; the effectiveness of the plant is influenced by its annual output, product, market prices, raw material prices and operating costs; the annual factory capacity is affected by amount of investment; the market price of product is a stochastic variable subject to a certain distribution; raw materials price is influenced by the whole layout; the goal is to maximize the difference between the sum of annual profit of all factory and the average annual amount of investment.

45.2.2 Mathematical Models

45.2.2.1 Model

In this paper, a bi-level optimization is used to describe the industry location problem. The upper level plan describes the optimal number, the location of industry and the total investment within the limits of decision-making; the lower level programming determines the capability of branch factory, aiming to maximize the total annual profit. The model is as follows[1]:

$$\begin{aligned} &\max_i (v - H \sum_{i \in I} C_i(y_i) x_i) \\ &\text{s.t.} \sum_{i \in I} x_i \leq \mathrm{N} \end{aligned} \tag{45.1}$$

$$\sum_{i \in I} C_i(y_i) x_i \leq \mathrm{I}_N \tag{45.2}$$

$$x_i = 0 \; or \; 1, \; i \in \mathrm{I} \tag{45.3}$$

[1] The full name of all s. t. in the paper are subject to.

$$\begin{aligned} V &= \max_{Y} \sum_{i \in I} P_p y_i - \sum_{i \in I} ((P_M(X) y_i + CE_i)) \\ \text{s.t.} &\sum_{i \in I} y_i \le Q \end{aligned} \tag{45.4}$$

$$y_i \ge Q_L \tag{45.5}$$

$$\sum_{i \in I} E(y_i) x_i \le E_M \tag{45.6}$$

$$y_i \ge 0 \tag{45.7}$$

where x_i means assigned 1 if the depot i is selected as the location factory, 0 opposite; X is upper decision variables $\{X_i\}$; y_i means production possibility of the factory; I is collection of the candidate location; $C_i(y_i)$ is construct costs when a branch was established at the depot i, is the function of y_i, $C_i(y_i) = a_1 y_i^{b1} + c_1$, a_1, b_1, c_1 are construction cost parameters; N is restriction number of industries; I_N is the restrictions total investment; v is total profits per year of lower level plant; H means unit coefficient which matches the upper plant construction costs and the lower branch's total profits.

The upper objective function is to maximize the difference between the annual profit and the average annual investment.

P_P is the product price, it is a random variable obedient to a certain distribution, gained from test of goodness of fit of market prices statistical analysis $\tilde{P}_P \sim N(L, R^2)$, $P_M(X)$ is the unit cost of the product, includes raw materials costs, human costs and logistics costs. Which is the function of X. Suppose S is the collection of the chosen location, then $P_M(X) = \frac{a_2(\sum_{j \in S, j \ne i} \frac{1}{b_{2j}}) + c_2}{\hat{u} S \hat{u}^{b_3}}$, a_2, b_2, b_3, c_2 is a normal number;

CE_i is constant which means other operating expenses of plant i; Q is total designed production possibility of the factory; Q_L means lower limit of branch's production possibility; E_M is human resource constraints, including management, technical force; $E(y_i)$ is a function of production possibilities, means the occupied human resources of the factory i, and $E(y_i) = a_3 y_i + c_3$, where a_3, c_3 are constant;

The lower objective function represents maximizing total operating profit, that is, to maximize the difference between the plant output value and the total cost;

Regulation (1): restrictions of branch number;
Regulation (2): restrictions on total investment;
Regulation (3): upper 0–1 decision variables;
Regulation (4): total plant designed production possibility is Q;
Regulation (5): lower limit Q_L of single branch's production possibility;
Regulation (6): total human resource constraints;
Regulation (7): factory production possibility is not less than 0.

45.2.2.2 Model Random Parameters Processing

As the above model expressed, the lower objective function contains a random variable $\hat{P}_P$ to process, we chose to follow processing methods: the objective function is transformed into constraint conditions, the probability of which is not less than the predetermined confidence level A [5] and [6]. Make $P_r\{v(Y) \geq \hat{v}\} \geq A$, then $\overline{v} = \max\{\overline{v}\hat{u}P_r\{v(Y) \geq \overline{v}\} \geq A\}$, that is, $\overline{v}$ is the maximum value of the objective functions when the confidence level is greater than or equal to A, so the lower objective function $v(Y)$ transform to:max $\overline{v}$

$s.t. P_r\{v(Y)\overline{v}\} \geq A$

Because $v(Y)$ is a linear function of P_P, and P_P is subject to normal distribution, so $v(Y)$ is also subject to normal distribution, and

$$\frac{\overline{v} - v(Y) - E(\overline{v} - v(Y))}{\overline{D(\overline{v} - v(Y))}} \sim N(0,1), D(\overline{v} - v(Y)) = (\sum_{i \in I} y_i)^2 D(P_p)$$

$$E(\overline{v} - v(Y)) = \overline{v} - (\sum_{i \in I} y_i E(P_P) - \sum_{i \in I} (P_M(X) y_i + CE_i)),$$

While $v(Y) \geq \overline{v}$ is equivalent to $\frac{\overline{v} - v(Y) - E(\overline{v} - v(Y))}{D(\overline{v} - v(Y))} \leq \frac{-E(\overline{v} - v(Y))}{D(\overline{v} - v(Y))}$, $P_r\{v(Y) \geq \overline{v}\} \geq A$ can be transformed into $P_r\left\{L \leq \frac{-E(\overline{v} - v(Y))}{D(\overline{v} - v(Y))}\right\} \geq A$, L is a stochastic variable, subject to a standard normal distribution. Make $X = 5^{-1}(A)$, 5^{-1}, is the inverse probability distribution function of standard normal distribution. From Ref. [7],

$$P_r\left\{L \leq \frac{-E(\overline{v} - v(Y))}{\overline{D(\overline{v} - v(Y))}}\right\} \geq A \Rightarrow \frac{-E(\overline{v} - v(Y))}{\overline{D(\overline{v} - v(Y))}} \geq X$$

After processing, the low-level programming can be described as:

$$\max \overline{v}$$

$$s.t. \overline{v} - (\sum_{i \in I} y_i E(\tilde{P}_P) - \sum_{i \in I} (P_M(X) y_i + CE_i)) + 5^{-1`}(A) \sum_{i \in I} y_i \overline{D(\tilde{P}_P)} \leq 0$$

Restrict (4) (5) and (6), let the upper planning unchanged, when lower-level programming is processed through the random parameters, industry location allocation model transforms to a deterministic bi-level programming problem, and the low-level programming is linear.

45.3 The Algorithm Design of Industry Location Allocation Problem

Because the location allocation problem is a NP-hard problem, the limit of the node number will have strict requirements if using exact algorithm. Consequently,

this research employs the tabu search algorithm to optimize the industry location allocation problem considering the characteristics of the model.

45.3.1 Solution Structure and the Initial Solution Generation

By using binary code, the coding length is the number of candidate industry locations, concerning the encoded depot i, 1 represents selected i as the industry location, and 0 means depot i is not selected. Make B = N/M, for any point i in the solutions, generating the random number D; if DγB, assigning 0 for depot i, or otherwise set to 1. Repeat the above steps m times, then get the initial solution.

45.3.2 Neighborhood Search

According to the characteristics of industrial location allocation, we designed three neighborhood operations: self-negated, 2-swap exchange and 2-opt exchanged.

1. Self-negated. Which is the operation of single point, that is doing self negated for the selected solution i.
2. 2-swap exchange. Which is two-point operation, that is exchanging two points of solutions while other points unchanged. For example, if selecting the points 2 and 6 in the solution to do 2-swap, solution 001101 will be turned into 011100.
3. 2-opt exchanged. Which is a Multi-point operation, that is exchanging two points, and reverses the value between two points. For example, if selecting the points 2 and 6 in the solution to do 2-opt exchanged, solution 001101 will be turned into a 010110.

Self negated, 2-swap exchange operations can be faster when searching a local optimal solution, while 2-opt exchange is able to expand the search space. Self-negated and 2-swap exchange operate were repeated in each cycle to search the local optimal solution. When the temporary optimal solution does not change, do 2-opt exchanged once to skip local optimal. The combinations of the above three neighborhoods will not only maintain the advantages of the TS local "mountain-climbing" ability, but also greatly enhance its global optimization ability.

45.3.3 Constraint Handling

The upper level constraints of the industry location allocation model include the number of branches and the amount of total investment limits. The number of factories is controlled by neighborhood operator: if the current solution violates the

branch number limitation after a neighborhood operation, the operation of this neighborhood will be canceled; for the total investment constraints, this paper employs the penalty function. If the value is beyond the constraints, then it should be multiplied by a penalty coefficient, which will be added into solutions for fitness as a penalty. The upper objective function F can be expressed as:

$$F = v - H\sum_{i\in I} C_i(y_i)x_i - p\tilde{o}(0, \sum_{i\in I} C_i(y_i)x_i - I_N)$$

where v is the total per year of lower level plant; H is unit coefficient which matches the upper plant construction costs and the lower branch's total profits; $C_i(y_i)$ is construct costs when a branch was established at the depot I; x_i means assigned 1 if the depot i is selected as the location factory, 0 opposite; I_N is the restrictions total investment; p is the penalty coefficient of the total amount of investment constraints.

45.3.4 Tabu List and Termination Criteria

According to the different natures of neighborhood operation in this tabu search algorithm, this paper designs two types of tabu list, namely the local tabu table and the global taboo list. The local tabu table stores itself negated-swap and 2-swap, exchange as the object of the neighborhood operation. That is, if a certain neighborhood is adopted to get the current solution, then in the next G-cycles it does not allow using the inverse operation of the neighborhood. Global taboo table stores the solution value for each cycle of the optimization process, and this taboo form can only be used in the selection of 2-opt exchange operation, so as to avoid the search repeat and accelerate the optimization speed.

If you get a better solution than the current best solution after a neighborhood operation, regardless of whether the neighborhood is taboo, the neighborhood should be regarded as the current neighborhood. Termination condition: If the optimal solution has not changed through the continuous K, the loop and the algorithm end.

45.3.5 Some Key Steps of Tabu Search Algorithm for Optimizing the Industry Location Problem with Pseudocode

```
Parameters. init();
For every factory
   f* = initSolution();
   if (the optimal value of continuous K generations did not change){
```

```
continue;//?
}
If (is Operated (f*.neighbor (2-opt))){
If (the optimal value of the continuous generation did not change){
nb solution values = searchSolution(f*.neighbor(2-opt));
} else {
nb_solution_values = searchSolution(!f*) + searchSolution(f*.neighbor
(2-swap));
}
}
nb solution values. orderByDesc ();
For each f in nb solution values
If (f > f*){
f* = f;
break;
}
If (f.violateRules() ==true){
contunue;
} else {
f = f*;
break;
}
End//the end of for each f in nb solution values
End//the end of for every factory
```

45.4 Numerical Results and Analysis

Calculate the multiple sets of examples by using TS in the industry Location Problem. Here are the results and their analysis of one representative numerical example.

45.4.1 Example Parameters and Results

Leader factory wants to establish one or more branch factories which the total capacity is 10 million units in the 10 candidate locations, and the limitation of total number of factory $N = 5$, the lower limit of single factory production capacity $Q_L = 100$ million units, factory construction cost parameters are: $a_1 = 1, b_1 = 0.7$, $c_1 = 10$, total investment restrictions $I_N = 220$; The parameters of raw materials cost are: $a_2 = 100, b_2 = 1.5, b_3 = 0.15, c_2 = 11$, human resource parameters are: $a_3 = 0.2, c_3 = 10, EM = 250$; matching coefficient $H = 5$, product price exhibits

Table 45.1 Candidate site coordinates

Serial number	1	2	3	4	5	6	7	8	9	10
X	22	32	45	52	61	61	50	3	28	5
Y	44	29	12	40	54	30	22	24	56	22

the normal distribution *N(15,4)*, confidence $A = 0.95$; candidate location coordinates generate randomly in *[0,100]* range, the results are shown in Table 45.1:

Based on the Matlab 7.0 to achieve the tabu search algorithm of the industry location problem, take the taboo algorithm parameters $K = 100$, $C = 30$, $G = 5$, calculate 20 times in pI2.0 model, converge to 100 % probability of this algorithm, we can get the optimal solution value of 992.98. The average convergence time is only 5.39 s. The optimal solution is found in the selection of 3,5and 8 factory locations. In these cases, the annual production capacity is 100,100 and 800 respectively.

45.4.2 Example Analysis

For the purpose of proving validity of the tabu search algorithm, the authors calculated the optimal solution of 3.1 by means of branch-and-bound method, which took 163.32 s. The value of the optimal solution is 992.98.

Analyze from computation results. With the problem that the tabu search algorithm this paper proposed, the results are stable and overall optimizing ability is strong; Analyze from computing speed. As shown in 3.1, its operation speed is 30.3 times the exact algorithm, it is practical and suitable for optimization of large-scale industry site problems.

45.5 Conclusion

With a description of industry location problem, this paper sets up a bi-level programming model for an industry location problem with the price changing randomly, and designs a taboo search algorithm for solving this model. The repeated calculations and results analysis demonstrates that the bi-level optimization with random parameters and optimal algorithm for industry location problem are effective.

References

1. William, G.T.: Timing market entry with a contribution-maximization approach to location-allocation decisions. Eur. J. Oper. Res. **4**(2), 95–106(1980)
2. HodgBon, J.: A flow-capturing location allocation model. Geogr. Anal. **22**(3), 270–279 (1990)
3. Willoughby, K.A., Uyeno, D.H.: Resolving splits in location/allocation modeling: a heuristic procedure for transit center decisions. Transport. Res. E. Logist. Transport. Rev. **37**(1), 71–83 (2001)
4. Hsieh, K.H., Tien, F.C.: Self-organizing feature maps for solving location-allocation problems with rectilinear distances. Comput. Oper. Res. **31**(7), 1017–1031 (2004)
5. Zhou, J., Liu, B.D.: New stochastic models for capacitated location-allocation. Comput. Eng. **45**(1), 111–125 (2003)
6. Louwers, D., Kip, B.J.: A facility location allocation model for reusing carpet materials. Comput. Ind. Eng. **36**(4), 855–869 (1999)
7. Niu, H-m: Class of traffic optimization model with stochastic parameters and genetic algorithm. J. Syst. Eng. **17**(2), 103–108 (2002)

Part III
Pattern Recognition

Chapter 46
Electrophysiological Correlates of Processing Visual Target Stimuli During a Visual Oddball Paradigm: An Event-Related Potential Study

Bin Wei, Bin Li and Yan Zhang

Abstract The aim of the study was to investigate temporal changes of brain to visual stimuli in earthquake-exposed survivors using event-related potentials (ERPs). The present study used ERP to explore whether trauma event and experience had much greater influence on earthquake-exposed survivors than controls, which could provide neuroscience evidences in the psychotherapy for earthquake-exposed survivors. After filling out a series of psychometric questionnaires, 13 earthquake-exposed middle school students and 13 healthy age and sex matched unexposed controls were investigated by using a visual oddball paradigm. One thousand visual stimuli trials were randomly presented with "target" stimuli occurring at a probability of 10 %, "novelty" earthquake-related stimuli 10 %, and "frequent" stimuli 80 %. Participants were asked to respond to "target" stimuli while ignoring other stimuli. Electroencephalogram (EEG) was continually recorded in order to assess P300 responses, an event-related potential (ERP) associated with attention processes. The results showed that the earthquake-exposed group had significantly reduced more P1 and P300 compared to the non earthquake-exposed control group. It indicated that the trauma event and experience had much greater influence on earthquake-exposed survivors than controls.

Keywords Earthquake-exposed · P300 · Oddball paradigm · ERPs

B. Wei · B. Li
Teaching Affairs Office, Mianyang Normal University, Mianyang, China

Y. Zhang (✉)
School of Educational Science, Mianyang Normal University, Mianyang, China
e-mail: zhangyan_psy@126.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_46,

46.1 Introduction

As we know, the brain is controlled to a large extent by chemical neurotransmitters, and it is also a bioelectric organ. The collective study of Event Related Potentials (ERPs) offers a window into brain physiology and function via computer and statistical analyses of traditional EEG patterns. It suggests innovative approaches to the improvement of attention, emotion and behavior, via high temporal resolution ERP technique. The focus of this study is on attention processing toward target stimuli in survivors who experienced trauma without PTSD symptoms. Thus, we hypothesized that the trauma event might have greater significantly influence on cognitive and brain functions in the earthquake-exposed group than the control group. It is yet unclear the extent to which modality of stimulus presentation is significant with respect to the activation of brain systems involved in the earthquake-exposed group, particularly to non-threatening neutral stimuli. In majority of studies of PTSD, it is now known that attention deficits are present to non-threatening neutral stimuli that are presented primarily in the auditory modality [1]. Several studies used visually presented trauma related stimuli [2–4]. Kimble, Kaloupek, Kaufman, and Deldin [5] found that attentional problems in PTSD may be activated either by heightened arousal to threat stimuli or by the requirement of sustained attention. As Miltner, et al. [6] showed higher P300 amplitudes to fearful stimuli than to neutral stimuli in spider phobics but not in healthy controls. A meta-analytic review of ERP studies in PTSD by Karl, Malta, and Maercker [7] revealed that PTSD patients showed greater P300 amplitudes to trauma-related cues compared to non-PTSD trauma controls. Also, Wessa, Jatzko, and Flor [8] found that PTSD subjects showed higher P300 to trauma-related materials than healthy controls. However, there are few studies that used only neutral visual stimuli in the earthquake-exposed group without PTSD.

Accordingly, we predicted that target stimuli would elicit attenuated amplitudes in the earthquake-exposed group compared to the control group, and this ERP effect would be related to a posterior greater positivity. Here, P1 and P300 amplitude were examined to determine whether the earthquake-exposed group had reduced amplitude to the target stimuli compared to controls in an oddball paradigm.

46.2 Method

46.2.1 Participants

The volunteer participants recruited from two middle schools, who were interested in our research. Firstly, participants were consented to participant our experiment by their guardians (their parents and teachers signed the informed consent). Secondly, participants were initially screened for selection criteria through an

interview and some psychometric assessments. Accordingly, 13 middle school student (7 girls, 6 boys) participants who had experienced the Wenchuan earthquake selected as the earthquake group, while 13 healthy age and sex matched participants (7 girls, 6 boys) that had not experienced the Wenchuan earthquake selected as the control group. The mean age of the earthquake group was 15.78 (±0.58, range 15–16), of the controls 15.45 (±0.26, range 15–16). All participants were right-handed and had normal hearing and no self-reported history of neurological or psychiatric disorder.

46.2.2 Procedures

Firstly, all subjects signed an informed consent in accordance with the Ethical Principles of Psychologists, and the study was approved by Mianyang Normal University ethics committee. The procedure consisted of two parts. In the first part, some psychometric assessments were carried out. In the second part, the EEG was recorded in an oddball paradigm. Finally, the psychological counseling about trauma was provided to all of the participants by our psychotherapists.

Psychometric Assessments We collected some demographics through a face to face interview [9]. There were 18 students from Mianyang, which is one of the three major cities immediately surrounding the earthquake's epicenter (Wenchuan, approximately 75 miles). Moreover, they reported had no loss of life of significant others, no physical injure, and no psychosomatic symptoms, but felt strong shaking in the earthquake, witnessed specific horrific events in the disaster area, and viewed earthquake-related television. Besides, there were 19 students from Chongqing located 186 miles from the epicenter (Wenchuan) of the 2008 Chinese earthquake [9]. However, only 16 students reported not had previous or current presence of traumatic experiences. Then, all of the participants selected by the interview were investigated by a post-traumatic stress disorder self-rating scale (PTSD-SS) [10]. The PTSD-SS was constructed based on the definition and diagnostic criteria of the post-traumatic stress disorder in the Diagnostic and Statistical Manual of Mental Disorders: 4th Edition (DSM-IV). It provides that individuals who have got the total score below 60 have no serious PTSD symptom [10]. According to its scores, 17 subjects from Mianyang and 15 subjects from Chongqing who have got the total score below 60 were selected as the earthquake group. The participants also filled out a modified version of the PTSD Cheeklist-CivilianVersion (PCL-C) [11]. This allowed for assessment of criteria A1 and A2 to assess the presence of a traumatic event as defined by current American Psychological Association (APA) criteria. Moreover, the total score within 17–37 was identified no obvious PTSD symptom. Accordingly, 13 subjects from Mianyang and 13 subjects from Chongqing were respectively selected as the earthquake group and control group. Furthermore, 26 subjects did not meet the A1 and A2 criteria for trauma in the DSM-IV. This paperwork was followed by the ERP task. All subjects were paid for their participation after they were investigated.

Experiment Tasks The participants were seated in a comfortable chair in a soundproof, electrically shielded, dimly lit room, at a distance of 80 cm from the computer screen. Initially, each participant performed 10 training trials. Then, they completed the oddball task. Altogether, 1000 trials divided into 5 blocks were presented for 1500 ms each with an inter stimulus interval (ISI) of 800 ± 50 ms. For the entire task, 80 % of the stimuli were repeating bell pictures ("frequents"), 10 % of the stimuli were rare repeating stool pictures ("target"), and the other 10 % of the stimuli were unique, non-repeating earthquake-related pictures ("novelty"). The earthquake-related pictures were from the database created by Zhang et al. [12], including the field, people, buildings, and environment related to earthquake. The subjects were asked to ignore all other pictures, and press "J" when they saw the "target" pictures. The stimuli were presented in randomized order. The response-to-hand assignment was counterbalanced across participants. Accuracy rates for all participants were over 97 %. The duration of the experiment was about 30 min.

46.2.3 EEG Recording and Data Processing

The electroencephalogram (EEG) was recorded from 64 scalp sites using tin electrodes mounted in an elastic cap (Brain Products, Munich, Germany), with the references on the left and right mastoids (average mastoid Ref. [13]. The ground electrode was on the medial frontal aspect. The vertical electrooculogram (EOG) was recorded with electrodes placed above and below the left eye. The horizontal EOG was recorded from the left versus right orbital rim. EEG and EOG were amplified using a 0.01–100 Hz bandpass and continuously sampled at 500 Hz/channel. Impedances were kept below 5 KΩ. Averaging of ERPs was computed off-line with rejecting those trials with eye movements, blinks, motion, or other artifacts at any of the channels. Trials contaminated with EEG artifacts (mean voltage exceeding ±80 μV) or those with artifacts due to amplifier clipping, bursts of electromyographic (EMG) activity, or peak-to-peak deflection exceeding ±80 μV were excluded from averaging.

Data were analyzed offline using BP software (Analyzer 1.0). For the ERP analysis, epochs of 1000 ms were generated offline from the continuous ERP records, starting 200 ms before target stimulus onset. Furthermore, incorrect behavioral classifications had to be rejected from further data processing. ERPs were aligned to a 200 ms baseline. Average ERP waveforms were calculated as described above in the test phase. There are more than 90 valid epochs for each individual ERP average.

46.2.4 Statistical Analysis

Consistent with previous research of Polich [14], the most obvious P300 effect to targets was parietal. In this study, the largest P300 to targets was maximal at Cz, Pz and Oz. Therefore, the amplitude and latency of the P300 to target stimuli were measured at frontal/central/parietal/occipital from 7 scalp positions including Fz, FCz, Cz, CPz, Pz, POz, and Oz.

Statistical analyses were performed using SPSS for Windows (version 11.0; SPSS Inc., 2002). For all analyses, *P* value was corrected for deviation from sphericity according to the Greenhouse–Geisser method. The results section only reported main effects and interactions that involved the group factor, based on the main hypothesis of the study.

46.3 Results

Base on the visual inspection of grand average amplitude (see Fig. 46.1) and on previous studies in the literature on oddball paradigms, ERP analyses were conducted on P1 (50–150 ms), N2 (150–250 ms), and P300 (350–400 ms) components. Accordingly, repeated measures analyses of variance (ANOVAs) on Group (Earthquake-exposed group/Control group) × Electrodes (Fz/FCz/Cz/CPz/Pz/POz/Oz) was conducted to determine if the latencies and amplitude differences were present.

P1 As shown in Fig. 46.1, the P1 was elicited by target stimuli. The ANOVA showed that there were no P1 latency main effect on Group, $F(1, 24) = 3.168$, $P = 0.291$, on Electrodes (POz/Oz), $F(1.00, 24.00) = 0.036$, $P = 0.852$, and no Group × Electrodes interaction, $F(1.00, 24.00) = 0.812$, $P = 0.376$.

The ANOVA revealed that there was main Group effect of P1, $F(1, 24) = 4.695$, $P = 0.040$, with the earthquake-exposed group (7.12 ± 1.35 μV) yielding reduced positivity compared to the control group (11.24 ± 1.35 μV). But the main effect of Electrodes, $F(1.00, 24.00) = 2.038$, $P = 0.166$, and Group × Electrodes interaction difference is not present, $F(1.00, 24.00) = 0.284$, $P = 0.599$.

P300 As shown in Fig. 46.1, the P300 was elicited by target stimuli. The ANOVA showed a main latency effect of P300 on Electrodes (Cz/CPz/Pz/POz/Oz), $F(2.16, 51.92) = 3.771$, $P = 0.027$. The results indicated that the latency on POz (359.23 ± 5.38 ms) and Oz (357.92 ± 6.20 ms) were earlier than that on Cz (381.31 ± 7.72 ms) and CPz (383.38 ± 7.80 ms), while Pz (365.46 ± 6.20 ms) was later than POz, and earlier than CPz. However, the Group × Electrodes interaction and Group effect were not significant, $F(2.16, 51.92) = 0.482$, $P = 0.635$, $F(1, 24) = 0.619$, $P = 0.439$.

The ANOVA showed a main amplitude effect of P300 on Electrodes (Cz/CPz/Pz/POz/Oz), $F(1.68, 40.31) = 5.190$, $P = 0.014$. The results showed that the

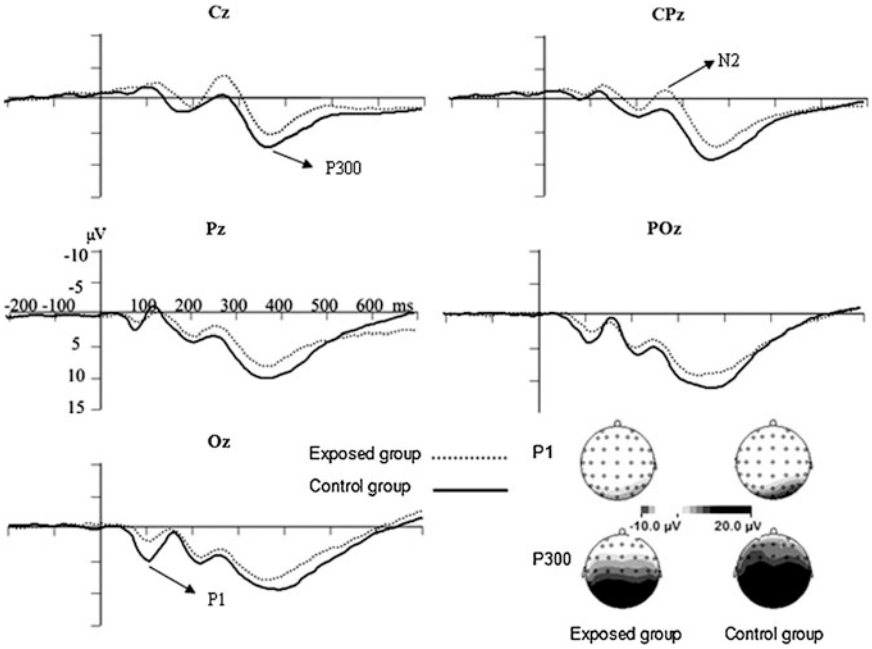


Fig. 46.1 Grand-mean ERP and topographies of differences curves at representative electrode sites to target stimuli in the earthquake-exposed group versus the control group

amplitude on Cz (12.77 ± 1.79 μV) and Oz (12.86 ± 1.70 μV) were smaller than that on Pz (18.54 ± 1.40 μV) and POz (18.39 ± 1.64 μV), while CPz (16.62 ± 1.46 μV) was larger than Cz. It indicated that the most enhanced P300 amplitude effects were on the parietal scalp. There was also a significant Group effect, $F\ (1,\ 24) = 5.539$, $P = 0.027$, with the earthquake-exposed group (13.12 ± 1.63 μV) producing reduced positivity compared with the control group (18.56 ± 1.63 μV).

46.4 Discussion

The present study examined electrophysiological measures of an earthquake-exposed group compared with a non earthquake-exposed control group to target stimuli during an oddball paradigm. It is generally accepted that the positive component emerged less than 150 ms after stimulus onset representing rapid processing in human visual system [15]. It suggested that the earthquake-exposed group allocated less attention resource to target stimuli than the control group. Moreover, P300 reflects more accentuated to high arousing or salient stimuli in the context of a given task [16]. In particular, the P300 component has been examined in studies of PTSD because it provides temporal information about attention

processes such as the ability to detect a target, the salience of the target to the participant, and the response to novel or distracter stimuli. For these reasons, the P300 is the most widely studied ERP in individuals with trauma histories. The findings suggested that PTSD may not be central to the attention disturbances found in traumatized samples, while trauma history may play a more important role.

Generally, oddball neutral targets, but not standard or novelty stimuli, elicit a subcomponent of the P300 called the P3b [7]. As for the findings in this study, recent brain imaging studies have shed light on the neural underpinnings of attention processing to target stimuli. Reduced P3b responses have been interpreted as an index of general cognitive impairment as well as more specific deficits in attention, working memory, and in the allocation of information processing resources to tasks [14]. As some researchers have presumed that abnormal function in these brain networks helps to explain deficits or everyday attention difficulties in PTSD [17], we could deduce that these brain functional difficulties or attention deficit may result in the ERP amplitude differences between the earthquake-exposed group and the non earthquake-exposed control group.

According to these studies, the decreased P1 and P300 may reflect more attenuated neural activity to target stimuli in the earthquake-exposed group than controls. Consequently, the findings are consistent both with the clinical presentation of the disorder and with theoretical notions that traumatized group demonstrates attention deficits towards virtual or auditory stimuli. Similarly, there were same cerebral processing mechanisms of attention to target stimuli with the traumatized PTSD group to the earthquake-exposed group without PTSD. As some studied have also demonstrated that attention tasks using neutral cognitive stimuli showed differences between PTSD and control participants [7], this study provides further evidence that the attention impairments in the earthquake-exposed group without PTSD are not confined to trauma-related stimuli.

46.5 Conclusion

The present study showed that the earthquake-exposed group exhibited reduced P1 and P300 amplitudes to target stimuli on posterior scalp compared with the control group during a visual oddball paradigm. It indicated that the trauma event and experience had much greater influence on earthquake-exposed survivors than controls.

Acknowledgments This project was supported by "Youth Project from Ministry of Education of Humanities and Social Sciences, China (11YJC190036)" to Yan Zhang.

References

1. Shucard, J.L., McCabe, D.C., Szymanski, H.: An event-related potential study of attention deficits in posttraumatic stress disorder during auditory and visual Go/NoGo continuous performance tasks [J]. Biol. Psychol. **79**(2), 223–233 (2008)
2. Attias, J., Bleich, A., Furman, V., Zinger, Y.: Event-related potentials in post-traumatic stress disorder of combat origin [J]. Biol. Psychiatry **40**(5), 373–381 (1996)
3. Bleich, A.V., Attias, J., Furman, V.: Effect of repeated visual traumatic stimuli on the event related P3 brain potential in post-traumatic stress disorder [J]. Int. J. Neurosci. **85**(1–2), 45–55 (1996)
4. Stanford, M.S., et al.: Houston, Impact of threat relevance on P3 event-related potentials in combat-related post-traumatic stress disorder [J]. Psychiatry Res. **102**(2), 125–137 (2001)
5. Kimble, M., Kaloupek, D., Kaufman, M., Deldin, P.: Stimulus novelty differentially affects attentional allocation in PTSD [J]. Biol. Psychiatry **47**(10), 880–890 (2000)
6. Miltner, W.H., et al.: Event-related brain potentials and affective responses to threat in spider/snake-phobic and non-phobic subjects [J]. Int. J. Psychophysiol. **57**(1), 43–52 (2005)
7. Karl, A., Malta, L.S., Maercker, A.: Meta-analytic review of event-related potential studies in post-traumatic stress disorder [J]. Biol. Psychol. **71**(2), 123–147 (2006)
8. Wessa, M., Jatzko, A., Flor, H.: Retrieval and emotional processing of traumatic memories in posttraumatic stress disorder: peripheral and central correlates [J]. Neuropsychologia **44**(10), 1683–1696 (2006)
9. Zhang, Y., et al.: Mental health and coping styles of children and adolescent survivors one year after the 2008 Chinese earthquake [J]. Child Youth Serv. Rev. **32**(10), 1403–1409 (2010)
10. Liu, X.C., et al.: Development of post-traumatic stress disorder self-rating scale and its reliability and validity [J]. Chinese J. Behav. Med. Sci. **7**(2), 93–96 (1998)
11. Yang, X.Y.: The research in formation and attribution of PTSD and intervention experiment for medical students. Ph.D. dissertation [D]. Liaoning Normal University, pp. 14–15 (2007)
12. Zhang, Y., et al.: A comparative study of earthquake-exposed middle school and undergraduate students on memory bias to threatening stimuli [J]. Psychol. Dev. Educ. **28**(2), 52–60 (2012)
13. Luck, S.J.: An introduction to event-related potentials and their neural origins. In: Luch, S. (ed.) An Introduction to the Event-Related Potential Technique [M], p. 107. MIT Press, Cambridge (2005)
14. Polich, J.: P300 clinical utility and control of variability [J]. J. Clin. Neurophysiol. **15**(1), 14–33 (1998)
15. Thorpe, S., Fize, D.: C. Marlot. Speed of processing in the human visual system [J]. Nature **381**(2), 520–522 (1996)
16. Johnson, R.Jr.: On the neural generators of the P300 component of the event-related potential [J]. Psychophysiology **30**(1), 90–97 (1993)
17. Weber, D.L., et al.: Abnormal frontal and parietal activity during working memory updating in post-traumatic stress disorder [J]. Psychiatry Res. **140**(1), 27–44 (2005)

Chapter 47
Realization of Equipment Simulation Training System Based on Virtual Reality Technology

Pin Duan, Lei Pang, Qi Guo, Yong Jin and Zhi-Xin Jia

Abstract According to the actual needs of equipment training, function and structure of simulation system present themselves by comparing past training systems. The ideas are elaborated which include solid modeling, 3D terrain rendering, fault databases generation and control rules setting. The system was realized based on 3D interactive devices and programming language. By using forces, the system meets the design requirements and the training needs.

Keywords Virtual reality · Simulation training · 3D terrain · Troubleshooting

47.1 Introduction

As the modern weapon updating from time to time, structure and function of the weapon are becoming more and more complicated. Therefore, there is a high request to efficient training. If the simulation training system can be developed before the weapon inputted to army, the training time will be decreased and the human resources will also be saved. By applying the virtual reality technology in training, trainees are able to get familiar with the operation platform and master the operation method in a short period which saves training time and expenses, improves the training efficiency and safety and avoid the limitation of the site as well [3]. Moreover, previous training systems are lack of the fault maintenance. So, the simulation system based on virtual reality technology is proposed.

P. Duan (✉) · L. Pang · Q. Guo · Y. Jin · Z.-X. Jia
General Armament Department of PLA, Wuhan Military Representative Office, Wuhan, China
e-mail: dpindpindpin@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_47,

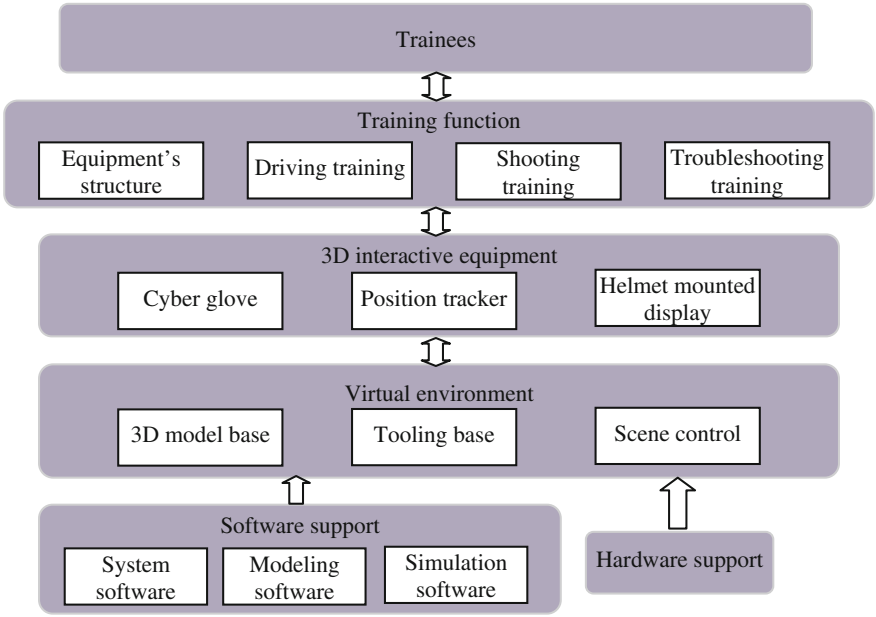


Fig. 47.1 The vertical type structure design chart of training

47.2 Needs and Functions

The equipment simulation training system aims at producing a living training environment and achieving training content, including equipment's structure study, simulation training, assessment and third part development. Simulation training includes driving training, shooting training and troubleshooting training. By the third part development interface, users can import other weapon system models, other 3D terrains and new battle plans based on help document [10]. Figure 47.1 shows the vertical type structure design of training.

47.3 Ideas

47.3.1 Solid Modeling

Besides the function and efficiency of training simulation system, the most important element is modeling for the trainees. Therefore, it is necessary to create a model which is attractive and easy to operate, which is one of significant factors in the design process [4].

Fig. 47.2 Equipment external structure

The paper models equipment by the software of 3D MAX, but it is necessary to generate the files as *.flt in operation process by the Vega software. Therefore, the 3D MAX software and the Creator software are united in the system [5]. Firstly, equipment models are produced in the 3D MAX software, and then imported into the Creator software as the *.dxf files. Finally, to imported into the Vega software and be able to function, the models are modified and simplified, and generated to the *.flt files. The models produced in this way are fine. Similarly, complex models of equipment showed in Fig. 47.2 can also be produced by this method.

47.3.2 3D Terrain Rendering

The method of terrain rendering is simply but time-consuming. Firstly, scan the contour maps into electronic topographic maps. Then workload is biggest to build paths of contour lines on the electronic topographic maps by the Photoshop software. It should be noticed to minimize the file size that paths are constructed in sparse areas of contour lines by intervals of 10 m while 50 m in dense areas can be increased. Next, fill the colors in the areas divided by paths from dark to light following the order of contour lines, and save the files as half-tone pictures. Finally, import it into modeling software, and generate the 3D terrain with slight adjustment [2]. Figure 47.3 demonstrates the finished 3D terrain.

47.3.3 Fault Databases Generation

Fault databases establishment is difficult for troubleshooting of virtual training system which includes expert solutions and animated simulations for troubles [8]. Based on equipment maintenance manuals and Feedback information from users, fault databases are finished including of Artillery types, parts names and No., fault

Fig. 47.3 3D terrain of a partial shooting range

codes, fault phenomenon, animated descriptions and solutions, as shown in Table 47.1.

According to fault codes from databases, virtual system shows the fault phenomenon in the maintaining process, such as breech block not in place and button malfunction. Trainees go to do other tasks after to solve the fault based on its phenomenon. In this way, trainees can be improved the ability to react and the capacity of dealing with sudden faults.

47.3.4 Control Rules Setting

Control rules are necessary no matter whether it concerns driving, shooting or troubleshooting. Control rules are the assembly of methods to describe the experience knowledge, which is the foundation of next step deduction. The rule is that "if R, then Q". R is a fact, and Q is the result caused by R.

The description as following:

<Rule>:: = IF< premise>THEN<conclusion>
<Premises>:: = AND|OR|NOT<condition>
<Conditions>:: = AND|OR|NOT<assertion>
<Assertions>:: = <verb><database><parameter>
<data>|<verb><database><parameter>
<Conclusions>:: = <Assertion><diagnose opinion>

Diagnose opinion is the final judging conclusion and operation method, e.g., "IF cannon is in march OR the included angle between barrel and gun is out of limits OR breechblock isn't at the close position, THEN not allow launching".

Take troubleshooting for example. In battle plan, assume that the bolt of oil filler hole leaks fluid in the recoil brake lever, after artillery arriving position with

Table 47.1 Fault database

Equipment types	Parts names and No.	Fault codes	Fault phenomenon	Animated description	Solutions	Ending animation
New type Equipment	Counter recoil mechanism 01	001	The bolt of oil filler hole leaks fluid in the recoil brake lever	01001.avi	Make sure the bolt of oil filler hole has been tightened; If sealing ring damaged, replace	01001 f.avi
		002	The bolt of oil drain hole leaks fluid in the recoil brake lever	01002.avi	Make sure the bolt of oil drain hole has been tightened; If sealing ring damaged, replace	01002 f.avi
		003	The recuperator charge valve leaks fluid and air	01003.avi	Confirm the charge valve has been tightened; Confirm sealing ring is OK; Replace the charge valve	01003 f.avi
		004	Counter-recoil too much	01004.avi	Check the recoil fluid volume; Adjust air pressure of recuperator	01004 f.avi
	Fire fighting and explosion suppression devices 02	001	The power light is out of work	02001.avi	Check all power switches; Check circuit	02001 f.avi
		002	Trouble lamp is on	02002.avi	Immediately turn off and check the cause of the malfunction	02002 f.avi
		003	When turn on switches, a few of lights aren't work at all	02003.avi	Make sure bulbs to be OK; Check circuit	02003 f.avi

Fig. 47.4 Simulation training environment

coordinates (x, y, and z); continue doing other tasks after checking and removing the fault. Control rules are set as following:

IF artillery coordinates equals x, y, and z, THEN set parameter S to 1;
IF parameter S equals 1, THEN transfer and display the file of 01001.avi in training window, AND set the bolt of oil filler hole to be loose OR set the sealing ring damaged;
IF the bolt has been tightened by trainees AND the sealing ring is OK, THEN set parameter S to 0;
IF parameter S equals 0, THEN transfer and display the file of 01001 f.avi in training window once, return artillery to the normal state and start other combat tasks.

47.4 Realization

The simulation training system can only provide information for trainees by their own senses of sight and hearing. The interactive devices used in this system include position tracker, data glove [9], HMD helmet, graphics workstations, and multimedia projector [1] and so on. Figures 47.4 and 47.5 show the hardware.

47.4.1 Program

Under the condition of Visual C++6.0 Integrated development environment, taking visual simulation tools Vega [6, 7], and introducing the MFC message response instrument as well, trainees can control the operation with help of interactive devices, which successfully make the equipment training process come true. The

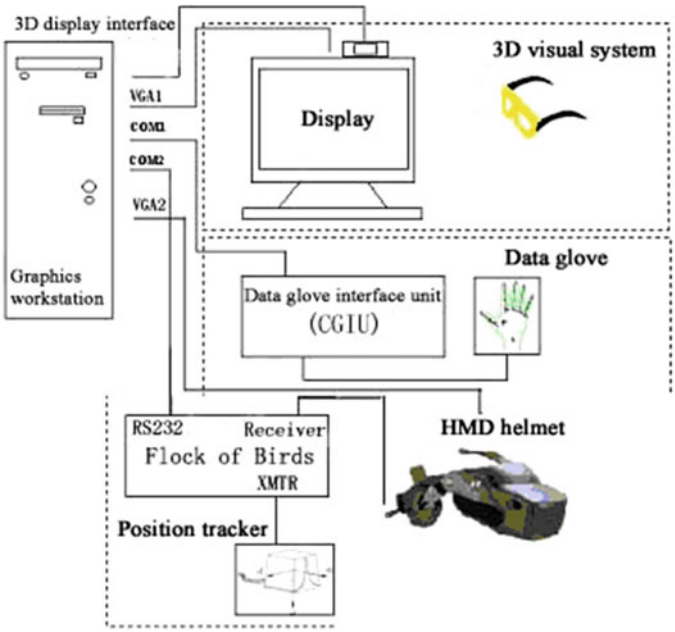


Fig. 47.5 Hardware configuration diagram of virtual operation environment

```
UINT runVegaApp (LPVOID pParam)
{
   CVirtualHandView* pOwner = (CVirtualHandView*) pParam;
   vgInitWinSys (AfxGetInstanceHandle (), pOwner–>GetSafeHwnd ());
   pOwner–>setVegaInitted (TRUE);
   pOwner–>postInit ();  // System initialization
   vgDefineSys (pOwner–>getAdfName ());  // Read the scene file, loading
virtual hand and equipment model
   pOwner–>setVegaDefined (TRUE);
   pOwner–>postDefine ();
   vgConfigSys ();  //system configuration
   pOwner–>setVegaConfiged (TRUE);
   pOwner–>postConfig ();
   while (pOwner–>getContinueRunning ())  // Enter the system main loop
   {
      if (pOwner–>bReceive)
       pOwner–>m_Comm.SetOutput (COleVariant (pOwner–>
    m_CommSendOut));
      vgSyncFrame ();  // Frame synchronization
      pOwner–>postSync ();
      vgFrame ();       // The current frame processing
      pOwner–>postFrame ();
   }
  pOwner–>setVegaInitted (FALSE);
  return 0;
}
```

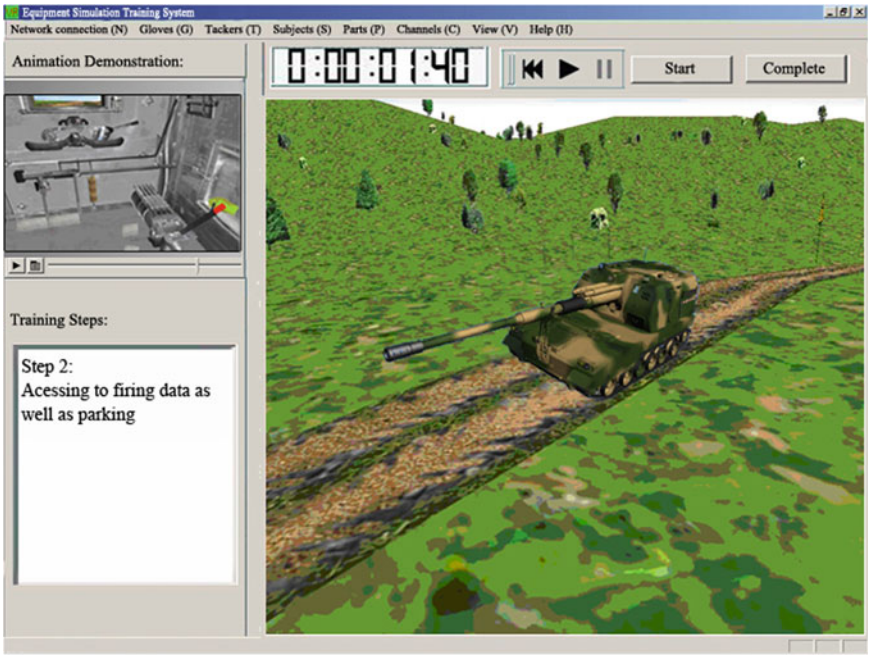


Fig. 47.6 Equipment model and battlefield morphology

code illustrated below is the main operating procedure which uses other functions to accomplish scene drawing, virtual hands import, virtual training proceeding and so on.

47.4.2 Interface

The system interface is provided with a big display window, a small display window and static text window. When learning, large and small display windows severally show operating processes from equipment external and internal, static text window provides a textual interpretation of the operation. When training and evaluation, trainees observe the battlefield through data helmet and operate the equipment through data gloves, moreover, to be monitored and to be directed by trainers through this interface. Figure 47.6 shows the system interface.

47.5 Conclusion

With the help of powerful virtual simulation software 3D MAX, Multigen Creator/Vega, the research has established simulation training system and accomplished the similar operation compared with real environment. It presents the ideas including solid modeling, 3D terrain rendering, fault databases generation and control rules setting. The system also reserves the third part of development interface for models, databases and battle plans. The system creates a favorable impression by using forces.

References

1. Ding, J.H., Wang, Y.G., Pan, Z.G.: Using large screen projection system the virtual Hefang roaming. J. Hangzhou Electron. Sci. Technol. Univ. **25**(1), 56–59 (2005) (In Chinese)
2. Guan, L., Shao, Y.W., Hao, C.Y. et al.: The research of virtual war training system based on direct3D. J. Proj. Rockets Missiles Guid. **26**(2), 107–109 (2006) (In Chinese)
3. Han, L.B., Hua, Y.Z.: Modern simulation technology constructs new weapons and equipment training simulator. J. Comput. Simul. **20**(10), 27–29 (2003) (In Chinese)
4. Li, H.Q., Hua, Y.Z.: OpenGL-based simulation visualization technology. J. Comput. Simul. **22**(6), 158–160 (2005) (In Chinese)
5. MultiGen Paradigm Inc. The multiGen creator desktop tutor. MultiGen Paradigm Inc, USA (2000)
6. MultiGen Paradigm Inc. Vega programmer's guide. MultiGen Paradigm Inc, USA (2001)
7. MultiGen Paradigm Inc. Vega man pages. MultiGen Paradigm Inc, USA (2001)
8. Sun, Y.K., Xiong, Z.Y.: Simulation design and realization for new type destroyer and frigate's operational system. J. Syst. Simul. **24**(4), 902–906 (2012) (In Chinese)
9. Zeng, F.F., Liang, B.L., Liu, Z. et al.: An interactive environment design based on data glove. J. Image Graph. **5**(2), 153–157 (2000) (In Chinese)
10. Zhou, Y.X.: Study for software architecture modelling. J. Softw. **9**, 866–871 (1998) (In Chinese)

Chapter 48
Super Sparse Projection Reconstruction of Computed Tomography Image Based-on Reweighted Total Variation

Gongxian Liu and Jianhua Luo

Abstract Sparse projection is an effective way to reduce the exposure to radiation during X-ray CT imaging. However, reconstruction of images from sparse projection data is challenging. In this paper, a novel method called reweight total variation (WTV) is applied to solve the challenging problem. And based on WTV, an iteration algorithm which allows the image to be reconstructed accurately is also proposed. The experimental results on both simulated and real images have consistently shown that, compared to the popular total variation (TV) method and the classical Algebra Reconstruction Technique (ART), the proposed method achieves better results when the projection is sparse, and performs comparably with TV and ART when the number of projections is relatively high. Therefore, the application of the proposed reconstruction algorithm may permit reduction of the radiation exposure without trade-off in imaging performance.

Keywords Sparse Projection Reconstruction · CT · Reweight total variation · Iterative Reconstruction

48.1 Introduction

X-ray computed tomography (CT) has played an important role in medical field. Nonetheless, exposing in strong X-ray intensity for a long time will do harm to people's health. An effective way to achieve the reduction of the radiation

G. Liu (✉)
School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
e-mail: liugxian2006@gmail.com

J. Luo
School of Aeronautics and Astronautics, Shanghai Jiao Tong University, Shanghai, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_48,

exposure is to reduce the number of projections required for reconstructing the image, but the image reconstructed from sparse projections often suffers from serious problems, such as blurring and artifacts.

Various algorithms have been developed to reconstruct image from sparse projections. There are mainly two methods, i.e., interpolating the missing data which is followed by image analytic reconstruction and iterative reconstruction. Numerous iterative algorithms have been developed for tomography image reconstruction. Among these algorithms mentioned, the widely used iterative algorithms for tomography imaging are the algebraic reconstruction technique (ART) and the expectation–maximization (EM) Algorithm. These methods differ in the constraints exposed on the image and the cost function that to be minimized [1]. Furthermore, they will result in artifacts as the projections reduce. For the case where the data is consistent yet is not sufficient to determine a unique solution to the imaging model, the ART algorithm finds the image that is consistent with the data and minimizes the sum-of-squares of the image pixel values [2]. In this paper, TV [3] and WTV [4] are introduced to reconstruct tomography images with super sparse projections. The two methods are actually iterative methods that differ in the cost function. WTV is a novel method for sparse image recovery and substantially less measurement is needed for exact recovery. It is a method that adds weights to TV, i.e., large coefficients in TV are penalized heavily than small coefficients [5].

The organization of this paper is as follows. Firstly, central slice theory and the main theory of TV and WTV are introduced in Sect. 48.2. Based on the theories mentioned in Sect. 48.2, experiments on both simulated images and real images are devised in Sect. 48.3. Finally, the conclusions are drawn and the future work is discussed in Sect. 48.4.

48.2 Theories

48.2.1 Central Slice Theory

The central slice theory is that if projecting an image to a line and doing a Fourier transform of the projection, it is equivalent to doing a two-dimensional Fourier transform to the image first and slicing through the original in the orthogonal direction. Then sparse projections can be converted into sparse radial spectral data according to the Fourier slice theorem. The Fourier transform of a parallel beam projection gives a slice of the two-dimensional (2-D) Fourier transform. Given an image, M projections in the image space are equivalent to sampling M radial lines of the image's spectrum, thus generating sparsely sampled radial spectral data. The smaller M is, the sparser the radial spectral data will be. Therefore, M is also referred to as sparse level of projections [6]. As Fig. 48.1 shows, suppose that there are 30 projections in the image domain, it is equivalent to getting data in the orthogonal directions in Fourier domain.

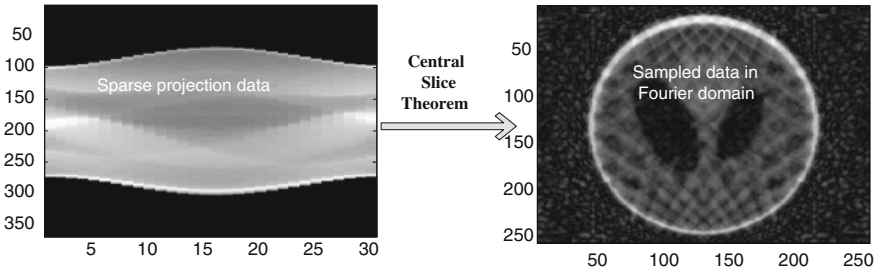


Fig. 48.1 Projection in image space is transformed to Fourier domain based on central slice theorem

48.2.2 Total Variation and Reweight Total Variation

Since CT image has a sparse or nearly sparse gradient, it is meaningful to search for the reconstruction with minimal TV norm, i.e.,

$$\min \|x\|_{TV} \quad s.t \quad y = \Phi x \tag{48.1}$$

In Eq. (48.1), $x_{n\times n}$is the image to be reconstructed and $\Phi_{n\times n}$ is the measurement matrix which is defined as $\Phi = MF$, where $M_{m\times n}$ is the mask matrix that samples data in the image in Fourier domain. The mask matrix is also called radial trace. $F_{n\times n}$ is Fourier masking operator, and $y_{m\times n}$ is the measured data. $\|x\|_{TV}$ is defined in Eq. (48.2):

$$\|x\|_{TV} = \sum_{1\le i,j\le n-1} \sqrt{\left(x_{i+1,j} - x_{i,j}\right)^2 + \left(x_{i,j+1} - x_{i,j}\right)^2} = \|Dx\| \tag{48.2}$$

And adding some weights to TV, it will get WTV which is defined as Eq. (48.3)

$$\|x\|_{WTV} = \sum_{1\le i,j\le n-1} W_{i,j}\sqrt{\left(x_{i+1,j} - x_{i,j}\right)^2 + \left(x_{i,j+1} - x_{i,j}\right)^2} = \|WDx\| \tag{48.3}$$

where $W_{i,j}$ is the weight of TV, which is defined in Eq. (48.4):

$$W_{i+1}^{l} = \frac{1}{\left\|x_{i,j}^{l}\right\|_{TV} + \varepsilon} \tag{48.4}$$

In Eq. (48.4), ε is set above zero to provide stability of the algorithm and this will ensure that a zero-valued component in x^{ℓ}does not strictly prohibit a nonzero estimate at the next step. Empirically, ε should be set slightly smaller than the expected nonzero magnitudes of x. In this paper, $\varepsilon = 0.1$. Then by solving the following equation, the exact reconstruction will be acquired.

$$\min \|x\|_{WTV} \quad s.t \quad y = \Phi x \tag{48.5}$$

Based on the above theory, the main steps of solving WTV are devised in the following.

- Step 1. Set $l = 0$and $W_{i,j}^0 = 1, \quad 1 \le i,j \le n$;
- Step 2. Solve the WTV minimization problem, $x^l = \arg\min \sum_{1 \le i,j \le n} \left\|x_{i,j}\right\|_{WTV}$ $s.t. \quad y = \Phi x$;
- Step 3. Update the weights for each $(i,j) \quad 1 \le i,j \le n,\ W_{i,j}^{l+1} = \frac{1}{\left\|x_{i,j}^l\right\|_{TV}+\varepsilon}$;
- Step 4. Terminate on convergence or when l gets the max iterations. Otherwise increases l and go to step 2.

A robust quasi-Newton method [7] is used in step 2. Consider Eq. (48.5), the energy functional is given by Eq. (48.6).

$$E(x,\lambda) = \|x\|_{WTV} + \frac{\lambda}{2}\|\Phi x - y\|^2 \tag{48.6}$$

where λ controls the tradeoff between solution sparsity and data fidelity. Then minima of E are yielded as solutions of the associated Euler–Lagrange Eq. (48.7).

$$L(x,\lambda) = \Psi^*\Psi x + \lambda\Phi^*(\Phi x - y) = 0 \tag{48.7}$$

where $\psi = WD$. Then consolidation of the target variable x yields

$$[\Psi^*\Psi + \lambda\Phi^*\Phi]x = \lambda\Phi^* y \tag{48.8}$$

Then robust quasi-Newton iteration is obtained for the computation of x.

$$x^{t+1} = x^t + \Delta^t \tag{48.9}$$

where

$$[\Psi^*\Psi + \lambda\Phi^*\Phi]\Delta^t = -L(x^t,\lambda) \tag{48.10}$$

48.3 Experimental Results

To evaluate the WTV and TV method, digital phantoms from popular Shepp-Logan image and one real image are used. The sparse projections are simulated by generating the specified number of uniformly distributed projections from the phantom and real images. And to evaluate the accuracy of the reconstructed image, we adopt the standard deviation (STD) of errors between the constructed image and reference image.

48.3.1 Measurement of the Accuracy

We consider the constructed image as $x(i,j)$ and the original image as $x_0(i,j)$, the STD is computed in Eq. (48.11).

$$STD = \sqrt{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}[e(i,j)-\overline{e}]^2} \tag{48.11}$$

where $e(i,j) = x(i,j) - x_0(i,j)$, and $\overline{e} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} e(i,j)$. The smaller the STD is, the less the reconstructed error is.

48.3.2 Reconstruction of Phantom Image

The performances of the ART, TV and WTV methods at different sparse level M are evaluated using both noise-free and noisy projections of the Shepp-Logan phantom image. The noisy projections are generated by adding to the noise-free projections zero-mean Gaussian noise with variance 0.0001. Sparse projections are simulated at the sparse levels ranging from 10 to 31 and images are reconstructed using the ART, TV and WTV methods. The experimental results are shown in Fig. 48.2. In Fig. 48.2b, c and d, WTV method can reconstruct the phantom image perfectly with only 10 projections, while image reconstructed by TV method suffers artifacts severely. In Fig. 48.2f, g and h, WTV performs a little better than TV method under noise condition. Notice from Fig. 48.2i, j that the WTV method constantly outperforms the ART and TV methods at all sparse levels in both noise-free and noisy cases. In Fig. 48.2j, The Y-Axis uses semi-log since the STD of WTV is approximate to zero.

48.3.3 Reconstruction of Real Images

To study the performance of TV and WTV methods in reconstructing images with complex structures, 99 axial slices of a real CT brain images (courtesy of North Carolina Memorial Hospital and University of North Carolina, http://www-graphics.stanford.edu/data/voldata/) are used. Fig. 48.3 shows the reconstruction results of the 61st slice. A line profile of a pertinent section of each image is shown in Fig. 48.3e, which shows WTV method performs better. Notice from Fig. 48.3f that the WTV method constantly outperforms the ART and TV methods at all sparse levels.

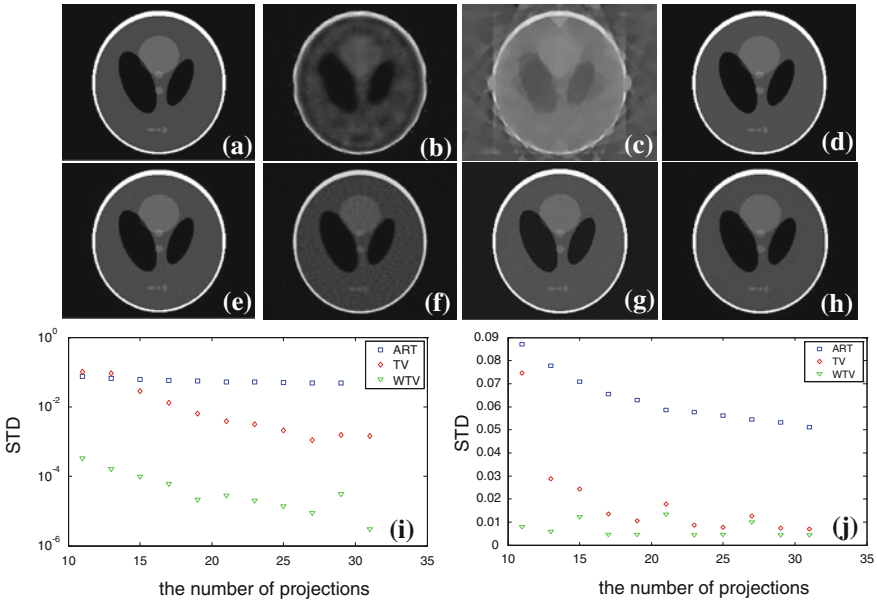


Fig. 48.2 Reconstruction of phantom image. **a** Is the noise-free image; **b**, **c** and **d** are images reconstructed by the ART, TV and WTV methods respectively with 10 projections; **e** is noised image, **f**, **g** and **h** are noised images reconstructed with 30 projections; **i** and **j** shows the STDs between reconstructed image and noise-free or noised image with 11–31 projections, respectively

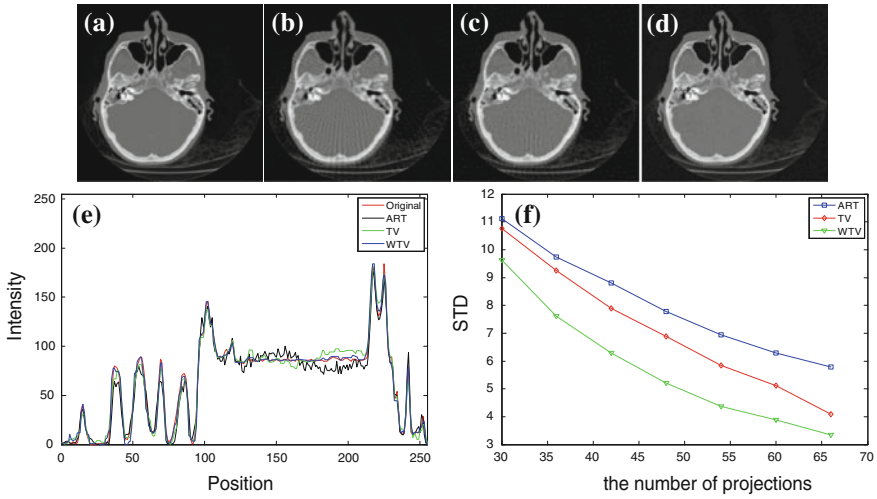


Fig. 48.3 Reconstruction of the slice 61 (brain), in which the number of projections is 66. **a** Is the reference image, **b**, **c** and **d** are the images reconstructed by the ART, TV and WTV methods, respectively. **e** Shows the line profile. **f** Shows the STD between the reference image and reconstructed image using 30–66 projections

48.4 Discussion and Conclusion

Numerical experiment results on both simulated and real data have shown that WTV has a better performance than classical TV method. WTV can reconstruct images more accurately by using fewer projections. As a consequence, the application of WTV method may permit reduction of the radiation exposure without trade-off in imaging performance. However, both methods perform unstably with increasing projections. Initial guess is that sampled data is uniform. And it is difficult to make general conclusions about the performance of WTV algorithm because its performance depends on the structure of the scanned object. During the process of iteration, ε is set to be a fixed value. How to get the best value of ε also needs to be researched. The future work is to search some algorithms to ensure the stability of reconstruction methods and doing more tests about other kinds of CT images. And testing the algorithm with adaptive ε to get the best one is another future task.

References

1. Sidky, E.Y., Kao, C.-M., Pan, X.: Accurate image reconstruction from few-views and limited-angle data in divergent-beam CT. J X-Ray Sci. Technol. **14**, 119–139 (2006)
2. Sidky, E.Y., Pan, X.: Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. Phys. Med. Biol. **53**, 4777–4807 (2008)
3. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Phys. D **60**, 259–268 (1992)
4. Candès, E.J., Wakin, M.B., Boyd, S.P.: Enhancing sparsity by reweighted $\ell 1$ minimization. J. Fourieer Anal. Appl. **14**, 877–905 (2008)
5. Candes, E., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theor. **52**, 489–509 (2006)
6. Luo, J., Liu, J., Li, W., Zhu, Y., Jiang, R.: Image reconstruction from sparse projections using S-transform. J. Math. Imaging Vision **43**, 227–239 (2012)
7. Trzasko, J., Manduca, A.: Highly undersampled magnetic resonance image reconstruction via homotopic $\ell 0$-Minimization. IEEE Trans. Med. Imaging **28**, 106–112 (2009)

Chapter 49
Sea Wave Filter Design for Cable-Height Control System of Anti-Submarine Helicopter

Yueheng Qiu, Weiguo Zhang, Pengxuan Zhao and Xiaoxiong Liu

Abstract The paper aims to solve the radio altitude signal mixed with sea wave noise which should be filtered as the helicopter executing the antisubmarine task. The sea wave is modeled based on the rational spectral approach and obtained in the form of white noise shaping filter as the sea wave color noise is changed into white noise. For the measurement equation and state equation can be formed including the estimated value, the sea wave filter is brought out according to the continuous Kalman filtering theory and added to the altitude channel. Lastly, the cable-height control system has taken radio altitude and normal acceleration which have been filtered as feedback signals. The effects of the filter in different wind speeds are separately verified by the digital simulations, and the results show the design approach is available and effective.

Keywords Radio altitude signal · The sea wave · Helicopter · Kalman filtering theory

49.1 Introduction

The helicopter should maintain the hovering state above the sea during the task of submarine [1]. In this case, the helicopter is not only required to maintain the attitude and speed stability, but also to keep the height stability which can be realized based on the cable-height control system.

Y. Qiu (✉) · W. Zhang · X. Liu
School of Automation, Northwestern Polytechnical University, Xi'an, China
e-mail: qiuyueheng@163.com

P. Zhao
Aircraft Design and Research Institute, Aviation Industry Corporation of china, Xi'an, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_49,

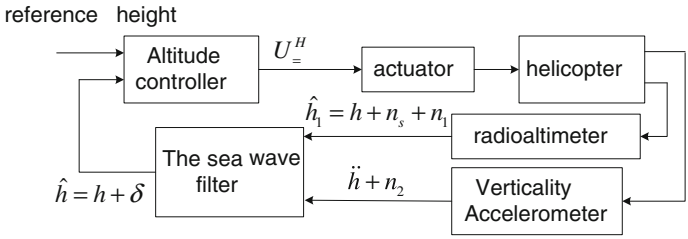


Fig. 49.1 The cable-height control scheme

Actually, the essential of the cable-height control system is the control of the flight altitude, but the input signal measured by the radio altimeter mixed with the sea wave noise will affect the precision of the control. So the sea wave filter is introduced into the altitude controller which can filter the measured values of the radio altimeter and get the altitude signal regardless of the interference of the sea wave [2]. The specific configuration of the cable-height control system is shown in Fig. 49.1.

The power spectrum can't be approximately described as the sea wave interference is a complex colored noise. The commonly used sea wave spectrum is Neumann spectrum, Pierso Moscowitz (PM) spectrum etc., and the modeling method of the sea waves are energy bisection and rational spectral [3–5].

In many historical documents, the commonly used methods of filtering are Wiener filter and Kalman filter [6–8]. But the Wiener filter only applies to the stationary stochastic process and the historical and current observational data are needed to estimate the current value of the signal. By contrast, the Kalman filter has no such restriction [9], and it doesn't need all the historical data and is suitable for the stable and unstable stochastic processes. In this paper, the Kalman filter algorithm is added into the altitude controller and the sea wave interference signal can be filtered to get the precision of height signal.

49.2 The Sea Wave Filter

Due to the error of initial value of integral, after two integral calculating of the normal acceleration signal measured by accelerometer, the precision of height of the helicopter will be reduced with the time-varying. This article used the filtering method combined with the two signals $\hat{h}_1$ and $\hat{\ddot{h}}$, and h is the actual height, $\ddot{h}$ is the actual normal acceleration, n_s is color noise, n_1 and n_2 is zero mean with white noise. The structure of the sea wave filter is shown in Fig. 49.2.

As shown in Fig. 49.2, the altitude signal $\hat{h}_2 = h + n_e + h_p$ can be obtained after two integral calculating of the output of accelerometer $\hat{\ddot{h}}$, and $n_e$is random error, h_p is the error due to the initial value of integral estimation. The estimate

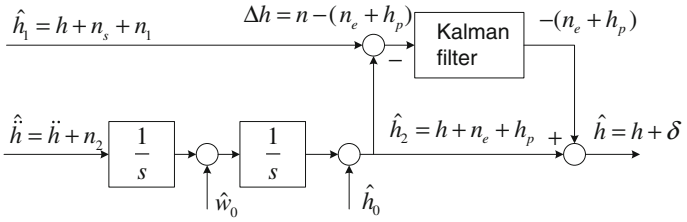


Fig. 49.2 The design of sea wave filter

value $-(n_e + h_p)$ obtained based on the difference Δh between $\hat{h}_1$ and $\hat{h}_2$ identified by the Kalman filter, and the value cancelled after $\hat{h}_2$ superimposed. Then the final output value is $\hat{h}$, and δ is caused by the incomplete filtering and initial value error of integral.

49.2.1 The Sea Wave Modeling

The PM spectrum can be described as follows,

$$S_{PM}(\omega) = \frac{\alpha g^2}{\omega^5} \exp\left[-\beta\left(\frac{g}{U\omega}\right)^4\right] \tag{49.1}$$

where $g = 9.81\ \mathrm{m/s^2}$, $\alpha = 8.1 \times 10^{-3}$, $\beta = 0.74$, U is the wind-speed 10 m above the sea surface.

The wave power spectrum described in Eq. (49.1) is not the form of rational spectral, and can be expressed as the form of approximately rational spectral using the least squares method. Using the rational spectrum [3], the shaping filter transfer function can be obtained.

The rational approximation spectrum of wave is

$$\hat{S}_{PM}(\omega) = \frac{a_0 + a_1\omega^2 + \cdots + a_m\omega^{2m}}{b_0 + b_1\omega^2 + \cdots + \omega^{2n}} = \frac{P_{PM}(\omega)}{Q_{PM}(\omega)} \tag{49.2}$$

If $n > m$, and

$$Q_{PM}(\omega)S_{PM}(\omega) = P_{PM}(\omega) + e \tag{49.3}$$

where e is the error, $S_{PM}(\omega)$ is the PM spectrum of wave expressed in Eq. (49.1). The Eq. (49.2) is taken into the Eq. (49.3), so

$$S_{PM}(\omega)\omega^{2n} = -S_{PM}(\omega)\omega^{2(n-1)}b_{n-1} - \cdots - S_{PM}(\omega)b_0 + \omega^{2m}a_m + \cdots + a_0 + e \tag{49.4}$$

The parameters $\theta = [b_{n-1} \cdots b_1 b_0 a_m \cdots a_1 a_0]^{\mathrm{T}}$ can be estimated, $z = S_{PM}(\omega)\omega^{2n}, h = \left[-S_{PM}(\omega)\omega^{2(n-1)} \cdots - S_{PM}(\omega)\omega^2 - S_{PM}(\omega)\omega^{2m} \cdots \omega^2 1\right]$, the Eq. (49.4) can be written as

$$z = h\theta + e \tag{49.5}$$

$S_{PM}(\omega)$ is divided into equidistant frequency ω_k with the energy interval $[0 \quad 0.95\omega]$, $k = 1, 2, \ldots, \mathrm{N}$ and $\mathrm{N} > \mathrm{n} + \mathrm{m} + 1$. Each of the corresponding frequency value is coincided with the Eq. (49.5), $z(k) = h(k)\theta + e(k)$, the matrix is written as

$$Z_N = H_N\theta + E_N \tag{49.6}$$

$$Z_N = [S_{PM}(\omega_1)\omega_1^{2n} \quad S_{PM}(\omega_2)\omega_2^{2n} \quad \cdots \quad S_{PM}(\omega_N)\omega_N^{2n}]^{\mathrm{T}}$$

$$E_N = [e_1 \quad e_2 \quad \cdots \quad e_N]^{\mathrm{T}}$$

$$H_N = \begin{bmatrix} -S_{PM}(\omega_1)\omega_1^{2(n-1)} & \cdots & -S_{PM}(\omega_1)\omega_1^2 & -S_{PM}(\omega_1) & \omega_1^{2m} & \cdots & \omega_1^2 & 1 \\ -S_{PM}(\omega_2)\omega_2^{2(n-1)} & \cdots & -S_{PM}(\omega_2)\omega_2^2 & -S_{PM}(\omega_2) & \omega_2^{2m} & \cdots & \omega_2^2 & 1 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -S_{PM}(\omega_N)\omega_N^{2(n-1)} & \cdots & -S_{PM}(\omega_N)\omega_N^2 & -S_{PM}(\omega_N) & \omega_N^{2m} & \cdots & \omega_N^2 & 1 \end{bmatrix}$$

The criterion function is

$$J(\theta) = \sum_{k=1}^{N} [e(k)]^2 = \sum_{k=1}^{N} [z(k) - h(k)\theta]^2 = (Z_N - H_N\theta)^{\mathrm{T}}(Z_N - H_N\theta) \tag{49.7}$$

Minimization of $J(\theta)$, the least squares estimation of the Parameter is

$$\hat{\theta} = \left(H_N{}^{\mathrm{T}} H_N\right)^{-1} H_N{}^{\mathrm{T}} Z_N \tag{49.8}$$

The approximation rational spectral function of wave is showed in Fig. 49.3.

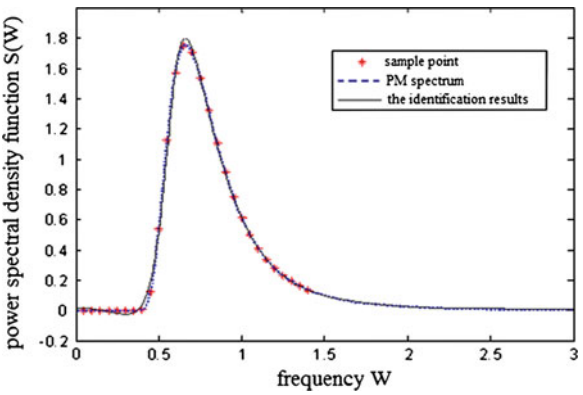


Fig. 49.3 The identification result

It can be seen from Fig. 49.3, the approximation rational spectral density function curve of wave using the method of least squares identification is very close to the PM spectral density curve.

49.2.2 The Shaping Filter

In order to apply the principle of Kalman filter to the design of wave filter, the wave noise should be changed into white noise by the method of shaping filter.

Firstly, calculate the zero-pole of $\hat{S}_{PM}(\omega)$, the zeros $(\omega_{z1}, \omega_{z3})$ and poles $(\omega_{p1}, \omega_{p3}, \omega_{p5}$ and $\omega_{p7})$ are selected in the upper half-plane of ω respectively according to the rational spectral theorem [3]. Then the transfer function of shaping filter can be obtained.

After the zero mean white noise n_d is taken into the shaping filter, the output n_s is the wave process,

$$n_s = \Phi(s) \cdot n_d \tag{49.9}$$

Finally, Eq. (49.9) is changed into the form of state equation

$$\begin{cases} \dot{x} = Ax + Bn_d \\ n_s = Cx \end{cases} \tag{49.10}$$

49.2.3 The Sea Wea Kalman Filter

The system and measurement equation is

$$\begin{cases} \dot{X}(t) = F(t)X(t) + G(t)w(t) \\ Z(t) = H(t)X(t) + v(t) \end{cases} \tag{49.11}$$

In Eq. (49.11), the system noise $w(t)$ and measurement noise $v(t)$ is zero mean and is not correlated with white noise.

The basic equation of Kalman filter is

$$\dot{\hat{X}}(t) = F(t)\hat{X}(t) + K(t)[Z(t) - H(t)\hat{X}(t)] = [F(t) - K(t)H(t)]\hat{X}(t) + K(t)Z(t) \tag{49.12}$$

$$K(t) = P(t)H^{\mathrm{T}}(t)r^{-1}(t) \tag{49.13}$$

$$\dot{P}(t) = P(t)F^{\mathrm{T}}(t) + F(t)P(t) - P(t)H^{\mathrm{T}}(t)r^{-1}(t)H(t)P(t) + G(t)q(t)G^{\mathrm{T}}(t) \tag{49.14}$$

The sea wave Kalman filter requires the statistical properties of the system equation, measurement equation, white noise excitation and measurement error.

The measurement,

$$Z = n - (s + h_p) = n_s + n_1 - \int_0^\tau \left[\int_0^\tau n_2(t)dt \right] dt - (\hat{w}_0 - w_0)t - (\hat{h}_0 - h_0) \tag{49.15}$$

In Eq. (49.15), $\hat{w}_0$ and $\hat{h}_0$ are the initial estimate values of the normal velocity w_0 and altitude h_0 respectively.

In order to translate Eq. (49.15) into the form of state space, two state variables $y = [x_1 \quad x_2]^T$ should be introduced, such as

$$\begin{aligned} x_1 &= \int_0^\tau [\int_0^\tau n_2(t)dt]dt + (\hat{w}_0 - w_0)t + (\hat{h}_0 - h_0), x_2 \\ &= \int_0^\tau n_2(t)dt + (\hat{w}_0 - w_0) \text{ and } x_1(0) = \hat{h}_0 - h_0,\ x_2(0) = \hat{w}_0 - w_0. \end{aligned}$$

$$\begin{cases} \dot{y} = \bar{A}y + \bar{B}n_2 \\ -(s + h_p) = \bar{C}y \end{cases} \tag{49.16}$$

Combined with the Eqs. (49.10) and (49.16), the system and measurement equations can be obtained as

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & \bar{A} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} B & 0 \\ 0 & \bar{B} \end{bmatrix} \begin{bmatrix} n_d \\ n_2 \end{bmatrix}$$

$$Z = \begin{bmatrix} C & \bar{C} \end{bmatrix} [x \quad y]^T + n_1$$

The variances of zero mean white noise n_1, n_2, n_3 are Var_1, Var_2, Var_d respectively. The parameter matrixes of basic equations of the Kalman filter are

$$F(t) = \begin{bmatrix} A & 0 \\ 0 & \bar{A} \end{bmatrix}, H(t) = \begin{bmatrix} C & \bar{C} \end{bmatrix}, G(t)q(t)G^{\mathrm{T}}(t) = \begin{bmatrix} B & 0 \\ 0 & \bar{B} \end{bmatrix} \begin{bmatrix} \text{Var}_d & 0 \\ 0 & \text{Var}_2 \end{bmatrix} \begin{bmatrix} B & 0 \\ 0 & \bar{B} \end{bmatrix}^{\mathrm{T}}$$

$$r(t) = \text{Var}_1$$

The Riccati equation can be solved and the $P(t)$, $K(t)$ are calculated, then the differential equation about $\hat{X}(t)$ is obtained, and the estimate of the reservation term $-(s + h_p)$ can be collected finally.

$$\dot{\hat{X}}(t) = [F(t) - K(t)H(t)]\hat{X}(t) + K(t)Z(t) - (s + h_p) = C_e(t)\hat{X}(t)$$

$$C_e = [0 \quad 0 \quad 0 \quad 0 \quad -1 \quad 0]$$

49.3 The Filtering Result

The variance of radio altimeter measurement noise n_1 is $\mathrm{Var}_1 = 2.5$, and the variance of accelerometer measurement noise $n_2$is $\mathrm{Var}_2 = 0.01$. The simulation implemented by the different values of the wind speed as shown in Figs. 49.4 and 49.5.

Scenario one: The wind speed is $U = 13$ m/s.

Scenario two: The wind speed is $U = 8$ m/s.

It is clear from Figs. 49.4 and 49.5 that the wave noise and radio altimeter noise are suppressed as the mean and variance of the height error approaching zero. The filtering result is satisfactory and it can be seen that the sea wave noise changes with wind speed.

The statistic characteristics of sea wave noise n_s and height error δ after two experiments can be seen in Table 49.1.

In Table 49.1, the statistical properties of height errorδ affected by the increased sea wave noisen_s, the Kalman filter coefficient matrix should be recalculated. The result of different intensity sea wave noise shown in Fig. 49.6, the effect of adjusted Kalman filter is satisfactory.

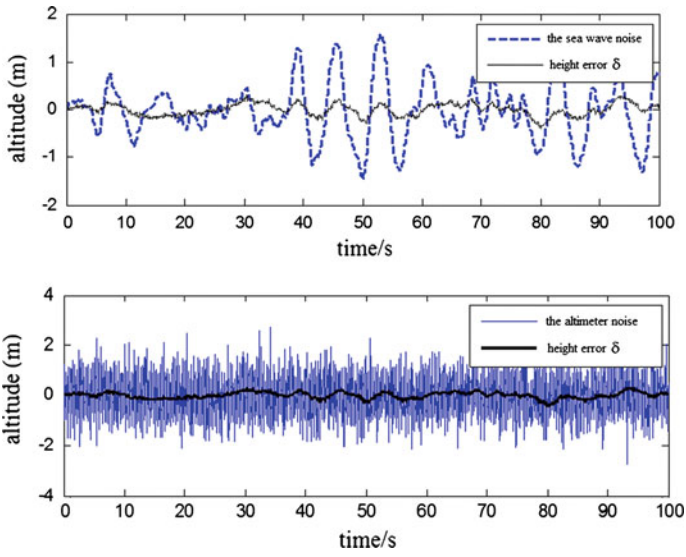


Fig. 49.4 The result of sea wave filter ($U = 13$ m/s)

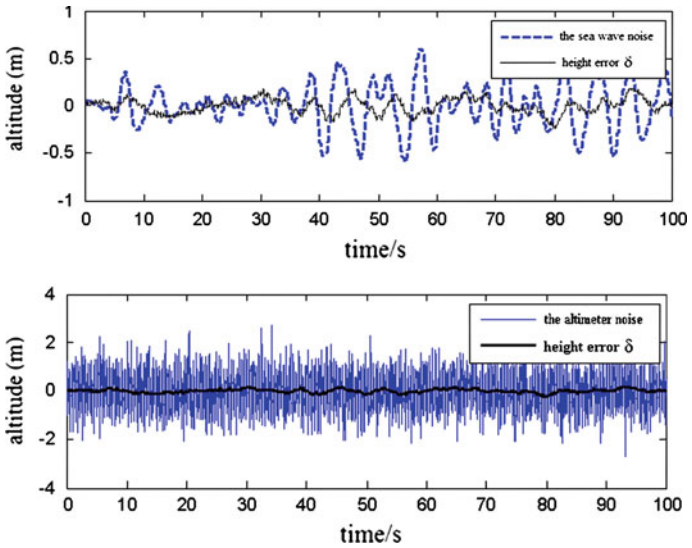


Fig. 49.5 The result of sea wave filter ($U = 8$ m/s)

Table 49.1 The statistic characteristic

Statistic		Scenario 1	Scenario 2
Sea wave noise n_s	Mean	−0.0297	−0.000095691
	Variance	0.3601	0.0548
Height error δ	Mean	−0.0030	0.00011903
	Variance	0.0155	0.0061

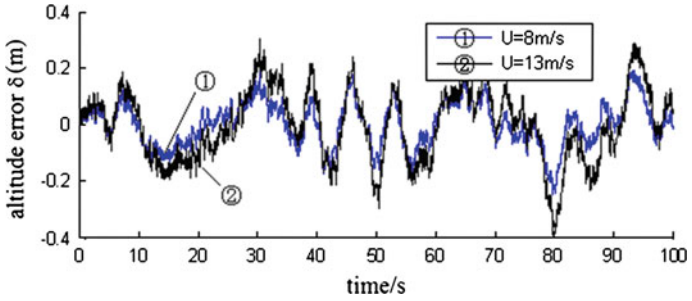


Fig. 49.6 The different result of sea wave filter

49.4 Conclusion

In summary, this paper describes an effective method to filter the measured values of the radio altimeter mixed with sea wave noise. Firstly, the PM spectrum function expression is given to describe the commonly used power spectrum and the sea wave is modeled based on rational spectral method. Secondly, the sea wave filter is designed based on the Kalman filter principle. Lastly, two sets of experiments designed in the situation of different wind speeds showed that the system has a good filtering effect.

References

1. Ning, D.: Design and simulate the helicopter anti-submarine control system. Helicopter Tech. **2**(1), 23–26 (2003)
2. Zhongjian, L., Jinwen, A.: Simulation of antisubmarine helicopter's cable-orientation and cable-height control system. Flight Dyn. **3**(18), 72–75 (2000)
3. Xiren, Z., Yan, Z., Yehe, H.: The Rational spectrum modeling of the ocean wave and its simulation method. J. syst. Simul. **2**(4), 33–39 (1992)
4. Qiu, H.: Establishing and simulation for random ocean state model. J. Syst. Simul. **3**(12), 226–228 (2000)
5. Belmont, M.R., Horwood, J.M.K., Thurley, R.W.F., Baker, J.: Filters for linear sea-wave prediction. Elsevier Ocean Eng. **17**(33), 2332–2351 (2006)
6. Alexander, S.T., Ghirnikar, A.L.: A method for recursive least squares filtering based upon an inverse QR decomposition. IEEE Trans. Signal Process. **1**(41), 20–30 (1993)
7. Douglas, S.C., Pan, W.: Exact expectation analysis of the LMS adaptive filter. IEEE Trans. Signal Process. **1**(43), 2863–2871 (1995)
8. Geromel, J.C.: Optimal linear filtering under parameter uncertainty. IEEE Trans. Signal Process. **1**(47), 168–175 (1999)
9. Qiu, Y., Zhang, W., Liu, X., Zhao, P.: EMMAE failure detection system and failure evaluation over flight performance. Int. J. Intell. Comput. Cybern. **3**(5), 401–419 (2012)

Chapter 50
Reversible Watermarking Based on Prediction-Error Expansion for 2D Vector Maps

Mingqin Geng, Yuqing Zhang, Puyi Yu and Yifu Gao

Abstract In order to increase the embedding capacity when one hides information in 2D vector maps in a reversible way, researchers present a novel algorithm based on prediction-error expansion. In two neighboring coordinate values, researchers use the latter as the predicted value of the former. The difference between the original value and the predicted one is the prediction error. Then, the prediction error is expanded to hide one watermark bit. To control the distortion, researchers preset a threshold. Only those satisfying the threshold condition are selected for embedding. The others are shifted so that the decoder can identify the marked coordinate values by the range that their prediction errors fall into. Experiment results show that bits per coordinate (bpc) on the test river map is 0.7. The algorithm has good performance in capacity and RMSE value. The proposed algorithm can be used in 2D vector maps for data integrity protection, etc.

Keywords Reversible watermarking · Prediction-error expansion · 2D vector map · Prediction-error shifting

50.1 Introduction

Reversible watermarking, which can exactly restore the original cover data after extracting the hidden data, is suitable for content integrity authentication of the map data. Reversible watermarking for raster images is an active research field.

M. Geng (✉) · Y. Zhang
School of Information Engineering, China University of Geosciences (Beijing), Beijing, China
e-mail: gengmq@cugb.edu.cn

P. Yu
Information Technology Center, BGP, CNPC, Beijing, China

Y. Gao
The Research Institute of Petroleum Exploration and Development, Beijing, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_50,

Many algorithms have been proposed and they are classified into two categories. Type-I algorithms [2, 6, 7] employ additive spread spectrum techniques combined with modulo addition and are robust. Type-II algorithms increase the embedding capacity by lossless compression technology [3, 5], difference expansion technology [1, 4, 9, 10] and histogram modification technology [8], but they give up the robustness. However, the research on reversible watermarking for 2D vector maps has not attracted attention. The first algorithm in this literature, which was proposed by Voigt et al. [11], used the highest frequency coefficient modification technique to hide the watermark bit. Wang et al. [12] applied Tian's difference expansion [10] to 2D vector maps and embedded the hidden data by expanding the differences between two adjacent coordinates. Zhou et al. [14] can control the capacity by difference histogram.

Similar to Tian's algorithm [10] based on difference expansion; Wang's algorithm uses the location map to record the expansion locations. Though the location map is compressed, it consumes large part of the embedding capacity. Zhou's algorithm [14] replaces the location map with histogram shifting, but it embeds data in disjoints pairs. And thus it does not make good use of the coordinates. In this paper, researchers present a novel high capacity reversible watermarking based on prediction error. Researchers embed the watermark bits into every coordinate value except the last one. In other words, researchers hide the data into consecutive coordinates instead of disjoint pairs. The prediction errors with smaller magnitudes are expanded to embed watermark bits. The others are shifted in order that the marked prediction errors and the unmarked ones fall into different ranges. Thus the decoder can differentiate them without any additional information. Hence, this algorithm does not need the location map. Accordingly, the computational complexity is low without any compression and the capacity is increased largely.

The rest of this paper is organized as follows. The key idea of Thodi's algorithm based on prediction-error expansion [9] is introduced in Sect. 50.2. In Sect. 50.3 researchers discuss the basic principle of the algorithm, including calculating the prediction error, prediction-error expansion and shifting, the embedding condition, and the embedding and extraction progresses. Experimental results are shown in Sect. 50.4. And at last, the conclusion is drawn in Sect. 50.5.

50.2 Thodi's Algorithm Based on Prediction-Error Expansion

Thodi et al. [9] proposed the algorithm based on prediction-error expansion. In this algorithm, the current pixel value is predicted by its three neighboring pixels. The prediction error between the current pixel value and its predicted one is expanded to embed one bit in the same way as Tian's difference expansion [10].

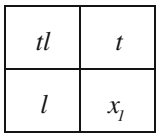


Fig. 50.1 Context of a pixel

This algorithm predicts the current pixel value x_1 from its context. The context is shown in Fig. 50.1. The predictor borrows from Ref. [13]. The predicted value $\hat{x}_1$ is calculated by (50.1), where t, l and tl are the original pixel values.

$$\hat{x}_1 = \begin{cases} \min(t,l), & tl \geq \max(t,l) \\ \max(t,l), & tl \leq \min(t,l) \\ t + l - tl, & otherwise \end{cases} \tag{50.1}$$

The prediction error, which is denoted as p_e, is the difference between the original value x_1 and the predicted one $\hat{x}_1$:

$$p_e = x_1 - \hat{x}_1. \tag{50.2}$$

The binary representation of the prediction error p_e is shifted left by one bit and one watermark bit b ($b \in \{0, 1\}$) can be appended to it as the new least significant bit (LSB). That is

$$p'_e = 2p_e + b. \tag{50.3}$$

After embedding one bit into the prediction error, the watermarked pixel value is calculated as:

$$x'_1 = \hat{x}_1 + p'_e. \tag{50.4}$$

At the decoder, the predicted value $\hat{x}_1$ and the prediction error p'_e are first calculated by (50.1) and (50.2), respectively. Then, the original prediction error p_e is calculated as

$$p_e = \left\lfloor \frac{p'_e}{2} \right\rfloor. \tag{50.5}$$

and the extracted watermark bit b is

$$b = p'_e - 2p_e. \tag{50.6}$$

At last, the original pixel value x_1 is recovered with the predicted value $\hat{x}_1$ and the restored prediction error p_e as

$$x_1 = \hat{x}_1 + p_e. \tag{50.7}$$

50.3 The Proposed Reversible Watermarking Scheme

In 2D vector maps, the coordinate values of the adjacent vertices appear high correlation. For two neighboring coordinates, the first coordinate value is predicted by the second one. The prediction error is expanded for embedding. To identify the embedded prediction errors, the others are shifted. Embedding condition is also investigated to ensure the distortion is not greater than the map precision tolerance.

50.3.1 Prediction-Error Expansion and Shifting

In 2D vector maps, let (x_1, x_2) be two neighboring coordinates. The current coordinate is x_1 and its predicted value is calculated by x_2, that is

$$\hat{x}_1 = x_2. \tag{50.8}$$

Its prediction error is calculated as

$$p_e = x_1 - \hat{x}_1 = x_1 - x_2. \tag{50.9}$$

The prediction error can be expanded and then one bit can be embedded into it. To control the distortion, researchers preset a threshold T_h. If the prediction error meets the condition $-T_h - 1 \le p_e \le T_h$, it is grouped into set E. Otherwise, it belongs to set S. For each prediction error in set E, researchers shift its binary presentation left by 1 bit and embed one bit as (50.3). After embedding, the modified prediction errors in set E will lie in the range $[-2T_h - 2, 2T_h + 1]$. On the other hand, the original prediction errors in set S occupy the range $(-\infty, -T_h - 2] \cup [T_h + 1, +\infty)$. Hence, the ranges that the new prediction errors in set E and the original ones in set S lie in will overlap in the range $[-2T_h - 2, -T_h - 2] \cup [T_h + 1, 2T_h + 1]$. To avoid the overlap, the prediction errors in set S are shifted left or right by $T_h + 1$. That is

$$p'_e = p_e + sign(p_e) \times (T_h + 1). \tag{50.10}$$

where the function $sign(n)$ returns the sign of number n. After shifting, the new prediction errors in set S will fall into the range $(-\infty, -2T_h - 3] \cup [2T_h + 2, +\infty)$. Thus, the decoder can distinguish the two sets by the ranges that their prediction errors fall into. In the same way, after shifting the modified coordinate value is calculated by (50.4).

At the decoder, for each prediction error in set S after calculating the predicted value $\hat{x}_1$ and the prediction error p'_e by (50.8) and (50.9) respectively, the original prediction error is recovered as

$$p_e = p'_e - sign(p'_e) \times (T_h + 1). \tag{50.11}$$

Then the original coordinate value is restored by (50.7).

50.3.2 Embedding Condition

Researchers hide the watermark information by modifying the coordinate values. This will introduce the distortion to the coordinate values. The distortion must be less than or equal to the map precision tolerance τ, that is

$$\left|x_1' - x_1\right| \leq \tau. \tag{50.12}$$

In this algorithm, researchers control the distortion by the predefined threshold T_h.

If researchers expand the prediction error and embed the watermark bit b, the embedded coordinate value is $x_1' = x_1 + p_e + b$ by (50.2), (50.3) and (50.4). Then condition (50.12) can be changed to $-\tau \leq p_e + b \leq \tau$. Researchers investigate the condition in two cases, i.e. $p_e \geq 0$ and $p_e < 0$. If $p_e \geq 0$, the maximum value of $p_e + b$ is $p_e + 1$ when $b = 1$, and hence, the condition $p_e + b \leq \tau$ can be simplified to $p_e + 1 \leq \tau$. Because of $p_e \leq T_h$, the previous inequality $p_e + 1 \leq \tau$ can be changed to $T_h + 1 \leq \tau$. Therefore, the embedding condition is $T_h \leq \tau - 1$ when $p_e \geq 0$. On the other hand, if $p_e < 0$, the minimum value of $p_e + b$ is $p_e + 0$ when $b = 0$. Thus, the condition $p_e + b \geq -\tau$ is equivalent to $p_e \geq -\tau$. Since $p_e \geq -T_h - 1$, the above inequality is changed to $-T_h - 1 \geq -\tau$, i.e. $T_h \leq \tau - 1$. When expansion embedding, the threshold should satisfy the condition:

$$T_h \leq \tau - 1. \tag{50.13}$$

If researchers shift the prediction error, the new coordinate value is calculated as $x_1' = x_1 + sign(p_e) \times (T_h + 1)$ by (50.2), (50.4) and (50.10). Then researchers change condition (50.12) to $T_h + 1 \leq \tau$. That is

$$T_h \leq \tau - 1. \tag{50.14}$$

To conclude the above analysis, the condition that the threshold T_h should obey when the prediction error is expanded and shifted is the intersection of inequalities (50.13) and (50.14):

$$T_h \leq \tau - 1. \tag{50.15}$$

50.3.3 Embedding Progress

In 2D vector maps, every vertex is composed of X and Y coordinates. Hidden data can be embedded in X and Y coordinates in the same way. Researchers take X coordinate for example. Suppose the X coordinate sequence in order is $\hat{x}_i$, where N is the number of the vertices. The embedding progress begins with x_1 and operates every X coordinate till x_{N-1}. The last coordinate x_N remains unchanged.

The watermark bit-stream is $\hat{x}_i$, where L is the length of the stream. For each coordinate value x_i, $\hat{x}_i$, the embedding progress consists of 3 steps.

1. Researchers calculate the predicted value $\hat{x}_i$ and the prediction error $p_{e,i}$ of the current coordinate value x_i by (50.8) and (50.9), respectively.
2. If the prediction error meets the condition $-T_h - 1 \leq p_{e,i} \leq T_h$, researchers apply (50.3) to expand the prediction error and embed one bit b_l in B,$1 \leq l \leq L$; Otherwise, researchers shift the prediction error as (50.10).
   The modified prediction error is denoted as $p'_{e,i}$.
3. Researchers calculate the new coordinate value x'_i by (50.4) with $\hat{x}_i$ and $p'_{e,i}$.

50.3.4 Extraction Progress

In the marked 2D vector map, X coordinates in the same order as embedding form the sequence $(x'_1, x'_2, \ldots, x'_{N-1}, x_N)$, where x_N is the original coordinate value. Unlike embedding progress, extraction and recovering progress is carried out in inverse order and begins with x'_{N-1}. For each coordinate value x'_i, $1 \leq i \leq N-1$, researchers extract the watermark and recover the original coordinate value as follows.

First, the predicted value $\hat{x}_i$ of the current coordinate value x'_i is calculated by (50.8). And then, researchers calculate its prediction error $p_{e,i}$ as (50.9).

If the prediction error falls into the range $[-2T_h - 2, 2T_h + 1]$, the original prediction error $p_{e,i}$ is recovered by (50.5) and researchers extract the watermark bit as (50.6);

Otherwise, researchers restore the original prediction error $p_{e,i}$ by (50.11).

Researchers restore the original coordinate value by (50.7) with $\hat{x}_i$ and $p_{e,i}$.

The extracted bit-stream is $b_L b_{L-1} \cdots b_2 b_1$, where L is the length of the stream. It is clear that the extracted bit-stream is in the opposite order.

50.4 Experimental Results

Researchers test the proposed algorithm in two maps, i.e. the river map and the contour map. The original river map is shown in Fig. 50.2 and the part of the original contour map is shown in Fig. 50.3. They have different features, including the scale, the number of vertices and so on. The features of the maps are listed in Table 50.1.

Researchers hide data in both X and Y coordinates. The proposed algorithm is compared with Wang's and Zhou's algorithms. The capacity is calculated as bits per coordinate (bpc) and the distortion is measured by root mean square error

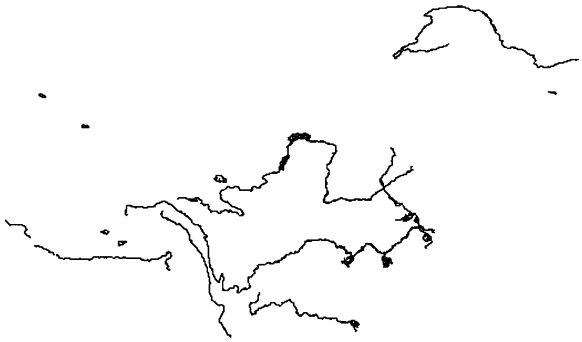


Fig. 50.2 Original river map

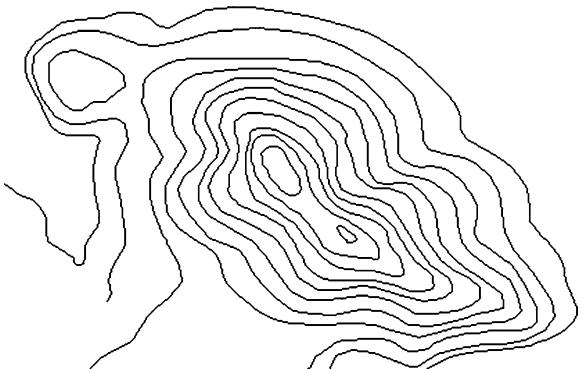


Fig. 50.3 Original contour map

Table 50.1 Features of original maps

Map	Scale	Number of vertices	T(meter)	Decimal digits
River	1:4000000	68925	500	6
Contour	1:10000	176526	6	7

(RMSE).The capacity versus RMSE value is shown in Figs. 50.4 and 50.5. Experimental results show that the proposed algorithm achieves higher capacity at lower distortion than Wang's [12]. When researchers compare the proposed algorithm with Zhou's [14], they have similar RMSE values at lower capacity, but the proposed algorithm increases embedding capacity largely.

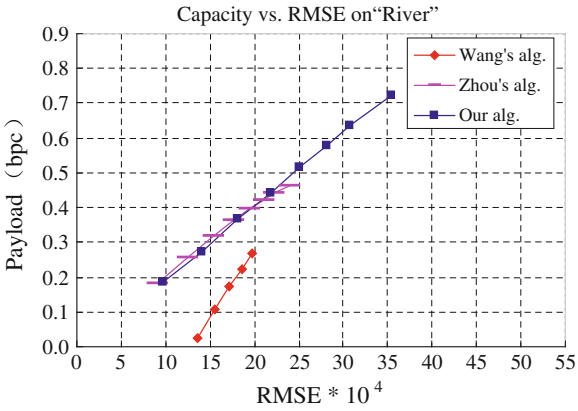


Fig. 50.4 Capacity versus RMSE on "River"

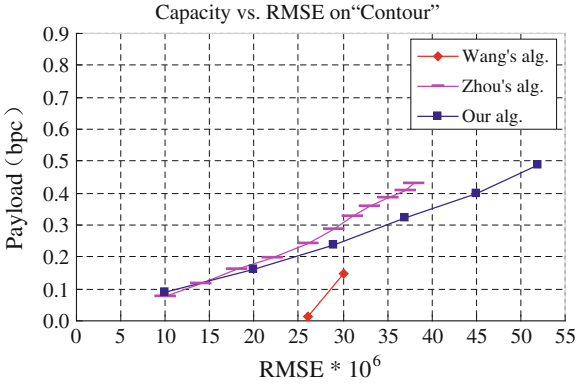


Fig. 50.5 Capacity versus RMSE on "Contour"

50.5 Conclusion

Researchers propose a novel reversible watermarking based on prediction-error expansion for 2D vector maps. For two adjacent coordinates, the predicted value of the first one is determined by the next one. The prediction error is expanded to embed data. They shift the unexpanded prediction errors in order that the decoder can distinguish the expanded locations without the location map. Experimental results show the proposed algorithm has good performance.

Acknowledgments This research is supported by the Fundamental Research Funds for the Central Universities.

References

1. Alattar, A.M.: Reversible watermark using the difference expansion of a generalized integer transform. IEEE Trans. Image Process. **13**, 1147–1156 (2004)
2. Bender, W., Gruhl, D., Morimoto, N., Lu, A.: Techniques for data hiding. IBM Syst. j. **35**, 313–336 (1996)
3. Celik, M.U., Sharma, G., Tekalp, A.M., Saber, E.: Lossless generalized-LSB data embedding. IEEE Trans. Image Process. **14**, 253–266 (2005)
4. Coltuc, D.: Low distortion transform for reversible watermarking. IEEE Trans. Image Process. **21**, 412–417 (2012)
5. Fridrich, J., Goljan, M., Du, R.: Invertible authentication. In: Proceedings of SPIE, Security and Watermarking of Multimedia Contents, vol. 4314, pp. 197–208 (2001)
6. Honsinger, C.W., Jones, P.W., Rabbani, M., Stoffel, J.C.: Lossless recovery of an original image containing embedded data. U.S. Patent 6, 278:791 (2001)
7. Macq, B.: Lossless multiresolution transform for image authenticating watermarking. In: Proceedings of EUSIPCO, pp. 533–536 (2000)
8. Ni, Z.C., Shi, Y.Q., Ansari, N., Su, W.: Reversible data hiding. IEEE Trans. Circuits Syst. Video Technol. **16**, 354–362 (2006)
9. Thodi, D.M., Rodríguez, J.J.: Expansion embedding techniques for reversible watermarking. IEEE Trans. Image Process. **16**, 721–730 (2007)
10. Tian, J.: Reversible watermarking by difference expansion. In: Proceedings of Workshop on Multimedia and Security, pp.19–22 (2002)
11. Voigt, M., Yang, B., Busch, C.: Reversible watermarking of 2D-vector data. In: Proceedings of the 2004 Multimedia and Security Workshop on Multimedia and Security, pp.160–165 (2004)
12. Wang, X.T., Shao, C.Y., Xu, X.G., Niu, X.M.: Reversible data-hiding scheme for 2-D vector maps based on difference expansion. IEEE Trans. Inf. Forensics Secur. **2**, 311–320 (2007)
13. Weinberger, M., Seroussi, G., Sapiro, G. LOCO-I: a low complexity, context-based, lossless image compression algorithm. In: Proceedings of the IEEE Data Compression Conference, pp. 140–149 (1996)
14. Zhou, L., Hu, Y.J., Zeng, H.F.: Reversible data hiding algorithm for vector digital maps. J. Comput. Appl. **29**, 990–993 (2009)

Chapter 51
An Object Tracking Approach Based on Hu Moments and ABCshift

Xingye Wang, Zhenhai Wang and Kicheon Hong

Abstract Robust visual tracking has become an important topic in the field of computer vision. The integration of cues such as color, shape features has proved to be a promising approach to robust visual tracking. In this paper, an algorithm is presented which integrates Hu moments and color histogram. Moreover, this paper integrates the ABCshift algorithm to overcome color features drawbacks which easily lead to loss of target object when the color of object is similar to the color of background. The proposed algorithm has been compared with other trackers using challenging video sequences. Experimental work demonstrates that the proposed algorithm has strong robust and improves the tracking performance.

Keywords Object tracking · Hu invariant moment · ABCshift · Fusion

51.1 Introduction

Object tracking is a challenging problem in computer vision. It is widely applied in various fields, such as intelligence surveillance and monitoring [1], perceptual user interfaces [2], smart rooms, smart city [3], and video compression etc.

In recent years, many object tracking approaches have been proposed, for example, Mean shift [4], Camshift [5] and particle filter [6]. Among of these, Mean shift and Camshift have been widely adopted because of their relative simplicity and low computational cost. Color is used as a feature for histogram-based

X. Wang · Z. Wang (✉)
School of Informatics, Linyi University, Linyin, Shandong, China
e-mail: lywzhh@163.com

K. Hong
Department of Information & Telecommunications Engineering, The University of Suwon
Wau-ri, Bongdam-eup, Hwaseong-si, Gyeonggi-do 445-743, Korea
e-mail: kchong@suwon.ac.kr

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_51,

appearance representation in these methods. Color histogram is scale and rotation invariant and it can handle occlusion in a certain degree. As global statistic information, color histogram does not provide discriminative localization ability [7]. When the color of target is similar to the color of background, tracking will become unstable and sometimes even lead to loss of target object.

A good tracking algorithm should be able to work well in various difficult situations such as various illuminations, background clutter, and occlusion. There are two technique trends in the computer vision tracking community. One is to develop more inherently robust algorithms and another is to employ multiple cues to enhance tracking robustness. To increase the robustness and generality of tracking, various image features must be employed. Every single cure has its own advantages and disadvantages [8].

In this paper, we propose an object tracking method fusing color feature and shape feature that overcome the above mentioned drawbacks of Mean shift or Camshift. Hu invariant moments [9] are the one of region based features and they are a very popular shape measure. They are invariant to translation, scale change and rotation. Moreover, its computation is simply. To improve the robustness of tracking, we use adaptive background Camshift (ABCshift) [10] as a kernel tracker.

The paper is organized as follows. Section 51.2 introduces the Hu moments. The Camshift and ABCshift algorithm is presented in Sect. 51.3. The proposed object tracking approach is investigated in Sect. 51.4. The experimental results and conclusion are finally described in Sects. 51.5 and 51.6, respectively.

51.2 Hu Invariant Moments

51.2.1 Extraction for Hu Moments

For an image $f(x, y)$, size is $M \times N$. The central moment of order $(p + q)$ is defined as:

$$m_{pq} = \sum_{x=1}^{M} \sum_{y=1}^{N} x^p y^q f(x, y), \quad p, q = 0, 1, 2, \cdots \tag{51.1}$$

Central moments are defined as:

$$u_{pq} = \sum_{x=1}^{M} \sum_{y=1}^{N} (x - \bar{x})^p (y - \bar{y})^q f(x, y) \tag{51.2}$$

where $\bar{x} = m_{10}/m_{00}$ and $\bar{y} = m_{01}/m_{00}$. The normalized central moments are denoted by η_{pq}. η_{pq} is defined as $\eta_{pq} = u_{pq}/u_{00}^{\gamma}$, where $\gamma = (p + q)/2 + 1$, $p + q = 2, 3, \cdots$.

A set of seven invariants moments is defined as:

$$\phi_1 = \eta_{20} + \eta_{02} \tag{51.3}$$

$$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \tag{51.4}$$

$$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \tag{51.5}$$

$$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \tag{51.6}$$

$$\begin{aligned}\phi_5 = {} & (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ & + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\end{aligned} \tag{51.7}$$

$$\begin{aligned}\phi_6 = {} & (\eta_{20} - \eta_{02})\left[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] \\ & + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})\end{aligned} \tag{51.8}$$

$$\begin{aligned}\phi_7 = {} & (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ & + (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\end{aligned} \tag{51.9}$$

51.2.2 Hu Moments Similarity Measure

$S_o = \{S_{ok} | k = 1, 2, \cdots 7\}$ denotes the feature vector of query image. $S_i = \{S_{ik} | k = 1, 2, \cdots 7\}$ represents the feature vector of ith image in the image feature database. We compute the dissimilarity value between Hu moments of any two images by Euclidean distance:

$$dist(i) = \sqrt{\sum_{k=0\cdots7} (S_{ok} - S_{ik})^2} \tag{51.10}$$

51.3 Camshift Algorithm and ABCshift Algorithm

51.3.1 Camshift Algorithm

Camshift is an object tracking method which is a modification of Mean Shift tracking method. Mean Shift itself is a robust nonparametric technique for finding the peak in a probability distribution. Camshift can deal with dynamically changing color probability distribution which is taken from the video frames. Because RGB color models are much more sensitive to lighting changes, so this

algorithm converts RGB color space to HSV color space in order to decrease illumination influence to tracking object. They use the HSV color system and using only hue component to make the object's color 1D histogram. This histogram is stored to convert next frames into corresponding probability of the object. The probability distribution image itself is made by back projecting the 1D hue histogram to the hue image of the frame. Camshift is then used to track the object based on this backproject image.

The Camshift algorithm is shown as below:

Step 1: Choose the initial region of interest, which contains the object we want to track.
Step 2: Make a color histogram of that region as the object.
Step 3: Make a probability distribution of the frame using the color histogram. As a remark, in the implementation, they use the histogram back projection method.
Step 4: Based on the probability distribution image, find the center mass of the search window using Mean shift method.
Step 5: Center the search window to the point taken from step 4 and iterate step 4 until convergence.
Step 6: Process the next frame with the search window position from the step 5.

Figure 51.1 shows the CAMSHIFT Algorithm [5].

51.3.2 ABCshift Algorithm

For each frame of an image sequence, the Camshift algorithm looks at pixels which lie within a subset of the image defined by a search window. The Camshift algorithm will be failed when the tracked object moves across regions of background with which it shares significant colors.

ABCshift is an Adaptive Background Camshift algorithm. It can be continuously relearned for every frame by using a background model.

The object location probabilities can now be computed for each pixel using Bayes' law as:

$$P(O|C) = \frac{P(C|O)P(O)}{P(C)} \tag{51.11}$$

where $P(O|C)$ denotes the probability that the pixel represents the tracked object given its color, $P(C|O)$ is the color model learned for the tracked object and $P(O)$ and $P(C)$ are the prior probabilities that the pixel represents object and has the color C respectively.

The denominator of Eq. (51.11) can be expanded as

$$P(C) = P(C|O)P(O) + P(C|B)P(B) \tag{51.12}$$

where $P(B)$ denotes the probability that the pixel represents background.

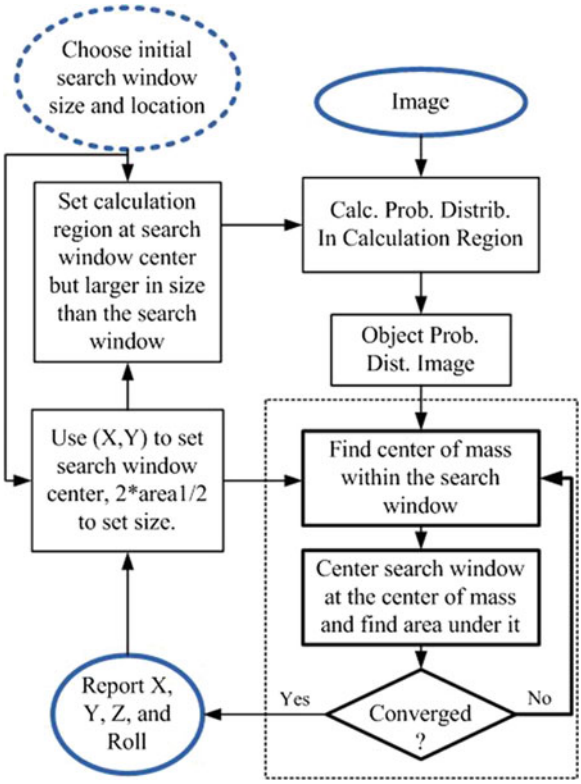


Fig. 51.1 The flowchart of CAMSHIFT algorithm

This paper assigns values to object priors in proportion to their expected image areas. If the search window is resized to be r times bigger than the estimated tracked object area, then $P(O)$ is assigned the value $1/r$ and $P(B)$ is assigned the value $(r - 1)/r$.

When object target enters into a background area which color is similar to a kind of color of the object, $P(C)$ values will increase. But $P(C|O)P(O)$ remains static. So, $P(C|O)$ values diminished.

The tracker will adaptively learn to ignore object colors which are similar to the background and instead tend to focus on those colors of the object which are most dissimilar to whatever background is currently in view.

51.4 The Proposed Object Tracking Approach

The proposed tracking algorithm combines the shape and color features through calculate the similarity between template region and candidate region. The likelihood function is defined as follow:

$$S = \alpha\rho_c + (1 - \alpha)d_{hu} \tag{51.13}$$

Color feather is scale and rotation invariant. It is more robust and stable than shape feature in tracking of colored objects. So, where $\alpha \in [0.5, 1]$, which is defined as the reliability factors for color features. ρ_c is the Bhattacharyya distance of ABCshift algorithm. d_{hu} is the distance of Hu moments. In this paper, the value of the factor α is set according to scene. It will be set smaller if the background includes clutter or the object appearance has geometric change, otherwise it will be set bigger.

The proposed algorithm is summarized as:

(1) Identify an object region in the first image and train the object model, $P(C|O)$. Computes the Hu moments.
(2) Center the search window on the estimated object centroid and resize it to have an area r times greater than the estimated object size.
(3) Learn the color distribution, $P(C)$ by building a histogram of the colors of all pixels within the search window. Calculate the Hu moments of the search window and the distance of Hu moments between template window and search window.
(4) Use Bayes' law, Eq. (51.11) to assign object probabilities, $P(O|C)$, to every pixel in the search window, creating a 2D distribution of object location.
(5) Estimate the new object position as the centroid of this distribution and estimate the new object size (in pixels) as the sum of all pixel probabilities (in pixels) as the sum of all pixel probabilities within the search window.
(6) Compute the Bhattacharyya metric between the distributions, $P(C|O)$ and $P(C)$. If this metric is less than a preset threshold then enlarge the estimated object size by a factor r.
(7) Calculate the likelihood between template windows with search window according to Eq. (51.13).
(8) Repeat steps 2–7 until the object position estimate converges.
(9) Return to step 2 for the next image frame.

51.5 Implementation and Experiments

To check the effectiveness of the proposed approach, we have implemented and tested it on a wide variety of challenging image sequences in different environments and applications. The experiments are performed on a computer which has an Inter(R) Core(TM) i3 processor (3.10 GHZ) with 3.00 GB memory. Our solution and other compared solutions are implemented in VC++ language.

In Fig. 51.2, the human head in a video sequence from http://vision.stanford.edu/~birch/headtracker/seq/ is moving to the left and right very quickly. The illumination on the face also changes. The background scene includes clutter and material of similar color to the face. Object turns around and is spatially occluded. As we can see in Fig. 51.2, at frame 33, the Camshift tracker fails to track because the background color is similar to the face. From frame 64, Camshift is trapped in

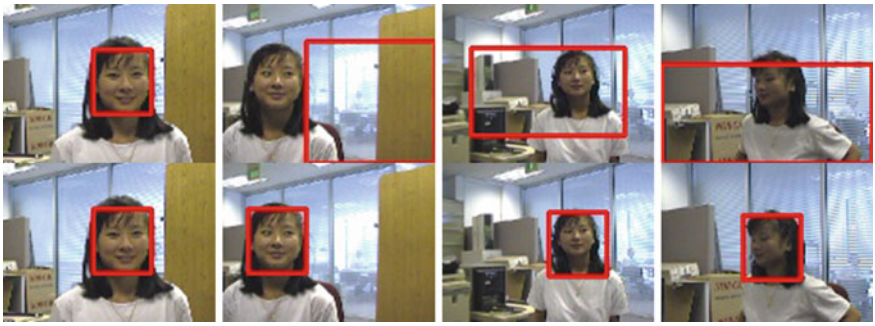


Fig. 51.2 Tracking results of the girl head moving sequence with Camshiftt tracker (*first row*) and our method (*second row*)

Fig. 51.3 The comparison of tracking result: Mean shift (*red rectangle*), our method (*blue rectangle*)

a false region. The tracker fails for most of the remaining frames because the object is distracted by similar color region. In our method, because fuse the shape features and colour features and ABCshift algorithm can adaptively adjust object probability distribution according to the background, the object never loses the target and achieves the most accurate results.

In Fig. 51.3, green rectangle shows the search window of Mean shift algorithm, red rectangle shows the result of Mean shift tracking. Blue rectangle shows the tracking result of proposed algorithm. Initial select object include two parts which have difference colours. When the person who wears a red shift enters to the region which includes red brick walls and doors, the Mean shift tracking lose the tracked person. Adaptive background method can successfully tracks throughout the sequence and is not distracted by red regions of background.

51.6 Conclusion

To increase tracking robustness and accurateness, the paper proposed an object tracking method based on combining Hu moments and ABCshift algorithm. It can not only overcome color features drawback, but also can adaptive background

color changing. Experiments show the proposed method has better performance than Mean shift or Camshift. Similarly to Mean shift and Camshift, proposed algorithm will fail to rapidly move object. In future work, the research attempt to investigate integration with Kalman filter or particle filter and adaptive features fusion mechanism to improve tracking robustness.

Acknowledgments This work was supported by the GRRC program of Gyeonggi Province, Korea [(GRRC SUWON 2012-B3), Development of Multiple Objects Tracking System for Intelligent Surveillance]. This research was also supported in part by Science and technology development planning project (No. 2012YD01052), Shandong province, China. Their support is gratefully acknowledged.

References

1. Yilmaz, A., Javed, O., Shah, M.: Object tracking: a survey. ACM Comput. Surv. **38**(4), 1–45 (2006)
2. Ikizler, N.: Human action recognition with line and flow histograms. In: Proceedings of the Pattern Recognition, ICPR 2008, pp. 1–4 (2006)
3. Qi, W., Zhang, X., et al.: A robust approach for multiple vehicles tracking using layered particle filter. Int. J. Electron. Commun. **65**(7), 609–618 (2011)
4. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE Trans. Pattern Anal. Mach. Intell. **25**(5), 564–577 (2003)
5. Bradski, G.: Computer vision face tracking for use in a perceptual user interface. Intel Technol. J. **2**(2), 13–27 (1998)
6. Isard, M., Blake, A.: CONDENSATION–conditional density propagation for visual tracking. Int. J. Comput. Vis. **29**(1), 5–28 (1998)
7. Wang, Z., et al.: Shape based appearance model for kernel tracking. Image Vis. Comput. **3**(3), 1–13 (2012)
8. Liu, H., Ze, Y., Zha, H., et al.: robust human tracking based on multi-cue integration and mean-shift. Pattern Recogn. Lett. **30**, 827–837 (2009)
9. Hu, K.M.: Visual pattern recognition by moment invariant. IEEE Trans. Inf. Theory **8**(2), 179–187 (1962)
10. Stolkin, R., Florescu, I., Kamberov, G.: An adaptive background model for camshift tracking with a moving camera. In: Proceedings of the 6th International Conference on Advances in Pattern Recognition. World Scientific Press, Kolkata (2007)

Chapter 52
Motion Blur Identification Using Image Statistics for Coded Exposure Photography

Kuihua Huang, Haozhe Liang, Weiya Ren and Jun Zhang

Abstract Coded exposure photography makes the traditional ill-posed motion deblurring problem well posed. However, how to accurately derive the motion blur length confused many researchers because of the non-smooth blur of the coded exposure image. This chapter proposes a novel approach to automatic estimate the motion blur length by analyzing the image statistics for coded exposure photography. The researchers analyze the image power spectrum statistics and observe that the motion blur length has some relations with the residual sums of squares (RSS) of the image power spectrum statistics in a least squares sense. That is, the power spectrum statistics of the obtained deblurred image using the correct estimated blur length corresponds to the lowest value of the RSS. Given an initial blur length, and using the high-speed direct deconvolution approach, researchers can easily find the correct blur length using a global search method. The experimental results demonstrate the validity of the proposed method.

Keywords Coded exposure photography · Blur length · Power spectrum statistics · Motion deblurring

52.1 Introduction

Motion blur is caused by relative motion between the captured scene and camera during the exposure time of images. Many algorithms are proposed to deblur the motion blur images. Recent work on deblurring algorithms [1, 2] has shown excellent results on images corrupted due to camera shake. However, for a fast moving object, whose image has a large blur length, the existing blind methods are

K. Huang (✉) · H. Liang · W. Ren · J. Zhang
Department of System Engineering, National University of Defense Technology, Changsha, China
e-mail: kuihhrgy@gmail.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_52,

almost impotent. Motion blur can be avoided using shorter exposure intervals with high-powered flashes, which is impossible for most cases.

Recently, some novel approaches belonging to computational photography are proposed to solve the motion blur problems [3–5]. Coded exposure photography as one of the examples was firstly proposed by Raskar et al. [6]. Coded exposure opens and closes the camera's shutter with a binary pseudo-random sequence during the exposure time. Through the mode of 'flutter shutter', coded exposure photography modulates the pattern of light and changes the band-limit box filter of conventional camera into a broad-band filter. Thus, the high frequency details are preserved and the ill-posed motion deblurring problem becomes well posed.

Though coded exposure photography has many advantages, it still needs accurate point spread function (PSF) estimation in the process of image deconvolution. For simplicity and without loss of generality, in this chapter we consider the constant velocity motions at the horizontal orientation. In this case, the estimation of PSF is equivalent to the estimation of blur length. Raskar et al. propose to obtain the blur length by the user manually but point out that a new algorithm is required to estimate the correct blur length automatically [6]. Agrawal and Xu [7] use the motion from blur (MFB) approach based on alpha matting to get the blur length. Alpha matting assumes that there exists a high contrast edge in the latent sharp image, which is rigorous for many applications. Tai et al. [5] present a blur estimation technique based on the projective motion blur model. However, it also needs some user interactions and requires the users to have some professional skills in image processing.

This chapter proposes an automatic blur length estimation method based on the analysis of image statistics for coded exposure photography. We compute the image power spectrum statistics and analyze the limitation of the linear power spectrum statistics model. We propose to automatic estimate the blur length directly through searching the lowest value of the residual sums of squares (RSS) of the deblurred image power spectrum statistics in a least squares sense.

52.2 Power Spectrum Statistics of Natural Images

For one-dimensional (1-D) motion blur, the process of blurring can be modeled as the following convolution:

$$g(x,y) = f(x,y) \otimes h(x) + n(x,y), \tag{52.1}$$

where $g(x,y)$, $f(x,y)$ and $h(x)$ are the blurry image, latent sharp image and the blurring kernel (PSF), respectively. $\otimes$ is the convolution operator and $n(x,y)$ is the additive noise.

If we ignore the additive noise, we can model the power spectrum of Eq. (52.1) as:

$$|G| = |FH| = |F||H|. \tag{52.2}$$

Van der Schaaf and van Hateren [8] have shown that the power spectrum of a natural image (without motion blur) falls off with the absolute frequency and they can be related by a power-law distribution. We calculate the power spectrum $|F(u,v)|$ for integer values of u and v using the discrete Fourier transform and parameterize the two-dimensional spatial frequencies in polar coordinates as $|F(\omega,\phi)|$, with $u = \omega\cos\phi$ and $v = \omega\sin\phi$, where u denotes the absolute spatial frequency, and ϕ the orientation. We then average $|F(\omega,\phi)|$ over ϕ, the resulting power spectrum statistics:

$$S_\omega(|F|) = \frac{1}{360}\sum_{\phi=1}^{360}|F(\omega,\phi)| \approx \frac{A}{\omega^\gamma}, \quad \omega = 1,2,\cdots,M, \tag{52.3}$$

where A and γ are constants, and as Ref. [8], we let $M = 127$, because higher frequencies suffer from noise.

We use the logarithm operator to Eq. (3) and let $x = \log(\omega)$, we can get a linear expression as follows:

$$\log(S_\omega(|F|)) = \log(A) - \gamma\log(w) = \log(A) - \gamma x. \tag{52.4}$$

We select five natural images randomly from Berkeley Segmentation Dataset (BSD), as shown in Fig. 52.1a, b shows corresponding power spectrum statistics and the line fitting results using Eq. (4) in a log–log scale. The line fitting results illustrate that the power-law distribution is accurate and robust.

McCloskey et al. [9] propose a linear power spectrum statistics model based on the power-law distribution to estimate the motion information. This model first computes the power spectrum statistics and estimates the parameters A and γ through a line fitting between $\log(S_\omega(|F|))$ and $\log(\omega)$, and then approximates the horizontal linear power spectrum statistics as follows:

$$R_u[|\mathrm{F}|] = E[\frac{1}{V}\sum_{v=0}^{V}|F(u,v)|] \approx \frac{1}{V}\sum_{v=0}^{V}\frac{A}{(u^2+v^2)^{\gamma/2}}. \tag{52.5}$$

Equation (52.5) is an approximate approach, and it is easy to see that the accuracy of the linear power spectrum statistics depends on an accurate estimation of parameters A and γ. So the linear power spectrum statistics model is a two-step approach. Of course, a little error of estimation of the parameters in the first step will be magnified in the second step and lead to a bad result. Figure 52.2a gives the power spectrum statistics of a natural image and the corresponding line fitting result. Figure 52.2b illustrates the linear power spectrum and the fitting line using Eq. (52.5). Although there is only a little error in Fig. 52.2a, the estimated linear power spectrum statistics has a big departure from the real observed data. Therefore, the linear power spectrum statistics model is not so accurate and robust.

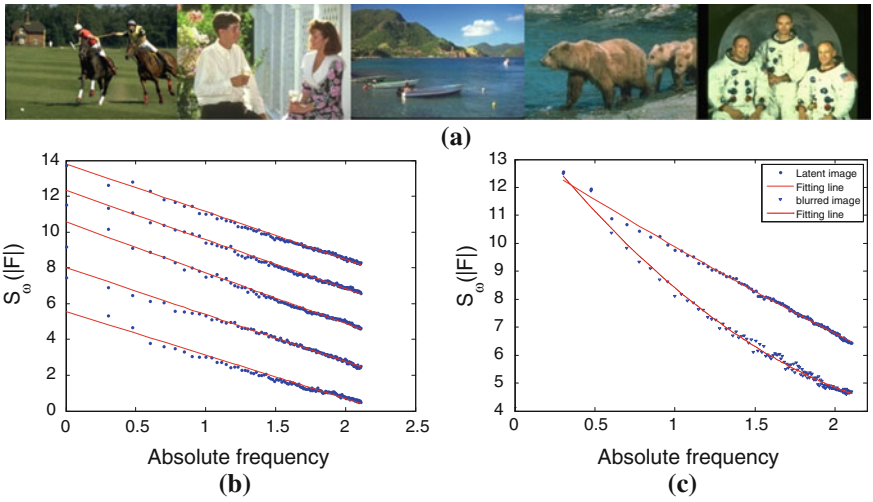


Fig. 52.1 Power spectrum statistics of five natural images. **a** Five natural images from the BSD. **b** Power spectrum statistics versus the absolute frequency in a log–log scale. The lines show the fits of the power-law distribution. For clarity, traces in the plot are shifted –2, –4, –6, –8 log-units. **c** Power spectrum statistics of the blurred image and the corresponding latent sharp image

52.3 Estimate the Blur Length

Other than using the linear power spectrum statistics, we estimate the blur length directly using the power-law distribution in this chapter.

We observe that the data points of the log power spectrum statistics of a blurred image is more disperse than the corresponding latent sharp image, and to a certain extent the lager the blur length, the more disperse the data points are. The difference between the power spectrum statistics of the latent sharp image and the corresponding blurred image is shown in Fig. 52.1c. If using the least square

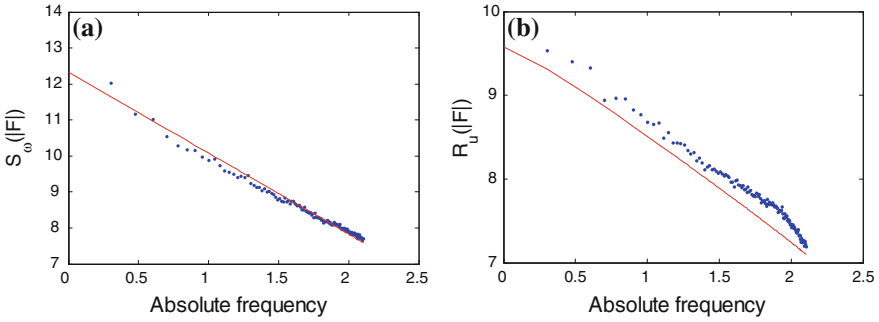


Fig. 52.2 The power law distribution and the corresponding linear power spectrum statistics. **a** Power law distribution of a natural image in a log–log scale. **b** The observerd linear power spectrum statistics (*dot*) and the corresponding line fitting result using Eq. (52.5)

estimator (LSE) to fit the observed data points, we can use the RSS to measure the disperse degree. Assuming the true data points of the log power spectrum statistics of an image are $(\omega_1, S_1), (\omega_2, S_2), \cdots (\omega_n, S_n)$, the fitted data points using the LSE are $(\omega_1, \hat{S}_1), (\omega_2, \hat{S}_2), \cdots (\omega_n, \hat{S}_n)$, the RSS can be expressed as follows:

$$RSS = \sum_{i=1}^{n} \left\| S_i - \hat{S}_i \right\|^2. \tag{52.6}$$

We simulate a coded exposure image with 52 pixel blur length, and use different blur length in the direct deconvolution method [6] to get the deblurred images. In the direct deconvolution method, we set the initial blur length to 40 pixels and increase by 1 pixel in each iterative step. If we set the biggest blur length to 60 pixels, we can get 21 deblurred images. We calculate the log power spectrum statistics of each image and use the LSE to get the fitted values. The RSS of each deblurred image is obtained using Eq. (6) and the results are illustrated in Fig. 52.3a.

From Fig. 52.3a, it can be seen that we get the lowest RSS when using a blur length of 52 pixels in the direct deconvolution method. So the lowest RSS corresponds to the correct blur length. It also can be seen that the bigger the distance between the estimated blur length and the correct blur length, the higher the RSS of the resulted deblurred image is. But this phenomenon is not always true when the estimated blur length is too far from the correct blur length. We use the same simulated 52 pixel blur length image, and set the estimated blur length with the range of 20–80 pixels, after deblurring using the direct deconvolution method, the resulted RSS is shown in Fig. 52.3b.

Although there will be some fluctuation when the estimated blur length is too far from the correct blur length, the lowest RSS still corresponds to the correct blur length.

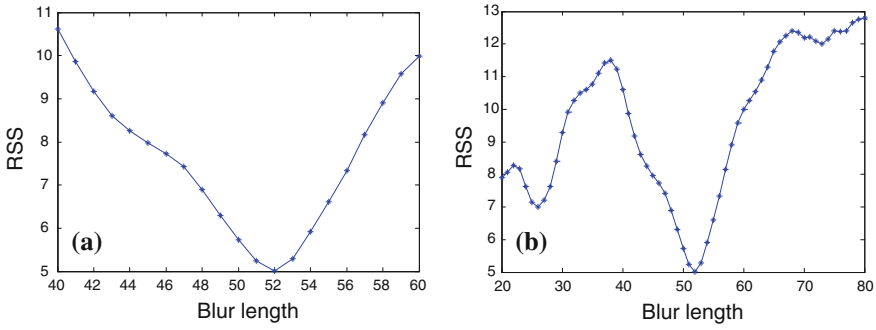


Fig. 52.3 The RSS of the deblurred images of the direct deconvolution method using different blur length

If we assume α is the estimated blur length, then a direct approach for testing if α is a good estimation of blur length would be to check if the RSS of the log power spectrum statistics in a least square sense achieves the lowest value. Given an initial blur length α_0, we can easily get the optimal estimated blur length in terms of RSS using a global optimization method in Matlab Toolbox, i.e. Direct search, Genetic algorithm, or Simulated Annealing, etc.

The main steps of our method are as follows:

1. Give an initial estimated blur length α_0;
2. Obtain the deblurred image I using the direct deconvolution method with the given blur length.
3. Compute the Fourier spectrum $|F(u, v)|$, and parameterize it to the polar coordinates as $|F(\omega, \varphi)|$.
4. Compute the power spectrum statistics of $|F(\omega, \varphi)|$ using Eq. (52.3), and use the LSE to obtain the corresponding fitted value.
5. Compute the RSS using Eq. (52.6).
6. Is the RSS the lowest value? If not, get a new α using the global optimization algorithm and turn to step 2; else, stop and get the optimal estimated blur length.

52.4 Experiments

In the first experiment, we used a real coded exposure image supplied by Raskar et al. [6], as shown in Fig. 52.4a. The blur length of the image had been identified manually and the identified result is 118 pixels. Figure 52.4b is the corresponding deblurred result with the identified blur length using the direct deconvolution method.

The result of Fig. 52.4b demonstrates the identified blur length is accurate. We use our method to automatically estimate the blur length and Fig. 52.4c is the result of RSS. The lowest value of the RSS corresponds to the correct blur length. This result demonstrates the validity of our method.

Fig. 52.4 Validity check of our method with Raskar's image. **a** Coded exposure image supplied by Raskar. **b** The deblurred image with correct blur length. **c** The RSS of deblurred images for different estimated blur length

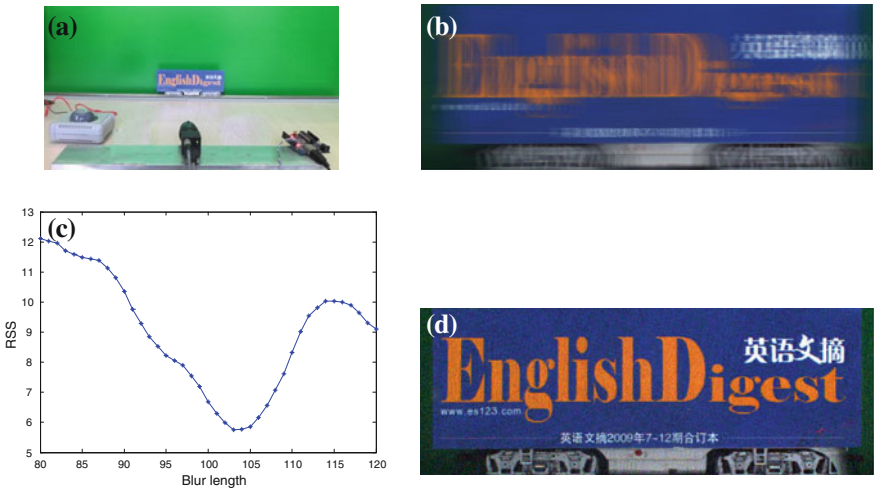


Fig. 52.5 Motion blur length estimation for real coded exposure image and our experimental setup. **a** Experimental setup. **b** Our coded exposure image. **c** The RSS of deblurred images resulted with different estimated blur length. **d** The deblurred image with the estimated blur length of the proposed method

In the second experiment, we used a PointGrey Flea2 camera to capture a toy train moves at uniform speed to get our coded exposure image, and in order to achieve accurate signal synchronization, we used an Arduino Duemilanove board to supply the external trigger pulse. Flea2 works in IEEE DCAM Trigger mode 5 ("Multiple Exposure Pulse Width Mode") which provides coded exposure function. The experimental setup is given in Fig. 52.5a.

Figure 52.5b is the captured coded exposure image, and using our proposed blur length estimation method the obtained result is 103 pixels, as illustrated in Fig. 52.5c. We use this blur length to deblur Fig. 52.5b, the deblurred result is shown in Fig. 52.5d. The deblurred result demonstrates that the proposed blur length estimation method is robust and accurate.

52.5 Conclusion

The researchers present an automatic motion blur length estimation approach based on the image power spectrum statistics. The contribution of the work is that the blur length affects the disperse degree of the image power spectrum statistics is found. Researchers employ the RSS to model the disperse degree and derive the correct blur length which corresponds to the lowest value of RSS. Because the direct deconvolution method is used in the process of image deconvolution, the proposed method is efficient.

References

1. Fergus, R., et al.: Removing camera shake from a single photograph. ACM Trans. Graph. **25**(3), 787–794 (2006)
2. Shan, Q., Jia, J., Agarwala, A.: High-quality motion deburring from a single image. ACM Trans. Graph. **27**(3), 1–10 (2008)
3. Yuan, L., Sun, J., Quan, L., et al.: Image deblurring with blurred/noisy image pairs. In: Proceedings of ACM SIGGRAPH, p. 1 (2007). doi: http://doi.acm.org/10.1145/1275808.1276379
4. Ben-Ezra, M., Nayar, S.K.: Motion-based motion deblurring. IEEE Trans. Pattern Anal. Mach. Intell. **26**(6), 689–698 (2004)
5. Tai, Y.-W., Kong, N., Lin, S., Shin, S.Y.: Coded exposure imaging for projective motion deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 10), pp. 2408–2415, IEEE Press (2010). doi: 10.1109/CVPR.2010.5539935
6. Raskar, R., Agrawal, A., Tumblin, J.: Coded exposure photography: Motion deblurring using a flutterd shutter. ACM Trans. Graph. **25**(3), 795–804 (2006)
7. Agrawal, A., Xu, Y.: Coded exposure deblurring: Optimized codes for PSF estimation and invertibility. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2066–2073 (2009)
8. van der Schaaf, A., van Hateren, J.H.: Modelling the power spectra of natural images: Statistics and information. Vis. Res. **36**, 2759–2770 (1996)
9. McCloskey, S., Ding, Y., Yu, J.: Design and estimation of coded exposure point spread functions. IEEE Trans. Pattern Anal. Mach. Intell. **34**(10), 2071–2077 (2012)

Chapter 53
Medical Images Fusion Using Parameterized Logarithmic Image Processing Model and Wavelet Sub-band Selection Schemes

Bole Chang, Wenbing Fan and Bo Deng

Abstract A novel wavelet sub-band selection scheme for medical image fusion, based on the Parameterized Logarithmic Image Processing (PLIP) model, is presented in this chapter which takes the characteristics of human visual system (HVS) and the spatial distribution of wavelet coefficients into account. The different fusion schemes are applied for the different frequency sub-bands. The visibility weighted average method is selected for coefficients in low-frequency band and a variance based weighted method is selected for coefficients in high-frequency bands. Subsequently, the fused coefficients are processed with consistency verification to guarantee the homogeneity of the fused image. Computer simulations illustrate that the proposed image fusion algorithms with the PLIP model is superior to some existing fusion methods, and can get satisfactory fusion results.

Keywords Image fusion · PLIP model · Wavelet transform · Sub-band selection scheme

53.1 Introduction

In the recent years, the study of medical image fusion attracts much attention including diagnosis, research, and treatment. Image fusion is the combination of multimodality source images [1]. Multimodality medical images mainly include the following images, computed tomography (CT) and magnetic resonance imaging (MRI) images and so on [2]. The aim of image fusion is to integrate complementary as well as redundant information from multiple images to create a fused image output, which should contain a more accurate description of the scene and is more suitable for human visual and further image processing and analysis task.

B. Chang (✉) · W. Fan · B. Deng
Department of Information Engineering,
Zhengzhou University, Zhengzhou, China
e-mail: cbl2388@gmail.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_53,

The existing fusion methods involve mainly Pixel Weighted Average fusion (PWA), Laplacian Pyramid (LP), Discrete Wavelet Transform (DWT) [3], Principal Component Analysis (PCA) [4], Contourlet transform (CT) [5], and Nonsubsampled Contourlet Transform (NSCT) [6]. Research shows that the PWA blurs feature information of image. The LP transform decompose fails to introduce any spatial orientation selectivity in the decomposition process. The PCA do not incorporate aspects of the human visual system in their formulation such as sensitivity to edges at their various scales and undesirable side contrast. The Contourlet transform, based on the Laplacian Pyramid and directional filter, is a kind of multi-scale and multi-direction discrete image transformation, in which the process of multi-scale analysis and direction analysis is successively disposed. But it is a varying shift transformation and ignores the relationship between contourlet coefficients. In order to obtain translation invariance, Cunha AL proposed NSCT, which consists of Nonsubsampled Directional Filter Bank (NSDFB) and Nonsubsampled Pyramid (NSP).The NSCT can extract the edge of the image contour information well. But, it captures image detail weakly, and fails to present the local characteristics of image. The DWT of image signal produces non-redundant image representations and provides better spatial and spectral localization of image information. In the DWT scheme, wavelet coefficient fusion rules directly influence the speed and quality of fusion [7].

The chapter is organized as follows: In Sect. 53.2, the PLIP model and the parameterized logarithmic multi-resolution image decomposition structure are described. And a new image fusion algorithm is proposed in Sect. 53.3. Experimental results and analysis are presented in Sect. 53.4. Finally, the conclusions are given with a short summary in Sect. 53.5.

53.2 Wavelet Decomposition Based on PLIP Model

The original concept and theory of wavelet-based multi-resolution analysis came from Mallat [8]. The wavelet transform is a mathematical tool that can detect local time–frequency features in a signal process.

53.2.1 PLIP Model

The PLIP model was proposed by Karen Panetta, which are appropriate particularly for image enhancement in both the spatial and transform domains. With defining a set of parameterized nonlinear operation to replace image of linear operators, it operates image gray value directly. The arithmetical operations of PLIP model are as follows [9].

$$\text{Gray tone calculation: } g(i,j) = M - f(i,j) \tag{53.1}$$

$$\text{Isomorphic transform: } \varphi(g) = -\lambda(M) \cdot \ln^{\beta}(1 - g/\lambda(M)) \tag{53.2}$$

$$\text{Inverse isomorphic transform: } \varphi^{-1}(g) = \lambda(M)[1 - \exp(-g/\lambda(M))^{1/\beta}] \tag{53.3}$$

where $g(i,j)$ is the same gray tone function and M is the maximum intensity value of input image $f(i,j)$. The parameter $\lambda(M)$ is a linear function of the type $\lambda(M) = AM + B$, with constant parameters A and B. The research shows that visually pleasing images can be got with $\beta = 1$ and $\lambda(M) = 896$.

53.2.2 Image Fusion Based on PLIP Model

The 2D Wavelet decomposition provides a framework in which two dimensional signals are decomposed different level sub-bands, which uses a quadrate mirror set analysis filters, h and g, and synthesis filters, $\tilde{h}$ and $\tilde{g}$ [7]. And the DWT based on the PLIP model (PLIP-DWT) is calculated by making use of the parameterized isomorphic transformation and is defined by the following equations.

$$\text{PLIP - DWT: } W_{\text{PLIP-DWT}}(f) = W_{\text{DWT}}(\varphi(f)) \tag{53.4}$$

$$\text{PLIP - IDWT: } W_{\text{PLIP-IDWT}} = \varphi^{-1}(W_{\text{IDWT}}(W_{\text{PLIP-DWT}}(f))) \tag{53.5}$$

The structures of 2D wavelet decomposition based on the PLIP model analysis and synthesis are shown in Fig. 53.1. The DWT decomposition based on the PLIP

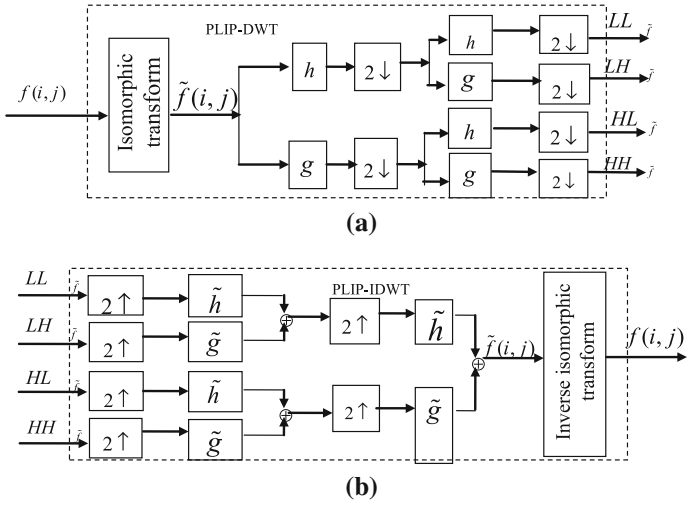


Fig. 53.1 The structures of PLIP 2D wavelet decomposition analysis and synthesis. **a** One stage of 2D PLIP-DWT multi-resolution image decomposition. **b** One stage of 2D PLIP-IDWT multi-resolution image fusion

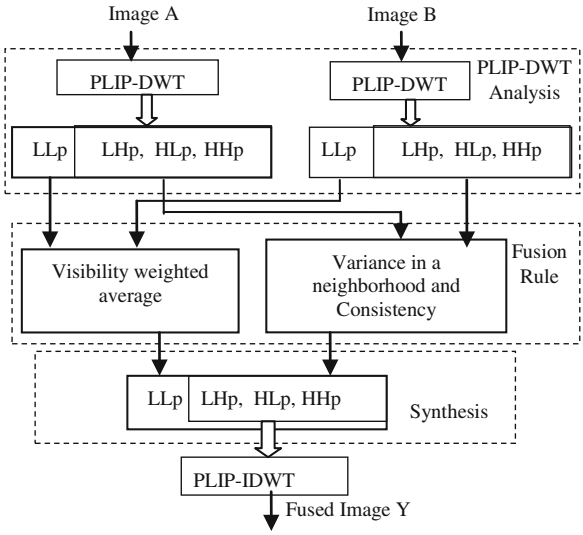


Fig. 53.2 Schematic diagram of the proposed fusion rule

model generates the low-frequency band coefficient $LL\tilde{f}$ and the high-frequency coefficient $HL\tilde{f}$, $LH\tilde{f}$, and $HH\tilde{f}$.

The overall flowchart of the DWT scheme based on the PLIP model can be illustrated in Fig. 53.2.

53.3 The Proposed Fusion Schemes for the Coefficients

The core technology based on the wavelet transform of the image fusion is the wavelet coefficient fusion rules. Therefore, a new type of wavelet coefficient fusion rules is proposed.

53.3.1 Weighted Average Method in Low Frequency Sub-band

For the low-frequency band, a fusion scheme selects the weighted average method, which is based on HVS to the sensitive degree of image [10]. Two original images, $A$and$B$, and their fused image Y are introduced. And the multi-scale decompositions of the original and fused images are denoted by C_{A}, C_{B} , and C_{Y}. Let $p = (\mathrm{m}, \mathrm{n}, \mathrm{k}, \mathrm{l})$ indicates the index corresponding to a particular coefficient. $C_{\mathrm{A}}(p)$ denotes the decomposition value of the corresponding coefficient at the

position (m, n) with decomposition level k and frequency band l (1 = LL, LH, HL, and HH). The visibility of wavelet coefficients is defined as:

$$\omega(p) = \frac{1}{N^2} \sum_{(i,j) \in W_N} \Upsilon(\mu(p)) \cdot (|C(m+i, n+j, k, l) - \mu(p)|/\mu(p)) \tag{53.6}$$

$$\Upsilon(\mu(p)) = (1/\mu(p))^{\alpha} \tag{53.7}$$

$$\mu(p) = (1/N^2) \sum_{(i,j) \in W_N} C(m+i, n+j, k, l) \tag{53.8}$$

where W_N is a $N * N$ block, $\Upsilon(\mu(p))$ is the weighting factor, α is a constant by perceptual experiment, and its range is from 0.6 to 0.7 [10]. After calculating the visibility of all the coefficients in the low-frequency band, the corresponding coefficients with higher magnitude of visibility are then chosen into the fused image as follows:

$$C_Y(p) = (\omega_A(p) \cdot C_A(p) + \omega_B(p) \cdot C_B(p))/(\omega_A(p) + \omega_B(p)) \tag{53.9}$$

53.3.2 Fusion Scheme in High Frequency Sub-bands

A scheme is proposed by computing the variance in a neighborhood to select the high-frequency coefficients [10]. The neighborhood variance of wavelet coefficient of input image is $\sigma_A(p)$, and the covariance is $\sigma_{AB}(p)$. The $\mu_A(p)$ denotes mean value. The procedure can be formulated as follow.

$$\sigma_A(p) = (1/N^2) \sum_{(i,j) \in W_N} (C_A(m+i, n+j, k, l) - \mu_A(p))^2 \tag{53.10}$$

$$\mu_A(p) = (1/N^2) \sum_{(i,j) \in W_N} C_A(m+i, n+j, k, l) \tag{53.11}$$

$$\sigma_{AB}(p) = (1/N^2) \sum_{(i,j) \in W_N} (C_A(m+i, n+j, k, l) \cdot C_B(m+i, n+j, k, l) - \mu_{AB}(p))^2 \tag{53.12}$$

$$\mu_{AB}(p) = (1/N^2) \sum_{(i,j) \in W_N} C_A(m+i, n+j, k, l) \cdot C_B(m+i, n+j, k, l) \tag{53.13}$$

The local matching coefficient measure of each sub-band between source images is given as:

$$M_{AB}(p) = 2\sigma_{AB}(p)/(\sigma_A^2(p) + \sigma_B^2(p)) \tag{53.14}$$

Comparing the matching measure to a threshold T determines if detail coefficients are to be combined by simple selection or by weighted averaging.

$$\delta(p) = \begin{cases} 1 - (M_{AB}(p) - T)/2(1 - T), & M_{AB}(p) > T, \sigma_A(p) > \sigma_B(p) \\ (M_{AB}(p) - T)/2(1 - T), & M_{AB}(p) > T, \sigma_A(p) < \sigma_B(p) \\ 1 & M_{AB}(p) < T, \sigma_A(p) > \sigma_B(p) \\ 0 & M_{AB}(p) < T, \sigma_A(p) < \sigma_B(p) \end{cases} \tag{53.15}$$

where $\delta(p)$ indicts the factor of multiplicative weight averaging. The fused coefficients are calculated using the following formulation.

$$C_Y(p) = \delta(p) \cdot C_A(p) + (1 - \delta(p)) \cdot C_B(p) \tag{53.16}$$

53.3.3 Consistency Verification

The proposed method cannot guarantee the homogeneity in the resultant fused image, especially for the high frequency sub-bands [10]. Therefore, a consistency verification scheme can ensure that the dominant features are incorporated into the fused image. The idea is likely to be a 3- by-3 median filter. And a window-based verification is applied to the fused high frequency coefficients. In the implementation, this rule is applied to a binary decision map, followed by the application of a median filter. This process can be formulated as follows:

$$C_A^m(p) = \underset{(i,j) \in W_N}{\text{median}}(|C_A(m + i, n + j, k, l)|) \tag{53.17}$$

$$\beta(p) = \begin{cases} 1, & C_A^m(p) > C_B^m(p) \\ 0, & \text{otherwise} \end{cases} \tag{53.18}$$

$$\beta'(p) = \sum_{(i,j) \in W_N} \beta(m + i, n + j, k, l) \tag{53.19}$$

$$\beta^*(p) = \begin{cases} 1, & \beta'(p) > N \\ 0, & \text{otherwise} \end{cases} \tag{53.20}$$

Refer to (53.16), the fused coefficients in the high frequency sub-bands are modified by:

$$C_Y^*(p) = C_Y(p) + \lambda \cdot [\beta^*(p) \cdot C_A(p) + (1 - \beta^*(p)) \cdot C_B(p)] \tag{53.22}$$

where λ is a constant by perceptual experiment?

53.4 Experimental Results and Analysis

In this section, the performance of the proposed method is compared with those of PWA, LP, PCA, Contourlet transform, NSCT and DWT. The first experiment is tested on MRI and MRA images, as shown in Fig. 53.3. And the information entropy (IE), cross entropy (CE), mutual information (MI) and fusion symmetry (FS) of the fused image are applied to evaluate the performance of the above fusion method [11]. The performances of the different methods are listed in Table 53.1.

Comparing the experimental results, the proposed method of fusion visibility is the best. As is shown from Table 53.1, the IE of the proposed method is higher than results of other methods, and the values of CE and FS are lower than the results of other methods. And the proposed fusion scheme removes the blurring information and the fused image has more resolution than other methods.

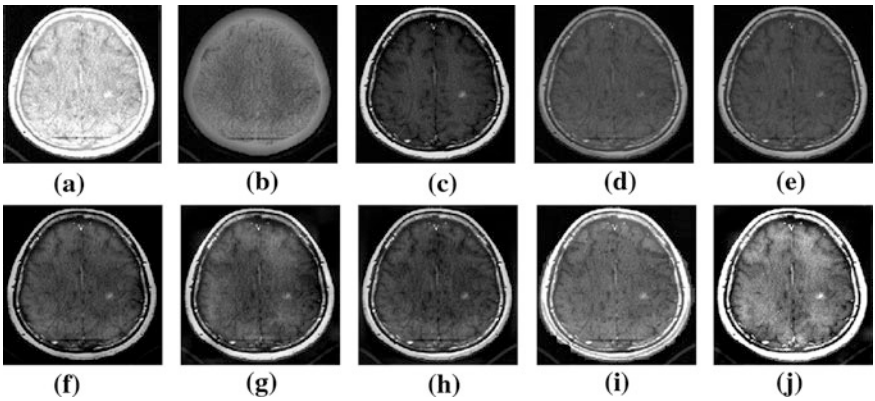


Fig. 53.3 Fusion results of the MRI and MRA images with different methods. **a** Reference image; **b** Original MRI image; **c** Original MRA image; **d** Fused image by PWA method; **e** Fused image by PCA method; **f** Fused image by LP method; **g** Fused image by DWT method; **h** Fused image by Contourlet transform method; **i** Fused image by NSCT method; **j** Fused image by the proposed method

Table 53.1 Quantitative evaluation results of different fusion methods in Fig. 53.3

Fusion methods	IE	CE	MI	FS
PWA	6.3195	1.4806	3.2325	0.0054
PCA	6.4163	1.3954	2.9406	0.0199
LP	5.8181	1.3732	2.2674	0.1035
DWT	6.0208	1.5649	2.6369	0.1623
Contourlet transform	6.4335	1.6156	2.7337	0.1845
NSCT	6.4862	1.3823	2.9891	0.0183
Proposed method	6.4919	1.3305	2.4366	0.0013

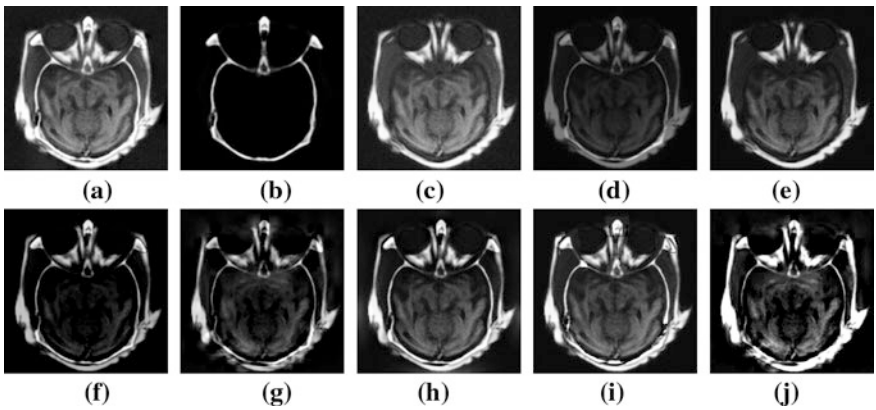


Fig. 53.4 Fusion results of the "brain" CT and MRI images with different methods. **a** Reference image; **b** Original CT image; **c** Original MRI image; **d** Fused image by PWA method; **e** Fused image by PCA method; **f** Fused image by LP method; **g** Fused image by DWT method; **h** Fused image by Contourlet transform method; **i** Fused image by NSCT method; **j** Fused image by the proposed method

Table 53.2 Quantitative evaluation results of different fusion methods in Fig. 53.4

Fusion methods	IE	CE	MI	FS
PWA	5.9152	0.7407	2.5326	0.1015
PCA	6.5814	0.4396	2.2358	0.0797
LP	3.8037	1.3318	2.0654	0.7059
DWT	5.9748	0.6800	1.6721	0.4652
Contourlet transform	6.9694	0.1332	1.9381	0.0733
NSCT	6.7896	0.1939	2.4606	0.0984
Proposed method	5.7291	0.1206	2.6832	0.0636

For further comparing, the proposed method is tested on CT and MRI images as shown in Fig. 53.4. The relevant performances of the different methods are listed in Table 53.2.

Comparing the experimental results in Fig. 53.4, the proposed method of fusion visibility is the best. As is shown from Table 53.2, the MI of the proposed method is higher than results of other methods, and the values of CE and FS are lower than the results of other methods. And the proposed fusion scheme contains more image information and the fused image has more resolution than other methods.

53.5 Conclusion

A novel wavelet-based sub-band selection approach with PLIP model is presented for medical image fusion—it consists of four steps: isomorphic transform, wavelet decomposition, coefficients fusion, wavelet synthesis, and inverse isomorphic

transform. The experimental results show that the proposed fusion method outperforms some existing fusion methods and it can get satisfactory fusion results.

References

1. Piella, G.: A general framework for multi-resolution image fusion: From pixels to regions. Inf. Fusion **4**, 259–280 (2003)
2. Yang, Y., Park, D.S., Huang, S., Rao, N.: Fusion of CT and MR images using an improved wavelet based method. J. X-Ray Sci. Technol. **18**(10), 157–170 (2010)
3. Wang, Y., Lohmann, B.: Multi-sensor image fusion: Concept, method and applications. Technical Report, Institute of Automatic Technology, University of Bremen, Bremen, Germany (2000)
4. Pradhan, P.S., et al.: Estimation of the number of decomposition levels for a wavelet-based multi-resolution multi-sensor image fusion. IEEE Trans. Geosci. Remote Sens. **44**(12), 3674–3686 (2006)
5. Do, M.N., Vetterli, M.: The contourlet transform: An efficient directional multi-resolution image representation. IEEE Trans. Image Process. **14**(12), 2091–2106 (2005)
6. Cunha, A.L., Zhou, J.P., Do, M.N.: The nonsubsampled contourlet transform: Theory, design, and applications. IEEE Trans. Image Process. **15**(10), 3089–3101 (2006)
7. Cheng, S.L., He, J.M., Lv, Z.W.: Medical image of PET/CT weighted fusion based on wavelet transform. In: iCBBE'08, pp. 2523–2525 (2008)
8. Mallat, S.G.: A theory for multi-resolution signal decomposition: The wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. **11**(7), 674–693 (1989)
9. Nercessian, S.C., Panetta, K.A., et al.: Multi-resolution decomposition schemes using the PLIP with application to image fusion. EURASIP J. Adv. Signal Process. **2011**(515084), 17p (2011)
10. Huang, J.W., Yun, Q.S., Dai, X.H.: A segmentation-based image coding algorithm using the features of human vision system. J. Image Graph. **4**(5), 400–404 (1999)
11. Shi, W.Z., et al.: Wavelet based image fusion and quality assessment. Int. J. Appl. Earth Obs. Geoinf. **6**(3–4), 241–251 (2005)

Chapter 54
Insect Extraction Based on the Improved Color Channel Comparison Method

Yan Yang, Sa Liu, Xiaodong Zhu, Shibin Lian, Huaiwei Wang and Tingyu Yan

Abstract Color Channel Comparison Method is an effective method to transform color images into gray ones. This method can enhance pests' characteristics and remove the background to some extent. However, some interfering background cannot be removed. In order to solve this problem, an improvement on Color Channel Comparison Method is realized in this paper. Comparisons between gray brightness and a threshold value determine whether the pixel is the interfering background. And the threshold value is determined according to the brightness of the image. Empirical results show that the interfering background in black or white pests' photo is effectively cleared, and black or white pests can be more effectively separated from the colored background by using the improved method. The improved color channel comparison method can effectively solve the interfering background problem of Color Channel Comparison Method.

Keywords Fruit tree pest identification · Color channel comparison method · Color-to-gray transform

54.1 Introduction

The pest, enemy of fruit growth, will cause significant losses and serious harm to the quality of the fruit, and incalculable damage to growers. Therefore, pest forecasting, which is particularly important for the effective prevention and

Y. Yang · S. Liu (✉) · X. Zhu · S. Lian · H. Wang
Department of Computer Science and Information, Beijing University of Agriculture, Beijing, China
e-mail: liusa@bac.edu.cn

T. Yan
Ministry of Basic Teaching, Beijing University of Agriculture, Beijing, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_54,

treatment, and computer vision technology used in pest identification become the hot spot of the current research [1].

Generally speaking, a color image needs to be transformed into a gray one when the computer vision technique is used to recognize the insect pests. The gray image having the brightness information without having the color information, the color-to-grey transform is one of the important steps of the image preprocessing, and therefore, the results directly affect the subsequent processing of the image [2–4].

The maximal value extraction method [5], the mean value extraction method [5], the probability coefficient extraction [5] method and the HLS brightness extraction method [6] are widely used to transform color images into the gray images. But, these methods does not consider the difference in color of black or white insect pests and background, and each pixel in image is transformed by one formula, therefore, these method can not effectively enhance pests' characteristics, and remove the background [7].

Recently, Fan etc. proposed color channel comparison method [7], this method is a more effective method to transform color images into gray ones than other methods. According to the difference in color of white or black insect pests and actual image backgrounds, Color Channel Comparison Method can enhance pests' characteristics, remove the background, but it also leaves some interfering background. In this paper, Improvement to the Color Channel Comparison Method is used to transform color images into gray ones, and the threshold value is determined according to the brightness of the image, and black or white pests can be more effectively separated from the colored background by using the improved method.

54.2 Color Channel Comparison Method

According to the RGB color model [8, 9], in RGB image, each of the RGB color pixel is represented by the values of the R, G, B, and different colors is composed by different values of R, G and B. The RGB color model is usually expressed as a unit cube, as shown in Fig. 54.1.

In Fig. 54.1, R, G and B are located at the three corners: Black at the origin, white is located in the corner furthest away from the origin. Equal to the amount of the respective elements on the main diagonal of the cube, generated from dark to bright white, that is gray. Different colors are located in or within the cube.

For color images, the three color channels is relatively large difference between the RGB, for example, the RGB value of pure red is (255, 0, 0), the difference of R channel and G, B channel is 255. RGB channel values of pure white or pure black are equal, the difference is zero. The smaller difference of the RGB channels, the closer to RGB model cube diagonal.

According to the RGB color model, color channel comparison method for white and black pests were two sets of processing formula:

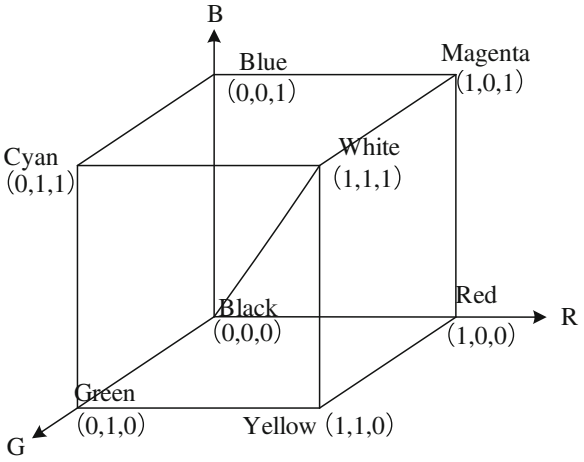


Fig. 54.1 The model of RGB

For white insect pests, if the difference between the highest (Max(R, G, B)) and lowest (Min(R, G, B)) values of R, G and B in a pixel is larger than 20, then let the gray value of the pixel equal to one fifth of the Min(R, G, B), otherwise let the gray value equal to five times of the Max(R, G, B). For the white insect pests, the formula is expressed as:

$$\begin{aligned} L &= \frac{\text{Min}(R,G,B)}{5} \text{ if } \text{Max}(R,G,B) - \text{Min}(R,G,B) > 20 \\ L &= 5 \times \text{Max}(R,G,B) \text{ if } \text{Max}(R,G,B) - \text{Min}(R,G,B) < 20 \end{aligned} \tag{54.1}$$

For black insect pests, if the difference between the highest and lowest values of R, G and B in a pixel is larger than 20, then let the gray value of the pixel equal to five times of the Max(R,G,B), otherwise then let the gray value equal to one fifth of the Min(R,G,B). For the black insect pests, the formula is expressed as:

$$\begin{aligned} L &= 5 \times \text{Max}(R,G,B) \text{ if } \text{Max}(R,G,B) - \text{Min}(R,G,B) > 20 \\ L &= \frac{\text{Min}(R,G,B)}{5} \text{ if } \text{Max}(R,G,B) - \text{Min}(R,G,B) < 20 \end{aligned} \tag{54.2}$$

As it is well known, the recognition of the brightness of the human eyes ranges from 10 to 20. Furthermore, in the Multithreshold model of the color and gray images, the lowest threshold is about 20 [10]. Thus the threshold in Eqs. (54.1) and (54.2) is considered to be 20.

By the difference in brightness of RGB channels in the pixel, color channel comparison method can enhance brightness contrast between color and gray. In Ref. [7], the test result shows that this method is a more effective method to transform color images into gray ones than the maximal value extraction method [5], the mean value extraction method [5], the probability coefficient extraction [5] method and the HLS brightness extraction method [5].

For white insect pests, if the difference between the highest and lowest values of R, G and B in a pixel is larger than 20, then the color of pixel is the color of background, so brightness needs to be reduced. Otherwise, the color of pixel is grey, brightness is enhanced. As a result, not only the brightness of the white is enhanced, the brightness of dark gray, like branches and gray spots, also be enhanced. For black insect pests, if the difference between the highest and lowest values of R, G and B in a pixel is larger than 20, then the color of pixel is the color of background, so brightness needs to be enhanced. Otherwise, the color of pixel is grey, brightness is reduced. As a result, not only the brightness of the black is reduced, the brightness of bright gray, such as heaven and bright spots, also be enhanced.

54.3 Improvement to the Color Channel Comparison Method

Through experiments, for white insect pests, if the difference between the highest (Max(R,G,B)) and lowest (Min(R,G,B)) values of R, G and B in a pixel is larger than 20, then the color of pixel is the color of background, in order to highlight the white pests, need to reduce the brightness of the background to enhance contrast, then let the gray value of the pixel equal to one fifth of the Min(R,G,B). If the difference between Max(R, G, B) and Min(R, G, B) in a pixel is less than 20, the pixel may be white pest, there may be interference background, like branches and gray spots. The brightness of white pest is much larger than the brightness of interference background. In image, the average value of each pixel's Max(R,G,B) is named avgMax, if the Max(R,G,B) in a pixel is lower than avgMax, this pixel is interference background, the brightness needs to be reduced, so, let the gray value of the pixel equal to one fifth of the Min(R,G,B). Otherwise, let the gray value equal to five times of the Max(R, G, B). For the white insect pests, the formula is expressed as:

$$\begin{aligned} L &= \frac{\text{Min}(R,G,B)}{5} \quad \text{if } \text{Max}(R,G,B) - \text{Min}(R,G,B) > 20 \\ L &= \frac{\text{Min}(R,G,B)}{5} \quad \text{if } \text{Max}(R,G,B) - \text{Min}(R,G,B) \leq 20 \& \ \text{Max}(R,G,B) < avgMax \\ L &= 5 \times \text{Max}(R,G,B) \ \text{if } \text{Max}(R,G,B) - \text{Min}(R,G,B) \leq 20 \& \ \text{Max}(R,G,B) \geq avgMax. \end{aligned} \tag{54.3}$$

For black insect pests, if the difference between Max(R, G, B) and Min(R, G, B) in a pixel is larger than 20, then the color of pixel is the color of background, in order to highlight the low brightness of the black pests, need to increase the brightness of the background to enhance contrast, then let the gray value of the pixel equal to five times of the Max(R, G, B). If the difference between Max(R, G, B) and

Min(R, G, B) in a pixel is less than 20, the pixel may be black pest, there may be interference background, like heaven and bright spots. The brightness of black pest is much lower than the brightness of interference background. In image, the average value of each pixel's Min(R, G, B) is named avgMin, if the Min(R,G,B) in a pixel is larger than avgMin, this pixel is interference background, the brightness needs to be increased, so, let the gray value of the pixel equal to five times of the Max(R,G,B). Otherwise let the gray value equal to five fifth of the Min(R, G, B). For the black insect pests, the formula is expressed as:

$$\begin{aligned} L &= 5 \times \mathrm{Max}(R,G,B) \ \text{ if } \ \mathrm{Max}(R,G,B) - \mathrm{Min}(R,G,B) > 20 \\ L &= 5 \times \mathrm{Max}(R,G,B) \ \text{ if } \ \mathrm{Max}(R,G,B) - \mathrm{Min}(R,G,B) \leq 20\& \\ &\quad \mathrm{Min}(R,G,B) > avgMin \\ L &= \frac{\mathrm{Min}(R,G,B)}{5} \ \text{ if } \ \mathrm{Max}(R,G,B) - \mathrm{Min}(R,G,B) \leq 20\& \\ &\quad \mathrm{Min}(R,G,B) \leq avgMin \end{aligned} \tag{54.4}$$

Using Eqs. (54.3) and (54.4), the white and black insect pests in gray images are enhanced relative to the background regions. As a result, the white and black insect pests are effectively recognized and extracted from the actual background images.

54.4 Results and Discussion

In this Section, the color channel comparison method and the improved method will be used to treat the actual pictures in which the white and black insect pests are shown. Firstly, the actual color pictures are transformed into the gray images. Then the gray images are transformed into the binary images by using Otsu's theory [11]. Their treatment effects are compared with each other, testing the advantage of improvement of the color channel comparison method.

Figure 54.2a shows a typical picture of Cnidocampa flavescens chrysalis. The colors of C. flavescens chrysalis are approximately white. Figure 54.2b, c shows a typical picture of Red Striped Longicorn and Drosicha Kuwana. The Red Striped Longicorn and Drosicha Kuwana are near black. In Fig. 54.1, interference background is marked as red circles, which is cleared by improved color channel comparison method.

Figure 54.3a shows the binary images of Fig. 54.2a, which was treated by the color channel comparison method related to Eq. (54.1). Figure 54.3b, c shows the binary images of Fig. 54.2b, c, which was treated by the color channel comparison method related to Eq. (54.2). In Fig. 54.3, interference background is marked as red circles, which is cleared by improved color channel comparison method.

In Fig. 54.3a, according to Eq. (54.1), the brightness L of the white chrysalis is equal to 5 × Max(R,G,B), i.e., is high because the R, G and B values are almost

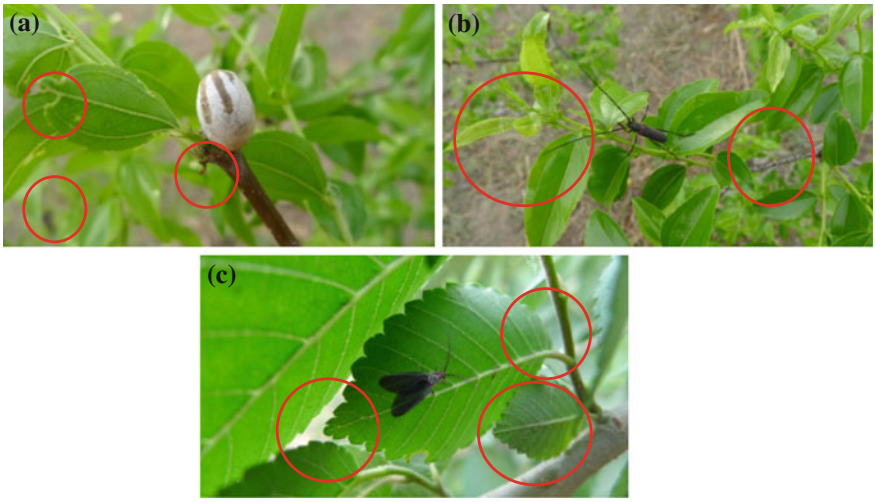


Fig. 54.2 Typical photo of *white* and *black* insect pests

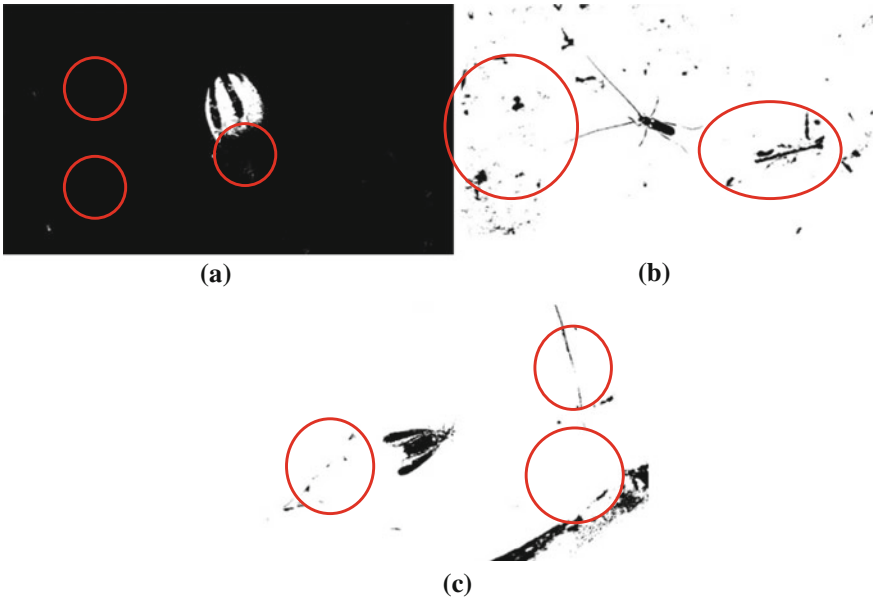


Fig. 54.3 The binary images of Fig. 54.1 treated by the color channel comparison method

same, i.e., Max(R,G,B) − Min(R,G,B) < 20. The brightness L of dark gray in red circle is equal to 5 × Max(R,G,B), i.e., is high because the R, G and B values are almost same, too. i.e., Max(R,G,B) − Min(R,G,B) < 20. The brightness L of

other background equals to Min(R,G,B)/5, i.e., is low because the background mainly is color, i.e., Max(R,G,B) − Min(R,G,B) > 20.

In Fig. 54.3b, c, according to Eq. (54.2), the brightness L of the black insect pest is equal to Min(R,G,B)/5, i.e., is low because the R, G and B values are almost same, i.e., Max(R,G,B) − Min(R,G,B) < 20. The brightness L of bright gray in red circle is equal to/5, i.e., is high because the R, G and B values are almost same, too. i.e., Max(R, G, B) − Min(R, G, B) < 20. The brightness L of the background equals to 5 × Max(R, G, B), i.e., is high because the background mainly is green or pale brown, i.e., Max(R, G, B) − Min(R, G, B) > 20.

Figure 54.4a shows the binary images of Fig. 54.2a, which was treated by improved the color channel comparison method related to Eq. (54.3). Figure 54.3b, c shows the binary images of Fig. 54.2b, c, which was treated by the color channel comparison method related to Eq. (54.4).

According to Eq. (54.3), the brightness L of the most background equals to Min(R,G,B)/5, i.e., is low because the background mainly is color, i.e., Max(R,G,B) − Min(R,G,B) > 20.the brightness L of the white chrysalis is equal to 5 × Max(R,G,B), i.e., is high because the R, G and B values are almost same and the Max(R,G,B) in a pixel is larger than avgMax, i.e., Max(R,G,B)–Min(R,G,B) ≤ 20 and Max(R,G,B) ≥ avgMax. The brightness L of Interference background in red circles, equals to Min(R,G,B)/5, i.e., is low because the R, G and B values are almost same and the Max(R,G,B) in a pixel is smaller than avgMax, i.e., Max(R,G,B) − Min(R,G,B) ≤ 20 and Max(R,G,B) < avgMax.

According to Eq. (54.4), the brightness L of the most background equals to 5 × Max(R,G,B), i.e., is high because the background mainly is color, i.e., Max(R,G,B) − Min(R,G,B) > 20.the brightness L of the Red Striped Longicorn and Drosicha Kuwana is equal to Min(R,G,B)/n, i.e., is low because the R, G and B values are almost same and the Min(R,G,B) in a pixel is smaller than avgMin, i.e., Max(R,G,B) − Min(R,G,B) ≤ 20 and Min(R,G,B) < avgMin. The brightness L of interference background in red circles, equals to Min(R,G,B)/5, i.e., is low because the R, G and B values are almost same and the Min(R,G,B) in a pixel is larger than avgMin, i.e., Max(R,G,B) − Min(R,G,B) ≤ 20 and Min(R,G,B) ≥ avgMin.

Comparing Fig. 54.3 with Fig. 54.4, the white and black insect pests in gray images are enhanced relative to the background regions. Interference background marked with red circles is effectively cleared by improved color channel comparison method. The white and black insect pests are effectively recognized and extracted from the actual background images.

54.5 Conclusion

According to the difference in color of white or black insect pests and actual image backgrounds, Color Channel Comparison Method can enhance pests' characteristics, removing the background. However, it also leaves some interfering

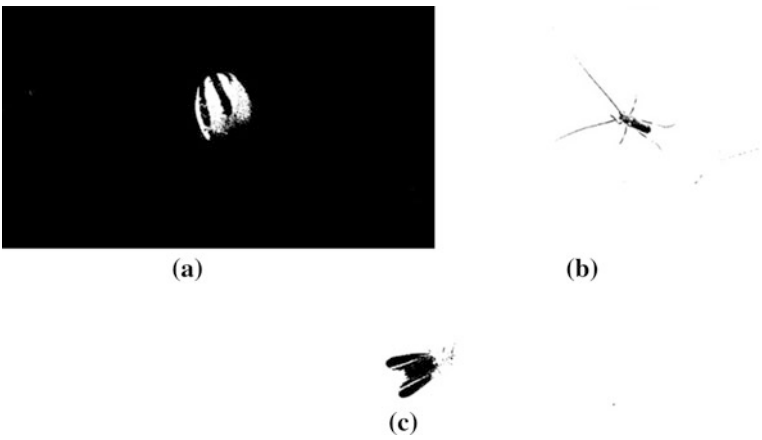


Fig. 54.4 The binary images of Fig. 54.1 treated by improved the color channel comparison method

background. This paper, aiming to the problem of the Color Channel Comparison Method, proposes the improvement on the Color Channel Comparison Method by using the brightness of the image. Experiments show that the white and black insect pests in gray images are enhanced relative to the background regions. Interference background can effectively cleared by improved color channel comparison method. The white and black insect pests are effectively recognized and extracted from the actual background images.

Acknowledgments This work is jointly supported by the Scientific Research Fund of Beijing Municipal Commission of Education Project (No. KM201110020013). The samples are provided by Professor Wang Jinzhong at Beijing University of Agriculture.

References

1. Liu, S., Yang, Y., et al.: Study on Recogniton of Jujube Insect by Computer Vision, ICIST (2012)
2. Guan, Z., Yao, Q., et al.: Application of digital image processing technology in recognizing the diseases, pests, and weeds from crops. Scientia Agricultura Sinica **42**(7), 2349–2358 (2009)
3. Granitoo, P.M., Navone, H.D., et al.: Weed seeds identification by machine vision. Comput. Electron. Agric. **33**, 91–103 (2002)
4. Aitkenhead, M.J., Dalgetty, I.A., et al.: Weed and crop discrimination using image analysis and artifical intelligence methods. Comput. Electron. Agric. **39**, 157–171 (2003)
5. Li, Z., Li, P., et al.: Three programming methods of gray processing based on GDI+. Comput. Technol. Dev. **19**(7), 73–75 (2009)
6. Liu, Q., Jiang, T.: A study of translation arithmetic between color image and grey image. J. Wuhan Univ. Technol. **27**(3), 344–346 (2003)
7. Fan, X., Liu, S., et al.: Study on extraction of insect by computer vision. BMEI (2012)

8. Zhao, Y., Wang, K., et al.: Research of maize leaf disease identifying system based image recognition. Scientia Agricultura Sinica **40**(4), 698–703 (2007)
9. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn, pp. 228–229. Publishing House of Electronics Industry, Beijing (2003)
10. Papamarkos, N., Strouthopoulos, C., Andreadis, I.: Multithresholding of color and gray-level images through a neural network technique. Image Vis. Comput. **18**, 213–222 (2000)
11. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–69 (1979)

Chapter 55
Combining Steerable Pyramid and Gaussian Mixture Models for Multi-Modal Remote Sensing Image Registration

Peng Ye and Fang Liu

Abstract Multi-modal remote sensing image registration is to align images acquired by different sensors and modalities. It is the fundamental step for following image analysis. Previous multi-resolution methods use spatial pyramids to achieve hierarchical registration with little consideration of the characteristics of pyramid transforms or robust point set registration methods after feature detection. Targeting at both problems, this paper proposes a novel image registration method by combining steerable pyramid and Gaussian mixture models. Steerable pyramid has been proved to be shift-invariant and outperforms traditional pyramid transform. Point set registration methods using Gaussian mixture model has been lately proposed and proved to be more robust and accurate than traditional point set registration methods. Experiments on real multi-modal remote sensing image pair demonstrate the feasibility of proposed method.

Keywords Steerable pyramid · Gaussian mixture models · Multi-modal image registration

55.1 Introduction

Multi-modal remote sensing image registration is the process to align images taken by different sensors and modalities. It provides insight into the analysis of the target otherwise cannot. For example, optical images usually have better resolution and understandable image features, but are greatly influenced by the imaging condition like illumination and clouds. Unlike passive imaging scheme of optical images, the active image mechanism makes SAR images rarely affected by the clouds or illuminations. SAR image are mainly determined by the reflection

P. Ye (✉) · F. Liu
ATR Lab, National Universtiy of Defense Technology, Changsha, China
e-mail: ye.peng.email@gmail.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_55,

characteristics of the target. However, SAR images usually suffer from speckle noise and have worse resolutions than optical image, making them more difficult to process. For better knowing of target, it is desirable to combine images of both modalities for further analysis. And multi-modal image registration is the crucial and fundamental step.

Remote sensing images usually adopt multi-resolution strategy to save the computation time and have a better convergence optimization. Previous multi-resolution methods used pyramid transforms like Gaussian pyramids or Laplace pyramids. Unlike steerable pyramid, these pyramid transforms have been proven to be shift-variant and have limited ability of extension [1]. Furthermore, as the basics of steerable pyramid, steerable filters provide a well theory-defined way for feature extraction, making steerable pyramid capable of being extended to detect feature [2]. The benefit of steerable pyramid is that the shift-invariant multi-resolution transformation and feature detection could be done at the same time.

For multi-modal image pair, it is the structural layout features that remain relatively unchanged and are the foundation for feature-based registration methods. It is therefore needed a point set registration method which preserves the structure layout of features and at the same time tolerates high amount of outliers. Traditional multi-resolution image registration methods used features extracted from different resolution as a scattered point set. Myronenko's research found out that previous point set registration methods lack analysis of the structure layout of point set or have little consideration of robustness [3]. The robustness of the point registration method in remote sensing was usually rather heuristically [4]. Point set registration methods developed on Gaussian mixture models have been lately proposed and proved to be robust and efficient than most traditional methods [3, 5]. The robustness against outliers is greatly valued in the multi-modal image registration case, where images usually present very different image features due to different imaging mechanism.

Remote sensing images are usually of large size with presence of a lot of details and complicated features. As a result, features in remote sensing image should be spatially diverse, easy to detect, high repeatable and less computation-load. Steerable pyramid transform not only could generate multi-resolution images but also could detect meaningful spatial diverse distribution features [4]. However this feature detection is rather coarse making a robust feature matching algorithm necessary. Also, multi-modal image registration makes outliers harder to deal with than mono-modal case. Point set registration methods based on Gaussian mixture models provide good ways to satisfy previous requirements. By combining steerable pyramid transform with Gaussian mixture models based point set registration methods, our method excels traditional registration methods in following aspects:

- the multi-resolution pyramid transform is shift-invariant;
- feature detection could be done within the multi-resolution transformation thus saving computation load;
- robust point set registration with focus on the structure layout of the image which is more suitable for the multi-modal image registration case.

55.2 Combining Steerable Pyramid and Gaussian Mixture Models

In this section, we first lay out the theory foundation of our method—namely steerable filters, steerable pyramid and Gaussian mixture model based point set registration; then we propose our new method.

55.2.1 Steerable Filters and Steerable Pyramid

Steerable pyramid transform, proposed by Simoncelli, is a linear multi-resolution, multi-orientation image decomposition transform [1]. It is developed in order to overcome the limitation of orthogonal separable wavelet decompositions. Because steerable filters are more robust to translation, rotation and noise than the standard Daubechies wavelet filters, they enable steerable pyramid to be shift-invariant and extendable for desired feature detection [1]. Table 55.1 shows the difference between popular pyramid transforms [1].

Figure 55.1 shows the decomposition (both analysis and synthesis) of steerable pyramid [1]. Initially, an image is separated into low- and high-pass subbands, using filters L_0 and H_0. The lowpass subband is then divided into a set of oriented bandpass subbands and a lower-pass subband. This lower-pass subband is subsampled by a factor of 2 in the X and Y directions. The recursive construction of a pyramid is achieved by inserting a copy of the shaded portion of the diagram at the location of the solid circle.

55.2.2 Gaussian Mixture Models for Feature Point Set Registration

Point set registration methods using GMM are various [3, 5]. We choose Coherent Point Drift (CPD) in this paper. CPD consider the alignment of two point sets as a

Table 55.1 Differences of popular pyramid transforms

	Steerable pyramid	Separable orthogonal wavelet	Laplacian pyramid	Gabor(octave)
Jointly-localized (space/frequency)	Yes	Yes (can be)	Yes	Not inverse
Translation-invariant (no aliasing)	Yes (approx)	No	Yes (approx)	No
Oriented kernels	Yes	No (not diagonals)	N/A	Yes
Rotation-invariant (steerable)	Yes (approx)	No	N/A	No
Tight frame (self-inverting)	Yes (approx)	Yes	No	No

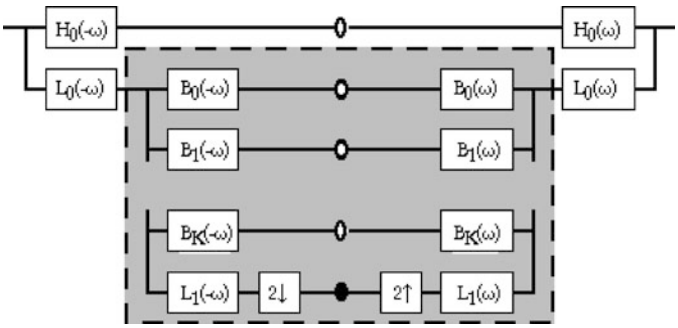


Fig. 55.1 Steerable pyramid

probability density estimation problem, where one point set represents the GMM centroids, and the other represents the data points [3]. Two point sets are aligned when the maximum GMM posterior probability is achieved. Denoting $X = (x_1, \ldots, x_n)^T$ the data point, $Y = (y_1, \ldots, y_n)^T$ the GMM centroids, the GMM probability density function is defined as:

$$p(x) = \omega p(x \mid M + 1) + (1 - \omega) \sum_{m=1}^{M} \mathrm{P(m)} p(x/m) \tag{55.1}$$

where $p(x/m) = \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{\|x - y_m\|^2}{2\sigma^2}\right)$. $p(x \mid M + 1) = 1/N$ is an additional uniform distribution with weight $\omega, 0 \leq \omega \leq 1$ added to the mixture models to account for noise and outliers. Equal isotropic covariance σ^2 and equal membership probabilities $P(m) = 1/M$ for all GMM components (m = 1,…,M) are used. And i.i.d (independent identical distribution) data assumption is made. Reparameterize the GMM centroid locations by a set of parameters θ and estimate them by maximizing the likelihood or, equivalently by minimizing the negative log-likelihood function. The GMM density estimation problem use EM algorithm as optimization method to solve the parameters. By deduction, for the case of affine transformation $T(y_m; B, t) = By_m + t$, where $B_{D \times D}$ is an affine transformation matrix, $t_{D \times 1}$ is the translation vector, the objective function takes the form

$$Q(B, t, \sigma^2) = \frac{1}{2\sigma^2} \sum_{n,m=1}^{N,M} p^{old}(\mathrm{m} \mid x_n) \|x_n - (By_m + t)\|^2 + \frac{N_p D}{2} \log \sigma^2 \tag{55.2}$$

where $N_p = \sum_{n=1}^{N} \sum_{m=1}^{M} p^{old}(m|x_n) \leq N$ (with N = N_p only if $\omega = 0$), p^{old} denotes the posterior probabilities of GMM components calculated using the previous parameter values,

$$p^{old}(m \mid x_n) = \frac{\exp\left(-\frac{1}{2} \left\| \frac{x_n - T(y_m, \theta^{old})}{\sigma^{old}} \right\|^2 \right)}{\sum_{k=1}^{M} \exp\left(-\frac{1}{2} \left\| \frac{x_n - T(y_k, \theta^{old})}{\sigma^{old}} \right\|^2 \right) + c}, with\ c$$
$$= (2\pi\sigma^2)^{D/2} \frac{\omega}{1-\omega} \frac{M}{N} \tag{55.3}$$

We can directly take the partial derivatives of Q, equate them to zero, and solve the resulting linear system of equations.

55.2.3 Proposed Method

The basic functions of steerable pyramid—steerable filters—are directional derivative operators which come in different sizes and orientations. The number of orientation may be adjusted by changing the derivative order. Steerable filters are highly potential basis function for lots of image processing tasks. Steerable filters are used to detect image features like canny criteria [2]. Mikolajczyk's research found out that for low-dimensional descriptor, steerable filters outperformed other descriptors like differential invariants [6]. In this paper, pixels with large values from the subband image are used as input feature point set. Netanyahu, etc. also used steerable pyramid as multi-resolution transform and initial feature detection, but the feature point registration method they used was heuristic adaption of least median of squares (LMS) estimation, which is vulnerable to a more difficult case like multi-modal registration [4]. Moreover their transformation model was restrained to be rigid and strict to a hand-chosen small value range and the image pairs to be registered were mono-modal, making their method limited to special cases. On the contrary, by introducing a robust point set registration method, our method has much less restriction on transformation model and it is used in multi-modal case. The benefits of CPD are twofold: 1. the point set moves coherently during transformation thus preserving the structure; 2. the robust registration process between estimation-step and maximization step ensures global minimum.

The registration process starts with the coarsest scale, namely the scale with smallest image size. For each scale, top 10 % pixels with the largest value of the subband image are input as initial feature candidates; a following coarse registration is done with this input feature set using CPD algorithm; and the result of this coarse registration serves as initial transformation value for the following scale. This scale-to-scale registration ends when it reaches the final scale with the original image size.

Also the transformation model is set to be increasing in our method. At a coarse scale, the transformation model between two images is chosen to be less complicate than the following scale. We choose rigid for small scales and affine for large scales. This is mainly because an early incorporation of complicate transformation model would actually degrade the image registration or even cause

failure [7]. At a coarse scale, the image suffers with less accurate details, thus the feature sets show relatively large structure similarity instead of detail local similarity. Therefore it is indeed needed to use a coarse transformation model at a coarse scale (Liu reviewed transformation models for image registration [8]). This hierarchical transformation model with hierarchical resolution image will ensure the optimization from falling into local minimum trap.

By combining steerable pyramid and Gaussian mixture model, we propose a hierarchical image registration method for multimodal image pair. Multi-modal image registration greatly suffers from outliers brought by feature detection method and different imaging modality. Traditional registration methods done on the original image with a high-order transformation model usually fails because of the outliers. Our method ensures an accurate results by: 1. a coarse registration at a coarse scale, saving time and ensuring a good initial guess for following scale; 2. a coarse transformation model at a coarse scale preventing falling into local optimization trap; 3. a robust point set registration method preserving the structure layout of the image which is more suitable for the multi-modal case.

55.3 Experiments and Analysis

Figure 55.2 shows the original optical-SAR image pair. It could be seen that optical image shows better resolution and better human-understandable features, while SAR image has better discrimination between different materials. The original optical image has a size of 1200*800, SAR image of size 600*400. Both images have been resized for displaying.

Figure 55.3 shows the multi-resolution and feature detection result. Each image has been decomposed into 3 levels. It can be seen that at a coarse scale, only the main structure features are present, thus making a registration with only a coarse transformation like rigid workable. At large scale, though more detail features are present, it is highly contaminated with outliers, which means it can easily fall into

(a) (b)

Fig. 55.2 Original optical-SAR image pair. **a** Optical image. **b** SAR image

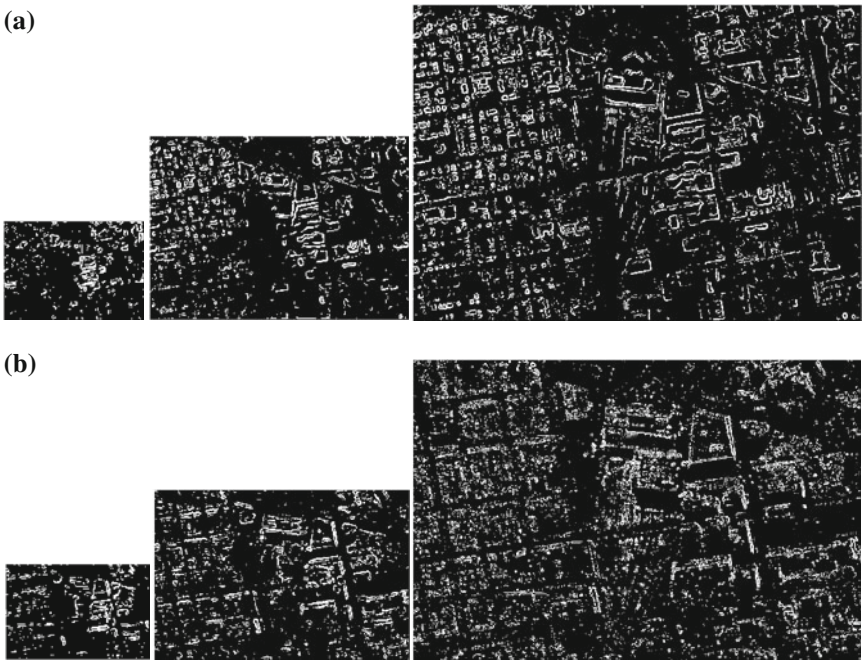


Fig. 55.3 Multi-resolution and feature detection of optical and SAR images. **a** Three level features representation of optical image. **b** Three level features representation of SAR images

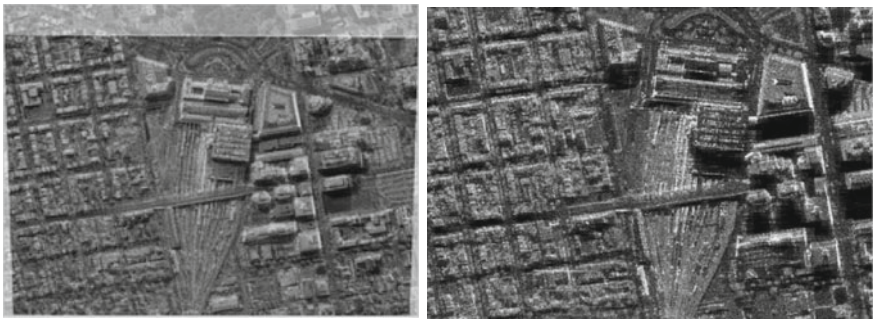


Fig. 55.4 Results of proposed method and CPD. **a** Our method. **b** Directly using CPD

local minimum trap without a good initial value. We set rigid transformation in the smallest size, and affine transformation in the following scales.

Figure 55.4 shows the registration result of using the proposed method and using CPD only. It could be seen that our method aligns the multi-modal image pair correctly, while the result of CPD is wrong (the result is not overlapped because it covers beyond the whole optical image). The RMSE (root mean square

error) of our method is 1.34 pixels. Direct implementation of CPD fails mainly because the highly-outliers-contaminated feature point set makes CPD method at the original scale fall into local minimum, while our method assures the correct result with a well computed initial registration result from coarse scale and coarse transformation model.

55.4 Conclusion

In this paper, researchers proposed a novel multi-modal image registration method by combing steerable pyramid and Gaussian mixture model based point set registration method. Steerable pyramid is translation-invariant and rotation-invariant and easy to extend to detect features. Gaussian mixture model based point set registration methods preserve the structure layout of the image and are more robust to outliers than previous point set registration methods. Both methods provide promising characteristics for multi-modal image registration which is bother by the easy trap in local minimum and high amount of outliers. By combining both methods, researchers have achieved better results otherwise could not been done by using any one. Experiments on real optical-SAR image pair demonstrated the method's feasibility. Further research could be done on speed-up computation and extendable feature extraction of steerable pyramid.

References

1. Simoncelli, E.P., Freeman, W.T.: The steerable pyramid: a flexible architecture for multi-scale derivative computation. In: IEEE Second International Conference on Image Processing. Washington (1995)
2. Jacob, M., Unser, M.: Design of steerable filters for feature detection using canny-like criteria. IEEE Trans. Pattern Anal. Mach. Intell. **26**(8), 1007–1019 (2004)
3. Myronenko, A., Song, X.: Point set registration: coherent point drift. IEEE Trans. Pattern Anal. Mach. Intell. **32**(12), 2262–2275 (2009)
4. Netanyahu, N.S., Le Moigne, J., Masek, J.G.: Georegistration of landsat data via robust matching of multiresolution features. IEEE Trans. Geosci. Remote Sens. **42**(7), 1586–1600 (2004)
5. Jian, B., Vemuri, B.C.: Robust point set registration using Gaussian mixture models. IEEE Trans. Pattern Anal. Mach. Intell. **33**(8), 1633–1645 (2011)
6. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell. **27**(10), 1615–1630 (2005)
7. Yang, G., Stewart, C.V., Sofka, M., Tsai, C.-L.: Registration of challenging image pairs: initialization, estimation, and decision. IEEE Trans. Pattern Anal. Mach. Intell. **29**(11), 1973–1989 (2007)
8. Liu, W., Ribeiro, E.: A survey on image-based continuum-body motion estimation. Image Vis. Comput. **29**(8), 509–523 (2011)

Chapter 56
Offset Modify in Histogram Testing of Analog-to-Digital Converter Based on Sine Wave

Chaotao Liu and Shirong Yin

Abstract This paper presented an offset modify method in histogram testing of ADC based on sine wave. The method of using histogram to estimate the code transition level is introduced firstly. Then the paper presented the modify method that use the difference between the cumulative histogram of low codes and the cumulative histogram of high codes. Simulation result demonstrated that the method introduced by this paper is efficient.

Keywords Histogram testing · ADC testing · Sine wave offset modify

56.1 Introduction

Histogram testing method is widely used for determination of nonlinearity errors of ADC (analog-to-digital converter). The histogram method involves the application of a given analog signal to the ADC input and the record of the number of times each code appears on the ADC outputs. Processing the measured data against a reference histogram then permits extracting the circuit's characteristics [1]. The excitation signals for ADC under test are usually a low-slope ramp signal. But a ramp generator is analog circuit which is normally very sensitive to noise and difficult built on chip. High-quality sine wave signal is easily generated by an all digital circuit. In addition, it is easy to improve sine wave purity by suitable filtering [2, 3]. So this paper researched how to use a sine wave stimulus and

C. Liu (✉) · S. Yin (✉)
School of Mechatronics and Automotive Engineering, Chongqing Jiaotong University, Chongqing, China
e-mail: liuchaotao@163.com

S. Yin
e-mail: yinsr@126.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_56,

histogram of ADC output codes to test the characteristics parameters of ADCs. Within the past few years, Histogram testing of ADC was paid much attention [4, 5]. But few approaches in the literature have proposed to resolve the bias of input sine wave parameter.

56.2 The Histogram of Sine Wave

The histogram test consists of stimulating the ADC with a known period signal and asynchronously acquiring a given number of samples. The result of the analog-to-digital conversion of those samples is then used to build a histogram, the x-axis is the possible codes appears on the ADC outputs, the y-axis is the number of each code. By comparing this number with the number expected of an ideal ADC, the actual transition voltages can be estimated.

For a sine wave input $v(t)$

$$v(t) = A\cos(\omega t) + C \tag{56.1}$$

where C, $A > 0$, ω is the offset (DC level), the amplitude, and the angular frequency respectively. The input signal $v(t)$ should exactly cover the full scale range of the converter in order to obtain the ideal reference histogram.

Uniformly sample the sine wave signal $v(t)$, the sampling phase is a random variable uniformly distributed in [0, 2π]. Then the sampled input signal V is a random variable. Thus $P(V)$, the distribution function of V that collected the samples with a value between $C - A$ and V, is represented by the fraction of the period 2π in which $v(t) < V$:

$$P\{V\} = \frac{1}{\pi}\arccos\left(\frac{C - V}{A}\right) \tag{56.2}$$

For a N-Bit ADC, if $V_{k+1} - V_k = V_{1LSB}$, where V_{1LSB} is the ideal code bin width, V_{k+1} and V_k are the $k+1$th and kth code transition level, show as Fig. 56.1. The probability function of the kth code appears on the ADC outputs that collected the samples with a value between V_{k+1} and V_k is:

$$p\{k\} = \frac{1}{\pi}\left[\arccos\left(\frac{C - V_{k+1}}{A}\right) - \arccos\left(\frac{C - V_k}{A}\right)\right] = \frac{\varphi(k)}{\pi} \tag{56.3}$$

where φ_k is the changed phase value when $v(t)$ changed from V_k to V_{k+1}.

When the phase of $v(t)$ changed from 0 to 2π, the time t changed from 0 to T, T is the period of the input sine wave, $T = 2\pi/\omega$. The sampling instance is a random variable uniformly distributed in [0, T] because the sampling phase is a random variable uniformly distributed in [0, 2π]. If the sampling period is T_s and the time of $v(t)$ changed from V_k to V_{k+1} is t_k, $\varphi_k = \omega t_k$, the probability function $p\{k\}$ can be represented as follow:

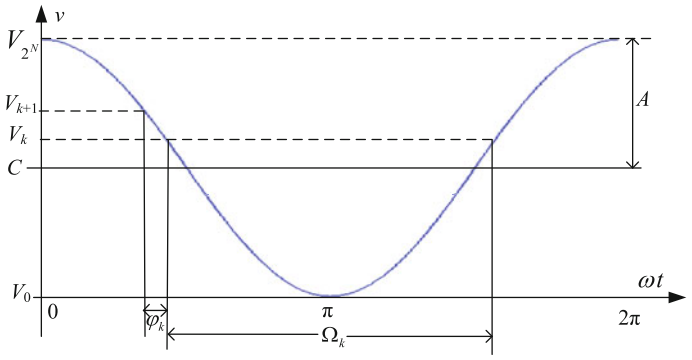


Fig. 56.1 The code transition level and the sampled input signal

$$p\{k\} = \frac{t_k}{T/2} \tag{56.4}$$

Let $h(k)$ be the total number of samples which yielded output code k as the result of the conversion. The total number of samples collected is

$$S = \sum_{i=0}^{2^N-1} h(i) \tag{56.5}$$

Because the input signal was uniformly sampled and sampling period is T_s, the probability function of the kth code appears on the ADC outputs $p\{k\}$ can be estimated by the relative frequency observed during the test:

$$\hat{p}\{k\} = \frac{h(k)}{S} \tag{56.6}$$

The kth cumulative histogram ch_k represent the number of samples that have a digital code equal to or lower than k,

$$ch_k = \sum_{i=0}^{k} h(k) \tag{56.7}$$

Then the distribution function of kth code transition level V_k can be estimated by the histograms of output code smaller than k:

$$\hat{P}(V_k) = \sum_{i=1}^{k} \hat{p}(i) = \frac{ch(k-1)}{S} \tag{56.8}$$

So the code transition levels can be estimated by $h(k)$

$$\hat{V}_k = C + A\cos\left[\pi \frac{ch(k)}{S}\right] \tag{56.9}$$

Once the code transition levels are determined, it is easy to obtain all the specification parameters which may be derived from the knowledge of the conversion characteristic, such as gain error, offset, Integral Nonlinearity (INL) and Differential Nonlinearity (DNL) [6].

56.3 Modify Offset

From Eqs. (56.3) and (56.8), (56.9) we can find that the estimated kth code transition level $\hat{V}_k$ and probability function $p(k)$ are the function of amplitude A and offset C of the input sine wave. The error of A and C will impact on the precision of the estimated V_k, so we have to estimated the A and C to modify the V_k. Provide that the A and C can be precisely controlled so as to $v(t)$ span all the code transition levels of the ADC, and synchronous sampling is preferred, the histogram of the ADC output will be a symmetry bathtub, show as Fig. 56.2a. The symmetry axis is $k = 2^{N-1} - 1$. The transition level of 2^{N-1} and 0 are positive peak and negative peak respectively.

If the C is biased form the regular value, the histogram of the ADC output will be a asymmetry bathtub, show as Fig. 56.2b. N_L and N_H represent the total number of samples which conversion codes are smaller than $2^{N-1} - 1$ (low codes), and represent the total number of samples which conversion codes are larger than $2^{N-1} - 1$ (high codes).

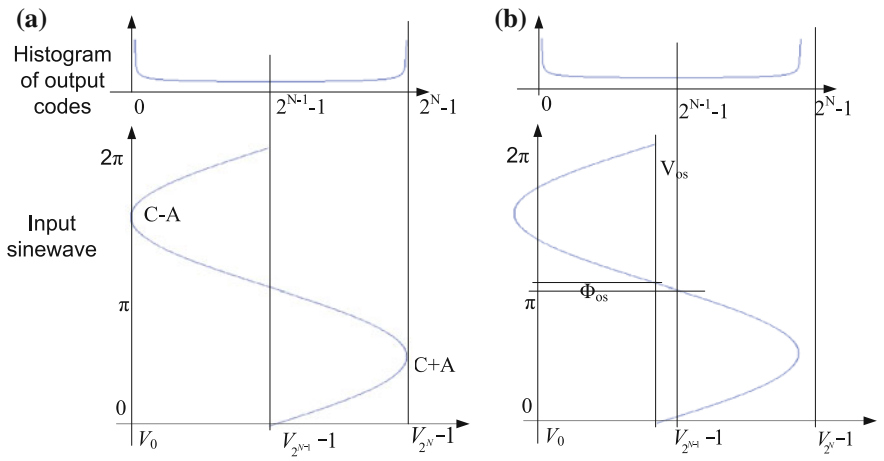


Fig. 56.2 The histogram of the ADC output that the input sine wave biased and not biased. **a** symmetry histogram C not biaed **b** asymmetry histogram C biaed

$$N_{\mathrm{L}} = \sum_{i=1}^{2^{N-1}-1} H(i) \tag{56.10}$$

$$N_{\mathrm{H}} = \sum_{i=2^{N-1}}^{2^{N}-1} H(i) \tag{56.11}$$

From Fig. 56.2b we can find that the phase difference between high codes and low codes is 4 Φ_{os}. From Eqs. (56.4)–(56.8) we can calculate the phase Φ_{os} is

$$\Phi_{os} = \frac{\pi}{4}\frac{N_H - N_L}{N_H + N_L} = \frac{\pi}{4}\frac{N_H - N_L}{S} \tag{56.12}$$

From Eq. (56.9) we can calculate the bias voltage is

$$\hat{V}_{os} = A\sin(\Phi_{os}) \tag{56.13}$$

Then the offset C can be modified by $\hat{V}_{os}$

$$\hat{C} = C + \hat{V}_{os} \tag{56.14}$$

56.4 Simulation Result

An 8-bit ADC of ADI corporation, the AD9289 is used to validate the efficiency of the method presented in this paper. The AD9289 is a monolithic, single-supply, 65 MSPS ADC with an on-chip, high-performance sample-and-hold amplifier and voltage reference, 1Vp-p to 2Vp-p input voltage range. We use ADIsimADC with Matlab on a PC platform to capture the ADC output codes and calculate the characteristics parameters. The behavioral modeling of the ADC and ADIsimADC are offered by ADI Corporation.

If the stimulus $v(t) = \sin(\omega t) + 1.5$, $A = 1$ and $C = 1.5$ not biased, $v(t)$ exactly cover the full scale range of the converter, the histogram and Differential Non-linearity (DNL) of the ADC output are shown as Fig. 56.3. The DNL is less than 0.4 LBS. Figure 56.4 shows the ideal code transition levels V_k and the estimated code transition levels $\hat{V}_{\mathrm{k}}$.

Make C biased 0.1, that is to say, $C = 1.4$, the histogram and DNL of the ADC output are shown as Fig. 56.5. The DNL is less than 4 LBS. Figure 56.6 shows the ideal V_k, estimated $\hat{V}_{\mathrm{k}}$ for C biased 0.1 and the modified code transition levels V_k. The estimated bias voltage using the method of this paper is 0.098. The DNL that calculated with the modified code transition levels V_k is shown as Fig. 56.7, the DNL is less than 0.6 LBS.

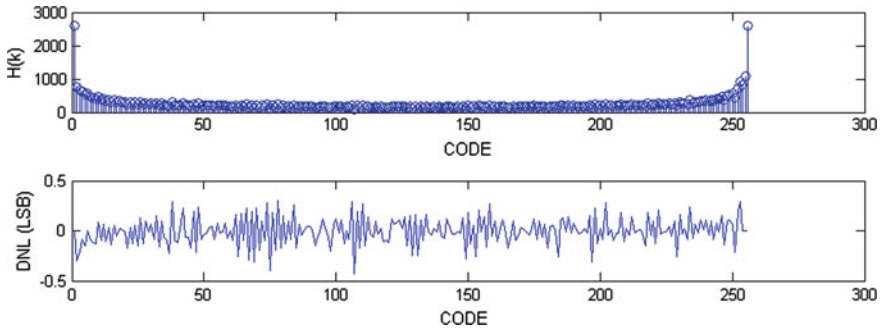


Fig. 56.3 The histogram and DNL of the ADC output, *C* not biased

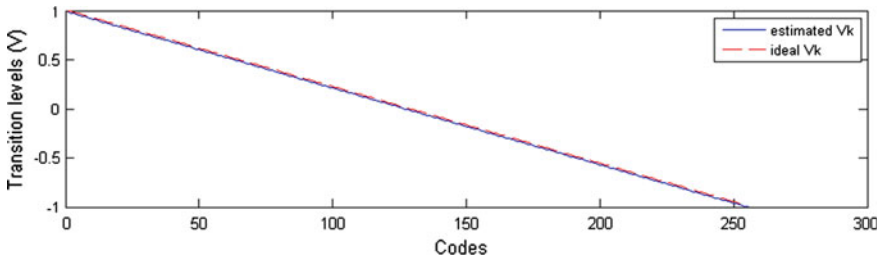


Fig. 56.4 The estimated V_k, *C* not biased

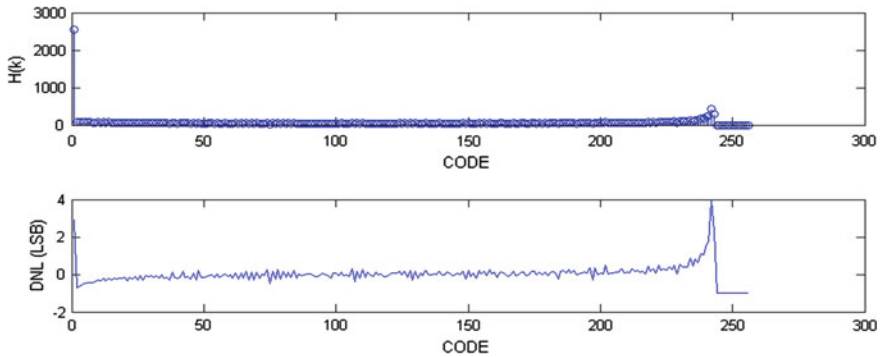


Fig. 56.5 The histogram and DNL of the ADC output, *C* biased 0.1

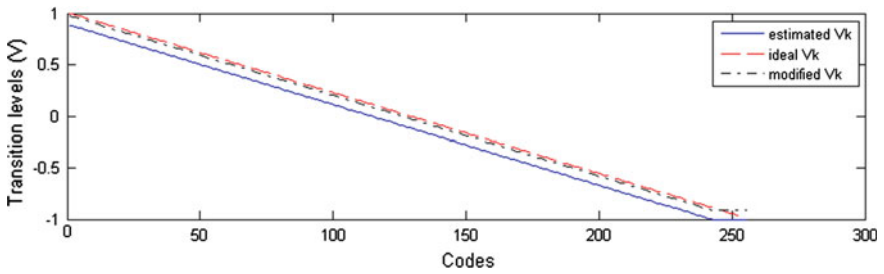


Fig. 56.6 The estimated V_k, C biased 0.1

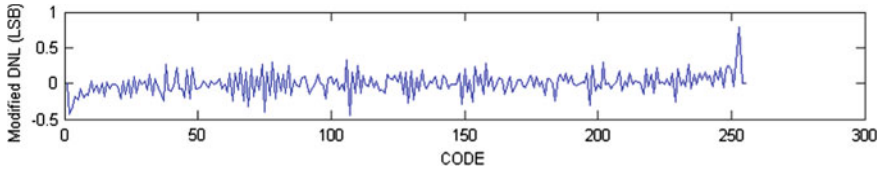


Fig. 56.7 The modified DNL, C biased 0.1

56.5 Conclusion

This paper presented an offset modify method that used the difference between the cumulative histogram of codes less than $2^{N-1}-1$ and the cumulative histogram of codes larger than 2^{N-1}. Simulation result demonstrated that the method introduced by this paper is efficient.

Acknowledgments This research is supported by the National Nature Science Foundation of Chongqing, China. (No. 2010BB2276).

References

1. Bernard, S., Azaïs, F., Bertrand, Y., Renovell, M.: Analog BIST generator for ADC testing. In: Proceedings of the 2001 IEEE International Symposium on Defect and Fault Tolerance in VLSI Systems (DFT'01), pp. 1–9 (2001)
2. Tone, M.F., Roberts, G.W.: A frequency response, harmonic distortion, and intermodulation distortion test for BIST of a Sigma-Delta ADC[J]. IEEE Trans. Circuits SystemS-11 Analog Digit. Signal Process. **43**(8), 608–613 (1996)
3. Lee, K.-J., Chang, S.-J., Tzeng, R.-S.: A Sigma-Delta modulation based BIST scheme for A/D converters. In: Proceedings of the 12th Asian Test Symposium (ATS'03), pp. 1–4 (2003)
4. Corrêa Alegria, F., Cruz Serra, A.: The histogram test of ADCs with sinusoidal stimulus is unbiased by phase noise[J]. IEEE Trans. Instrum. Meas. **58**(11), 3847–3854 (2009)

5. Korhonen, E., Kostamovaara, J.: An improved algorithm to identify the test stimulus in histogram-based A/D converter testing. In: Proceedings of the 13th European Test Symposium, pp. 149–154 (2008)
6. Dallet, D., da Silva, J.M.: Dynamic Characterisation of Analogue-to-Digital Converters[M]. Science Press in China (2007)

Chapter 57
Image Text Extraction Based on Morphology and Color Layering

Zhen Zhang and Feng Xu

Abstract In order to extract text regions in the color image, this paper proposes an image text extraction method based on morphology and color layering. Firstly, we extract the edges with Sobel operator, and then extract rectangular regions and non-rectangular regions according to text features by using morphological methods. Finally, we handle these two types of regions respectively where in Color Layering Algorithm is implemented when dealing with non-rectangular regions. This paper also proposes Large Characters Repair Algorithm so that the method can also be applied to the images with texts of different fonts. During the research, we have found that Color Layering Algorithm can separate text from complex background effectively, which makes background removed more easily. The experimental results show that the proposed method has a high accuracy rate.

Keywords Image text extraction · Morphology · Color layering · Image character recognition

57.1 Introduction

With the development of Internet, image search becomes more and more popular. Optical Character Recognition (OCR) is an important part of image search, image annotation and text-graphic conversion applications. Image text extraction in OCR technology has great significance. In recent years, scene image text extraction has become popular. Complex layout of the content and background are common in scene image. We can't get satisfactory results by using OCR directly, so we need

Z. Zhang · F. Xu (✉)
College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China
e-mail: njxuf@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_57,

to remove background and extract text regions. Scholars have proposed many methods but no one can solve this problem well [1–4].

Hasan et al. used morphological method to remove small regions [5]. But it will result in losing a lot of signals. Calculating the number of region's corners is used in this method to remove the region, which has poor effect when it is applied to the image which contains small font texts. Kim et al. extracted text regions using low-level image features and high-level text features [6]. Wang et al. used morphology to extract text regions [7]. But the threshold used in the BB method to remove regions is too large which is easy to cause the deletion of small font texts mistakenly. Lee et al. combined binarization method with color clustering method to extract text regions [8].

In this paper, we propose an image text extraction method based on morphology and color layering. In this method, we extract the edges with Sobel operator and binarize the image, then do morphological processing, mark and repair connected regions. After it, we mark rectangular regions and non-rectangular regions, apply Edge Color Algorithm to rectangular regions. Finally, Color Layering Algorithm is implemented to non-rectangular regions. Color layering is similar to human vision that focuses on the entire color blocks or the relationship between color blocks. The experimental results show that the proposed method has a good effect on the distortion images and can also be applied to the lower contrast images.

57.2 Image Text Extraction

The input of the method is a color image and the output is a binary image. The output can be used as the input of OCR system. The method can be divided into five steps: extract edge and binarize image, morphological processing and calculate connected regions, mark rectangular region and non-rectangular region, process rectangular region and non-rectangular region.

57.2.1 Extract Edge and Binarize Image

Texts in the image have edges, so we use Sobel operator to extract the edges firstly. Sobel operator is a finite-difference operator. It has a good effect on the image edge extraction. The operator contains two matrices to extract the edges of the horizontal and vertical directions respectively. In this paper, we use two matrices:

$$Horizontalx = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{57.1}$$

$$Verticaly = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{57.2}$$

After being processed by Sobel operator, the image is still not a binary image. In order to improve binarization effect, we set threshold T = 128 and then calculate Tx and Tm respectively. Tx is the mean value of the pixels greater than T and Tm is the mean value of the pixels smaller than T. Then we let T = (Tx + Tm)/2 and repeat these steps ten times to get a final threshold Tu. Finally we obtain the binary image using formula (57.3).

$$RGB(i) = \begin{cases} (0,0,0), AVG(i) < Tu \\ (255,255,255), AVG(i) \geq Tu \end{cases} \tag{57.3}$$

where RGB(i) is RGB value of the pixel i and AVG(i) is the mean value of the color components in the pixel i (Fig. 57.1).

57.2.2 Morphological Processing and Calculate Connected Regions

The morphological technology is a special image processing and analysis method developed from Set Theory Method based on mathematical morphology. Erosion and dilation are two basic operations in morphology. It is defined as follows:

$$\text{Erosion: } X = E \odot B = \{x : B(x) \in E\} \tag{57.4}$$

$$\text{Dilation: } Y = E \oplus B = \{y : B(y) \cap E \neq \emptyset\} \tag{57.5}$$

where B is a structural unit and E is the image. Erosion makes the main region thinner and dilation dilates the main region.

Fig. 57.1 **a** Original test image; **b** result of binarization

After the first step, the image has black background and white edges. White edges are key data. Here we use morphological methods to do dilation once and do erosion once, and then do dilation three times. After that, white regions will cover all the edges and black regions are the background. Next, we calculate all connected regions (CRs) of white pixels and get connectivity mask denoted as G1. Connected regions contain texts and interference background.

Some images may contain large font texts. Through edge extraction and limited times of dilation, these texts may be hollow-carved and cannot all be covered in connected region (CR). Large Characters Repair Algorithm (LCRA) we proposed can solve this problem. Firstly, we mark black background to get connectivity mask G2. The hollow part CR2 in G2 must be wrapped by CR1 in G1. Then we add this kind of CR2 to repair set (RS). Finally we merge CR2 of RS into G1 (Fig. 57.2).

57.2.3 *Mark Rectangular Region and Non-rectangular Region*

There are many complex and irregular interference outlines in the binary image. We divide these interference outlines into two categories: one separated from text outline; the other overlapped with text outline. The text has its geometric features. After processing Chinese or English text with morphological method, the shape of text region will be rectangular. Interference outline and text region with interference outline will be irregular graphics. In order to increase the accuracy, we divide the regions in G1 into rectangular regions and non-rectangular regions according to formula (57.6).

$$RJV = \frac{NUMpix(CR)}{AREA(CR)} \tag{57.6}$$

where RJV is rectangular judgment value to judge the region, and NUMpix (CR) calculates the numbers of effective pixels in CR. AREA (CR) calculates area of enclosing rectangle of CR.

Fig. 57.2 **a** Not use LCRA; **b** use LCRA; **c** result of test image after this step processing

57.2.4 Process Rectangular Region

The image may contain non-text regions which are rectangular, such as the straight line. The length or width of these regions is very small. If the length or width is smaller than the preset threshold, we mark the region as non-text region.

The text which is not overlapped with interference outline will be covered in rectangle region. This kind of text is surrounded by a thin layer of background pixels. We don't know whether the color of text is whiter or blacker than background. But the color of pixels on the edge of the region is background color. According to this point, we calculate the mean value of the color of pixels on the edge of the region (ECA) as inspiring data to judge pixels to be background or text. We call it Edge Means Algorithm (EMA).

$$ECA = \frac{\sum_{i=1}^{num} PSum_i(R, G, B)}{num * 3} \tag{57.7}$$

where num is the number of edge pixels and $PSum_i$ (R, G, B) calculates the sum value of the color components in the pixel i.

Various colors of texts may also exist in one image. In this case, the traditional method using a single threshold will not be proper. The proposed algorithm is able to deal with this situation. The algorithm is applied to each region and can improve flexibility and accuracy (Fig. 57.3).

57.2.5 Process Non-rectangular Region

Through morphological processing, the texts are surrounded by the background in non-rectangular region. And we will do a simple processing to remove the determined background pixels preliminary in order to reduce the complexity of the

Fig. 57.3 Result of test image after this step processing, where the letter "E" is marked as rectangular region

subsequent steps. Firstly, we traverse the edge of the region and record the colors into edge color set (ECS), and then traverse the entire region. If the color of the pixel is in ECS, we mark the pixel as background.

Due to non-rectangular features, we propose Color Layering Algorithm (CLA). Each non-rectangular region is layered, and each layer represents a color. After being layered, non-text part will be separated into different layers. The interference outlines overlapping with text are also separated into different layers so that we can eliminate the interference outlines effectively. If we layer the region based on RGB space, we will need to handle a huge amount of data. RGB space is sensitive to noise and illumination. So we use HSI space, and only use the H (hue).

$$H = \begin{cases} \theta, G \geq B \\ 2\pi - \theta, G < B \end{cases} \tag{57.8}$$

$$\theta = \cos\left(\frac{(R-G)+(R-B)}{2\sqrt{(R-G)^2-(R-B)(G-B)}}\right) \tag{57.9}$$

H in HSV ranges 0 from 360 which is still too large. The colors between 15 consecutive values of H are similar, and have less impact on the algorithm accuracy. We merge these layers into one layer and finally get 24 layers.

In fact, due to the influence of noise and other software processing, one text may be separated into different layers. If the pixels are not completely destroyed, these layers will be continuous in H. According to this feature, we calculate connection pixels degree (CPD) in adjacent layers. If CPD is greater than the preset threshold, we merge these layers.

$$NPN = \sum_{i=1}^{num} ISCEN(Pi) \tag{57.10}$$

$$CPD = \frac{NPN}{SUM(GLM)} \tag{57.11}$$

where NPN is the number of adjacent pixels between two layers and ISCEN (Pi) judges whether pixel Pi is adjacent pixel. GLM is the layer which has less effective pixels of the two layers. SUM (GLM) calculates the number of effective pixels.

After using CLA, we need to extract text regions. Firstly, we use morphological methods to do dilation five times on each layer, and then remove the regions with too small width or height. Secondly, we calculate rectangular judgment value (RJV) of each region, and then remove regions according to RJV. Finally, we merge all the layers and obtain binary text extraction image (Fig. 57.4).

Fig. 57.4 Result of test image after this step processing

Table 57.1 Extraction results

	Total	Correct	Precision (%)
Simple background image	249	215	86
Complex natural image	286	210	73
ICDAR	536	354	66

57.3 Experiment

For evaluating the performance of the proposed method, we used the dataset of ICDAR 2002 Robust Reading and Text Locating and the pictures downloaded from the network. The pictures from the network are divided into simple background images and complex natural images. Simple background images have less interference outlines and have high contrast between background color and text color.

The experimental results are given in Table 57.1.

The results show that the proposed method has good performance in lower color contrast and can be applied to the images with texts of different fonts. Due to using Color Layering Algorithm, complex background of natural images can be removed well. Many errors occurred because the interference background color and text color are the same or too close. These interference backgrounds also have their own outlines. Therefore using Color Layering Algorithm to remove them is difficult.

57.4 Conclusion

In this paper, the authors propose a method that divides regions into rectangular regions and non-rectangular regions. After morphological processing, text region with less interference will show as a rectangular shape which has a good extraction

effect. For non-rectangular regions, Color Layering Algorithm separates background and text into different layers. So the authors can handle these layers with methods similar to the method handling rectangular regions. The experimental results show that the proposed method has a high accuracy rate.

Acknowledgments This work is supported by the China Aviation Science Foundation (No. 20101952021).

References

1. Liu, W.P., Fu, X.L., Zhao, H.Q., Li, X.L.: Automatic text extraction from color image. Comput. Eng. Appl. **21**(41), 79–82 (2005)
2. Chethan, H.K., Hemantha Kumar, G., Raghavendra R.: A novel edge based method to extract text in camera captured images. ICACCT, pp. 853–855 (2009)
3. Kim, E., Lee, S.H., Kim, J.H.: Scene text extraction using focus of mobile camera. ICDAR, pp. 166–170 (2009)
4. Angadi, S.A., Kodabagi, M.M.: Text region extraction from low resolution natural scene images using texture features. In: IEEE 2nd International Advance Computing Conference 2010, pp. 121–128 (2010)
5. Hasan, Y.M.Y., Karam, L.J.: Morphological text extraction from images. IEEE Trans. Image Process. **9**(11), 1978–1983 (2000)
6. Kim, K.C., Byun, H.R., Song, Y.J., Choi, Y.W., Chi, S.Y., Kim, K.K., Chung, Y.K.: Scene text extraction in natural scene images using hierarchical feature combining and verification. In: Proceedings of the 17th ICPR, pp. 679–682 (2004)
7. Wang, Y.M., Tanaka, N.: Text string extraction from scene image based on edge feature and morphology. In: Proceedings of Eighth IAPR Workshop on Document Analysis Systems, pp. 323–328 (2008)
8. Lee, S.H., Seok, J.H., Min, K.M., Kim, J.H.: Scene text extraction using image intensity and color information. CCPR 2009, pp. 1–5 (2009)

Chapter 58
Face Detection Using Ellipsoid Skin Model

Wei Li, Fangyuan Jiao and Chunlin He

Abstract In the presence of unequal lighting conditions and complex backgrounds, this paper proposes a novel face detection algorithm for color images which consists of four pivotal parts primarily: image preprocessing based on color balance and light equalization, skin region segmentation and extraction based on CbrCbgCgr ellipsoid skin model, image post-processing based on morphology, as well as face and facial feature detection based on AdaBoost classifier and facial geometry. Experimental results demonstrate that the algorithm can be effectively applied to the cases of unequal light, complex background and multi-face conditions.

Keywords Face detection · Face recognition · Skin model

58.1 Introduction

Face detection is a very import process of face recognition. In recent years, face detection in unequal light and complex background had become research focus in the fields of pattern recognition. There were many kinds of methods of face detection such as wavelet transform and neural network, which can be classified by three types approximately: the method of geometric features which requires of high quality image but has a limited range of application [1], the method of template matching which may consume a lot of time for computing [2], and the method based on classification which makes use of features such as skin color and brightness to segment face region and extract facial features for realizing face location and detection [3]. The third method avoids the explicit description of details of facial features, and it has the good real-time performance and stability,

W. Li (✉) · F. Jiao · C. He
Department of Computer Science, China West Normal University,
Nanchong, Sichuan, China
e-mail: nos036@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_58,

but is affected by unequal light and complex background easily. The algorithm in this paper was based on the third method. This paper present a face detection algorithm that is able to handle color images which are in conditions of unequal lighting, complex scene and multi-face conditions, based on color balance, light equalization, CbrCbgCgr ellipsoid skin model, morphology post processing, as well as AdaBoost classifier and facial geometry.

58.2 Face Detection Algorithm

In this paper, the flow of face detection algorithm is depicted in Fig. 58.1 which contains five modules: image input, image preprocessing, skin region segmentation and extraction, image post-processing, face and facial feature detection. The first module is responsible for inputting the color image containing faces, the second module is responsible for removing the color bias and balancing the unequal light, the third module is responsible for building the CbrCbgCgr ellipsoid skin model, calculating the skin likelihood based on this ellipsoid model, and acquiring the skin segmented and extracted images, the fourth module is responsible for removing noise and filling holes of facial features, acquiring the denoised images of the skin extracted image contained the facial holes (eyes and mouth), and the skin extracted image contained the entire face separately, the fifth module is responsible for acquiring the face region and facial features based on AdaBoost classifier and facial geometry.

58.2.1 Image Preprocessing

Because of the influence of the unequal light and input devices, the input image is usually noisy and the difference between skin color of face and background is probably not obvious. The distribution of skin region and the appearance of human face depend on lighting conditions greatly [4, 5]. The preprocessing was presented to depress the influence of varying and unequal light.

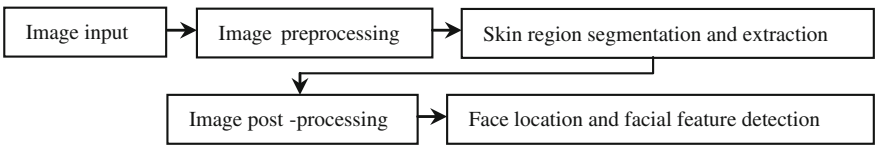


Fig. 58.1 Face detection algorithm flow

58.2.1.1 Color Balance

To solve the color bias problem caused by colored light source, the values of R, G, and B components of an input color image were adjusted by using adjusted Y, Cb, Cg, and Cr components. In this way, color bias of the input image can be removed and the original color characteristics of image scene can be restored appropriately. The steps of color balance are described as follows:

1. Acquiring Y, Cb, Cg, and Cr color components by (58.1).

$$\begin{bmatrix} Y \\ Cr \\ Cb \\ Cg \end{bmatrix} = \begin{bmatrix} 0 \\ 128 \\ 128 \\ 128 \end{bmatrix} + \begin{bmatrix} 0.299 & 0.578 & 0.114 \\ 0.500 & -0.4187 & -0.0813 \\ -0.1687 & -0.3313 & 0.500 \\ -0.316 & 0.500 & -0.184 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{58.1}$$

2. Acquiring the respective sum of Y, Cb, Cg, and Cr components (sY, sCb, sCg, and sCr), then acquiring the respective adjustment coefficient of Y, Cb, Cg, and Cr components (cY, cCb, cCg, and cCr) by (58.2), as well as calculating the respective average of Cb, Cg, and Cr components (aCb, aCg, and aCr) by (58.3):

$$\begin{aligned} cY = \frac{sCbgr}{sY},\ cCb = \frac{sCbgr}{sCb},\ cCg = \frac{sCbgr}{sCg},\ cCr = \frac{sCbgr}{sCr},\ sCbgr \\ = \frac{(sCb + sCg + sCr)}{3} \end{aligned} \tag{58.2}$$

$$aCb = sCb/(H \times W),\ aCg = sCg/(H \times W),\ aCr = sCr/(H \times W) \tag{58.3}$$

where H and W are the length and width of the input image.

3. Adjusting the value of Y, Cb, Cg, and Cr color components as follows:

$$\begin{bmatrix} Y \\ Cb \\ Cg \\ Cr \end{bmatrix} = \begin{bmatrix} Y \\ Cb \\ Cg \\ Cr \end{bmatrix} [cY \quad cCb \quad cCg \quad cCr] \quad \begin{array}{l} if \ |aCb - aCr| > 20 \\ or \ |aCb - aCg| > 20 \\ or \ |aCg - aCr| > 20 \end{array} \tag{58.4}$$

$$\begin{aligned} &Y = 255 \text{ if}(Y > 255),\ C_b = 255\ if(C_b > 255),\ Cg = 255\ if(\text{Cg} > 255), \\ &\text{Cr} = 255\ if(\text{Cr} > 255) \end{aligned} \tag{58.5}$$

4. Acquiring the value of YCb, YCg, YCr three components separately by (58.6).

$$YCb = (Y + Cb)/2,\ YCg = (Y + Cg)/2,\ YCr = (Y + Cr)/2 \tag{58.6}$$

5. Calculating respective maximum and minimum of YCb, YCg, and YCr color components (maxYCb, minYCb, maxYCg, minYCg, maxYCr, and minYCr),

and normalizing YCb, YCg, and YCr components to the range [0, 255] by (58.7).

$$\begin{cases} YCb = 255 \times \frac{(YCb - \min YCb)}{(\max YCb - \min YCb)} & if\ (\max YCb - \min YCb) > 0 \\ YCg = 255 \times \frac{(YCg - \min YCg)}{(\max YCg - \min YCg)} & if\ (\max YCg - \min YCg) > 0 \\ YCr = 255 \times \frac{(YCr - \min YCr)}{(\max YCr - \min YCr)} & if\ (\max YCr - \min YCr) > 0 \end{cases} \tag{58.7}$$

6. Calculating the maximum value of all adjusted YCb, YCg, and YCr color components (maxYCbgr), and adjusting R, G, and B components of the input color image by (58.8).

$$B = 255 \times YCb/\max YCbgr,\ G = 255 \times YCg/\max YCbgr,\\ R = 255 \times YCr/\max YCbgr. \tag{58.8}$$

58.2.1.2 Light Equalization

Light equalization for color balanced image can equalize the light further; especially can depress the influence of white light source, which is convenient to skin region segmentation and extraction. The light equalization is as follows:

$$B = 255 \times B/(B+G+R),\ G = 255 \times G/(B+G+R),\\ R = 255 \times R/(B+G+R). \tag{58.9}$$

Figure 58.2 shows the example of image preprocessing in this paper. Note that the blue bias color in Fig. 58.2a has been removed (see Fig. 58.2b), and the influence of unequal light in Fig. 58.2a has been reduced (see Fig. 58.2c). With preprocessing, this algorithm detects fewer non-skin pixels and more skin pixels (see Fig. 58.2e).

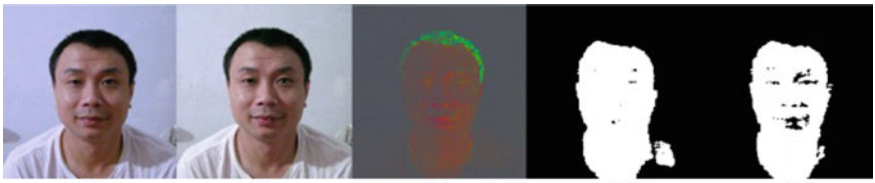


Fig. 58.2 Image preprocessing: **a** the color biased image, **b** the color balanced image, **c** the light equalized image, **d** the skin segmented image of (**a**), **e** the skin segmented image of (**c**)

58.2.2 Skin Segmentation and Extraction

The appearance of skin color of human face is not easily affected by factors such as facial expression, position and orientation, and meanwhile the face skin has stable features, centralized distribution and obvious differences from the background objects, which makes face segmentation based on skin color become a classic approach for face detection. The methods of skin segmentation at present are often based on skin model with statistical ideas. But because of some factors such as varying and unequal light, complex background and different race, 2D color space (e.g., YIQ, HSV, and YCbCr) may lost the information of facial features, so the 2D skin model based on 2D color space can not obtain satisfactory effect of skin segmentation as it has not too high preciseness. This paper built the CbrCbgCgr ellipsoid skin model in 3D color space to get the skin-likelihood image, and got skin segmented image and extracted image finally.

58.2.2.1 Skin Likelihood Calculation Based on CbrCgrCbg Ellipsoid Skin Model

In recent years, there are some mainstream skin model such as histogram model [6], Gaussian model [7] and ellipse skin model [3]. Among these models, the mixture Gaussian model and ellipse skin model based on YCbCr color space are used increasingly by more people. This paper built a novel CbrCbgCgr ellipsoid skin model according to the statistically analysis of the distribution characteristics of skin color clustering in CbrCbgCgr 3D color space, and calculated the skin likelihood for segmenting and extracting the skin region of color face image. Considering transforming RGB color space of the light equalized image (see Fig. 58.2c) into the CbCgCr color space by (58.1), and getting Cb, Cg, and Cr three color components, then calculating Cbr, Cbg, and Cgr by (58.10), which are the averages of every two color components.

$$Cbr = \frac{(Cb + Cr)}{2},\ Cgr = \frac{(Cg + Cr)}{2},\ Cbg = \frac{(Cb + Cg)}{2} \tag{58.10}$$

As the original input images, the 50 images in the MIT single face test set [8] included 10 subjects with the different color bias, races, light and poses. After image preprocessing, the light equalized images of original input images were chosen to acquire the skin samples. The statistics of skin samples showed that the distribution of skin color cluster was in the shape of a three-dimensional ellipsoid in the CbrCbgCgr 3D color space (see Fig. 58.3). Therefore, this paper proposed to build the CbrCbgCgr ellipsoid skin model, calculate skin likelihood like1 and get the gray image of skin likelihood based on this ellipsoid skin model as follows:

$$like1 = \frac{(\mathrm{Cbr} - 135.5)^2}{12^2} + \frac{(\mathrm{Cgr} - 132)^2}{6^2} + \frac{(\mathrm{Cbg} - 120)^2}{10^2} \tag{58.11}$$

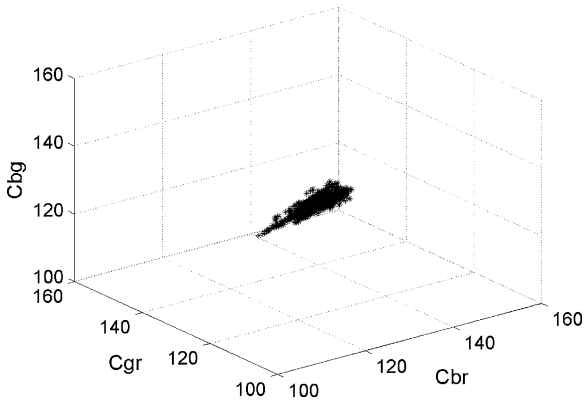


Fig. 58.3 The skin color cluster in CbrCbgCgr 3D color space

$$P_{i,j} = \begin{cases} like1 \times 0.5 \times 255; & \text{if like1} \leq 1 \\ 255; & \text{if like1} > 1 \end{cases} \tag{58.12}$$

where the $P_{i,j}$ is the value of pixel of skin-likelihood image.

Figure 58.4 shows the example of skin segmentation and extraction based on CbrCbgCgr ellipsoid skin model. The skin-likelihood image (see Fig. 58.4d) and the skin segmented image (see Fig. 58.4g) based on CbrCbgCgr ellipsoid model indicate the ellipsoid model had better preciseness than the CbCr Gaussian model (see Fig. 58.4b and e) and the CbCr Ellipse model (see Fig. 58.4c and f).

58.2.2.2 Skin Segmentation and Extraction Based on Skin Likelihood

After acquiring the skin likelihood and skin-likelihood image based on CbrCbgCgr ellipsoid model, by judging the non-white pixels of skin-likelihood image (see Fig. 58.4d) and reserving the corresponding region of color balanced image (see Fig. 58.2b), segmented the skin-likelihood image to get the skin segmented image and the skin extracted image (see Fig. 58.4g and h).

58.2.3 Image Post-Processing

This paper used the technology of morphology in the image post-processing, which used the close operation to remove noise of skin and non-skin region, and to get the skin segmented and extracted images contained the real eyes and mouth holes for following facial features detection. Then with variable template, it used open operation to fill the facial holes such as eyes and mouth, so it got the skin segmented image and extracted image that contained the entire face region, without facial holes. The examples of image post-processing based on morphology are shown in Fig. 58.4i–l. The denoised image of skin segmented image and the

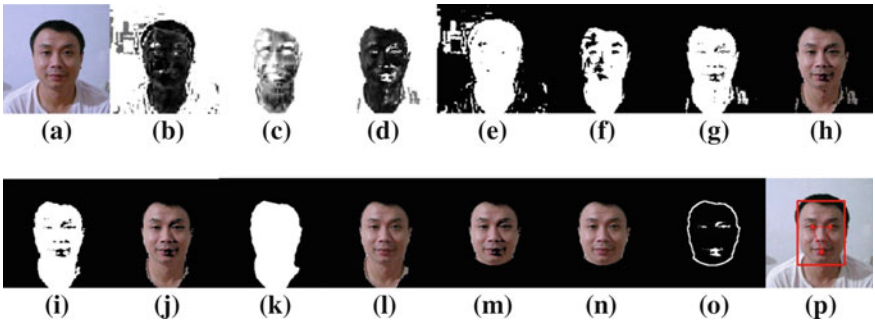


Fig. 58.4 Face detection: **a** the input image, **b** the skin-likelihood image of CbCr Gaussian, **c** the skin-likelihood image of CbCr ellipse model, **d** the skin-likelihood image of CbrCbgCgr ellipsoid model, **e** the skin segmented image of (**b**), **f** the skin segmented image of (**c**), **g** the skin segmented image of (**d**), **h** the skin extracted image based on ellipsoid model, **i** the denoised image of (**g**), **j** the denoised image of (**h**), **k** the skin segmented image contained the entire face, **l** the skin extracted image contained the entire face, **m** the face region image contained the facial holes, **n** the face region image contained the entire face, **o** the facial features image, **p** the face detection image

denoised image of skin extracted image, which both contained the facial holes, are shown in Fig. 58.4i and j. The skin segmented image and the skin extracted image, which both contained the entire face region, are shown in Fig. 58.4k and l.

58.2.4 Face and Facial Features Detection

The face and facial features detection is to get face region and detect facial features by using the skin segmented image and the skin extracted image which both contained entire face. This part consisted of face localization based on AdaBoost and facial features detection based on geometry and gradient.

After post-processing consisted of noise removing and holes filling, the skin segmented image and the skin extracted image both contained some candidate face region, one of which was the entire face region (see Fig. 58.4k and l). The principle of face localization is regarding the skin extracted image as an input image (see Fig. 58.4l) in this part, which will be detected by trained AdaBoost cascade classifier. In this way, the non-face skin region can be removed further and the real face region can be got, the more accurate face localization can be realized. As the skin connected region which probably contained the real face can be scanned in turn, this method can improve the speed of face localization and be applied to the conditions of multi-face, complex background and unequal light without needing to scan all the pixels of the whole picture. The face region image contained the facial holes and the face region image contained the entire face are shown in Fig. 58.4m and n.

Fig. 58.5 Multi-face detection: **a**, **c**, **e** the color input images, **b**, **d**, **f** the face detection images

Table 58.1 The average detection time on the test sets

Stage	Single face test set	Multi-face test set	Total
Image processing (s)	0.0935	0.454	0.274
Skin region segmentation and Extraction + Image processing (s)	0.055	0.265	0.16
Face location and facial features detection (s)	0.289	0.9	0.595

Table 58.2 Detection rate and detection time on the test sets

Algorithm	Hsu [3]	This algorithm
DR (%)	91.5	92.3
Time(sec):average ± s. d.	2.5 s ± 0.5 s	1.0 s ± 0.5 s

Facial features detection detected facial features such as eyes, mouth and face contour. The principle of detection is to use the face region image that contained the eyes and mouth holes, and to extract the eyes, mouth, and face boundary by judging the geometric position and morphological gradient of facial features. The results of face and facial features detection are shown in Fig. 58.4o and p.

58.3 Experiment Results

The algorithm was tested on a self-built face recognition database, which contained the MIT single face test set [8] and a multi-face test set. The multi-face test set contained 100 images with complicated background, unequal light, glasses. The experimental platform is: P4, 2.10 GHz CPU, 2G memory, WinXP OS, and VC++6.0. The results of multi-face detection are shown in Fig. 58.5. The average detection time in different stages on the test sets are shown in Table 58.1, the Table 58.2 listed the detection rate and the time of this algorithm as well as the algorithm of Hsu's [3]. From the Table 58.2, it is easy to find that the detection rate of this algorithm is higher than Hsu's, but the detection time is less than Hsu's.

58.4 Conclusion

A novel face detection algorithm for color image has been presented in this paper. In the process of algorithm, the color balance removed the color bias, then the light equalization depressed the influence of unequal light, and then the CbrCbgCgr ellipsoid skin model was built to decrease the error segmentation in order to improve the accuracy of skin segmentation or skin extraction, the morphological post-processing got the skin region contained facial features holes and skin region contained the entire face, which was helpful to localize face region faster by using AdaBoost cascade classifier, and to detect the facial features faster. The experiments shown in Fig. 58.5 demonstrated that this algorithm had satisfactory processing results for color image with single face, multiple faces, unequal light and complex background.

References

1. Lam, K.M., Hong, Y.: Location and extracting the eye in human face images. Pattern Recognit. **5**, 771–779 (1996)
2. Maio, D., Maltoni, D.: Real-time face location on gray-scale static images. Pattern Recognit. **9**, 1525–1539 (2000)
3. Hsu, R.L., Mottaleb, M.A., Jain, A.K.: Face detection in color images. IEEE Trans. Pattern Anal. Mach. Intell. **5**, 696–706 (2002)
4. Liu, L.Y., Sang, N., Yang, S.Y., Huang, R.: Real-time skin color detection under rapidly changing illumination conditions. IEEE Trans. Electron Devices **3**, 1295–1302 (2011)
5. Choi, S.I., Jeong, G.M.: Shadow compensation using fourier analysis with application to face recognition. IEEE Signal Process. Lett. **18**, 23–26 (2011)
6. Liu, Q., Peng, G.Z.: A robust skin color based face detection algorithm. In: Proceedings of 2010 2nd International Asia Conference Informatics in Control, Automation and Robotics, pp. 525–528 (2010)
7. Li, Q., Ji, H.B.: Face detection in complex background based on Gaussian models and neural networks. In: Proceedings of International Conference on Signal Process. (ICSP' 06) (2006)
8. CBCL face recognition database. http://cbcl.mit.edu/software-datasets/heisele/facerecognition-database.html

Chapter 59
Emergency Pre-Warning Decision Support System Based on Ontology and Swrl

Baohua Jin, Qing Lin, Huaiguang Wu and Zhongju Fu

Abstract Emergency resource is too huge to make quick decision of pre-warning issues. Aiming at this problem, ontology and Swrl (Semantic Web-Rule Language) rules are introduced in emergency pre-warning decision support system to express and integrate the pre-warning resource. And intelligent reasoning is also provided in this system. The validity of ontology model and Swrl rules are verified feasible in experimental results. The shortage of this system is also exposed in practice. Improvement is needed in future research.

Keywords Emergency pre-warning · Ontology model · Swrl rules · Reasoning method

59.1 Introduction

In recent years, food safety, weather disasters, and any other kinds of emergency events happened frequently. Qingdao "2.10" snow cooling event, Wenchuan "5.12" violent earthquake event and Yangzhou "9.1" hydrogen sulfide poisoning accidents fully expose deficiency of early warning application in emergency events. Pre-warning is a series of warning messages and corresponding measures according to comparison of current situation or assuming situation and normal situation. Pre-warning has wide application in the field of financial risk and credit fraud analysis. To ensure security of people and living standard of people's life, knowledge of emergency cases should be sufficiently applied to intelligent warning decision to establish emergency pre-warning decision support model. And knowledge representation and knowledge reasoning are two significant aspects of establishing this

B. Jin · Q. Lin · H. Wu (✉) · Z. Fu
School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, China
e-mail: hgawu@126.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_59,

system. Knowledge representation is the basis of establishing pre-warning decision support system. Knowledge reasoning takes emergency knowledge representation as the foundation. Logic rules are formed by identifiable data structure in the field of emergency pre-warning. Logic judgments based on semantic and solutions are provided by aspects of knowledge organization, intelligent retrieval and reasoning in emergency pre-warning decision support system.

In current academic cognition, knowledge decision support system are divided into several types which are the method based on relation, the method based on object-oriented, the method based on models (framework model, neural network model, ant colony model, etc.) and the method based on rules. In situation of domestic and international research, the typical emergency information system knowledge model are ERP (Enterprise Resource Planning) model, scene knowledge model, knowledge model based on common-KADS (Common Knowledge Acquisition Documentation and Structuring) method, etc. After basic research on ontology model, ontology and Swrl (Semantic Web-Rule Language) are introduced into emergency decision support system [1]. And the presentation and reasoning rules of this system are lucubrated.

Chinese natural language text is stressed as the research object [2]. And causal knowledge and dictionary drive rules are extracted by matching method from emergency field. And the establishment of emergency knowledge base which would be decomposed into several simple problem solving is the emphasis [3]. While in this paper, the experiment is more pertinence to the circumstances of issuing pre-warning decision support in emergencies. Emergency pre-warning decision support ontology is edited in protégé according previous cases. Swrl rules are extracted and would be consummate by absorbing experience for reasoning of Jena inference engine. The application of emergency pre-warning decision support system is attempting to provide new train of thoughts of intelligent decision support in the emergency field.

59.2 Related Concepts about Ontology and Rules

59.2.1 Ontology and Its Descriptive Language Rdf/Owl

The concept of ontology firstly comes from philosophy. Ontology is used for describing essence of abstract entities. Neches firstly proposed the deep meaning of ontology: “ontology is constituted of basic terminology and relationship in corresponding field, which are used for forming the definition of extensional rules” [4]. While Gruer proposed clearer definition: “ontology is the conceptual model of clear specifications” [5]. The later definition is improved by Borst: “ontology is sharing the conceptual model of the formal specification” [6]. In a word, the essence of ontology is the static conceptual model description in some field. And terminology and the relationship between terminologies are used to reflect the knowledge and knowledge structure.

59.2.2 Swrl

Swrl is a language which is presented by semantics [7]. The concept of Swrl rules evolve from Rule-ML (Rule Markup Initiative) and are formed by OWL ontology. Swrl has been one of the W3C standards. To combine Horn-like rules with OWL, Swrl integrate the description of Unary/Binary Datalog Rule-ML based on OWL DL and OWL Lite. The Swrl API provides a mechanism to create and manipulate Swrl rules in an OWL knowledge base. Thus Swrl makes up the insufficiency of OWL in description and reasoning.

59.3 Establishment of Emergency Pre-Warning Decision Support System Based on Ontology

59.3.1 Establishment of Emergency Pre-Warning Decision Support System Ontology

59.3.1.1 Descriptive Language of Emergency Pre-Warning Decision Support System Ontology

The establishment of emergency pre-warning decision support system ontology relies on the descriptive language of ontology. RDF/OWL is a kind of ontology language which is formed based on RDFS. The RDF/OWL is taken advantage of DAML + OIL [8]. Specific relationship of ontology languages is shown as Fig. 59.1 The latest formal version of RDF/OWL is issued and recommended by the W3C organization [9]. Compared with other ontology languages, the OWL is superior in semantic mechanism. According to the relationship between father-class and sub-class in the semantic mechanism of OWL, hierarchical structure of emergency pre-warning decision system ontology is great presented, which is shown as Fig. 59.2. In order to more accurately define emergency pre-warning resource and the relationship of resource, Defining Properties is introduced. A property is a type of binary relation. It is divided into Object Properties and Data Properties. And it is restrained by domain and range of properties to lay the foundation for subsequent semantic retrieval and reasoning [10]. All in all, the flexible expression mechanism of OWL can greatly deal with resource location, object instance of functional expression and relationship between the concepts.

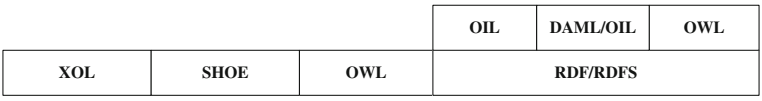


Fig. 59.1 The frame figure of ontology language

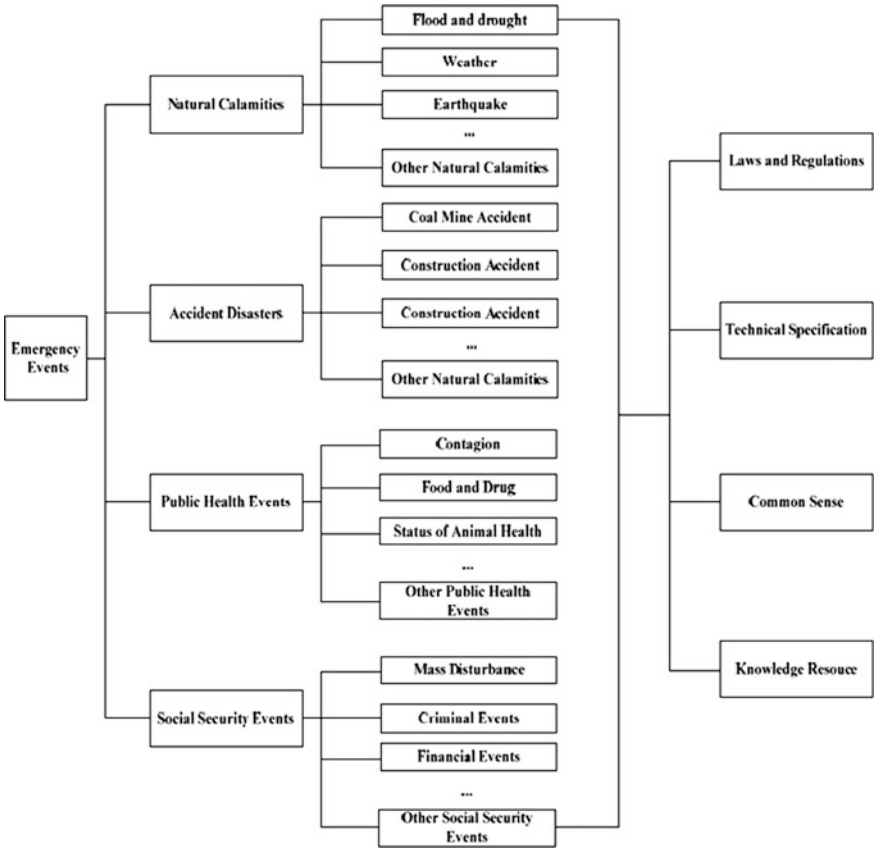


Fig. 59.2 Hierarchical structure figure of emergency pre-warning decision support system

59.3.1.2 Definition of Emergency Pre-Warning Decision Support System

Based on deep analysis of records in the field of emergency pre-warning, the machine-learning approach is used for extracting corresponding terminology, definition, abbreviation, standard jargon and frequently-used thesaurus in the field of emergency pre-warning (including plans, cases, emergency organization and emergency resource). Among the words, crucial conceptual words should be defined according to the minimum unit of ontology (records). The connotation of records contains type of event, definition of events' level, triggered condition, prevention measures, emergency scheme, post-recovery method and knowledge source. The records are related according to the connotation of records. For example, the records of weather disaster relate the type of event to cases is used for judging the rules the level of pre-warning in a weather event. And the conceptual words are defined in connotation table of weather disaster as follows: Table 59.1

Table 59.1 Level of pre-warning in weather disaster pre-warning events

Owl:thing		
Pre_warning	Warning_type	Typhoon
		Torrential_rain
		Blizzard
		Cold_wave
		Sand_storm
		High_temperature
		Drought
		Thunder
		Fog
	Level	Blue
		Yellow
		Orange
Weather_symptom	Symptom	Windy
		Rainy
		Snowy
		Sunny
		Thundery
		Foggy
	Degree	Heavy
		Light
Temperature	Tem_type	Warming
		Cooling
	Tem_Area	Initial_ temperature
		Final_ temperature
	Tem_range	Detailed
	Variation	Wide
		Narrow
Interval	Daytime	
	Night	

59.3.1.3 Description of the Relationship

According to the description of RDF/OWL, property is used for assertion of the facts of general instance and individual concrete facts. That is the descriptive factor between concepts, which is also used for concept and data [11]. Moreover, sub-property can inherit father-property. That means inherited rules exist between properties. The domain of a property is used for describing which class or instance is the subject of the property. And range of property is used for describing which class or instance is the object of the property. Base on this, level of pre-warning in emergency pre-warning decision support system is described by the properties of hasWeatherSymptom and its sub-property, hasTemVar, hasTemRange and so on. The specific relationship figure is showed in Fig. 59.3.

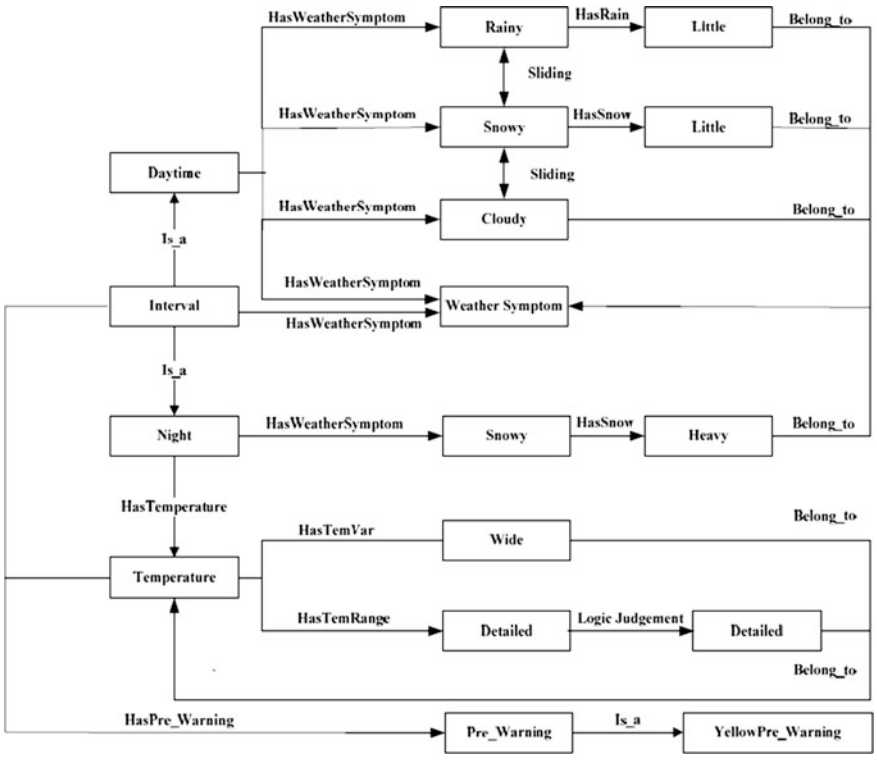


Fig. 59.3 Concept relation in emergency pre-warning decision support system

59.3.2 Emergency Pre-Warning Decision Support System Reasoning Rules Based on Swrl

59.3.2.1 Swrl Reasoning Rules

Semantic representation of RDF/OWL is taken advantage by emergency pre-warning decision support system to convey internal relations of class, property and instance. Moreover, certain reasoning is gotten by inheritance of property and anonymous class. For example, it can be deduced that two instances are corresponding to the same set or the same range by Functional Property. But the semantics mechanism of emergency pre-warning knowledge which established by RDF/OWL cannot fully meets the needs of the decision maker. And it also cannot get the integrated reasoning connotation in the emergency pre-warning field. The reasoning mechanism of emergency pre-warning rules which is established by Swrl not only can realize seamless connection with ontology model of RDF/OWL, but also preferably meets the need of presentation of rules.

Swrl can effectively expand the semantic of OWL, especially the capability of establishment of rules based on OWL. The uniform resource locator URI of each written Swrl rules is specified [12]. Swrl is made of the premise part antecedent and the conclusion part consequent. Swrl is consisting with syllogism in logic description. Swrl is made up of multiple atoms. The specific expression of Swrl is presented as follows:

rule :: = 'Implies(' [URIreference] {annotation} antecedent consequent')'
Antecedent:: 'Antecedent('{atom}')'
Consequent :: = 'Consequent (' {atom} ')'

Each atom is made of function C(x) of data and function P(x, y), different From (x, y), builtIn(r, x) of relationship. To combine with OWL, C is the description or data range of a class. And P is an OWL property. BuiltIn is the relationship of comparison between OWL instances or OWL data.

59.3.2.2 Establishment of Emergency Pre-Warning Decision Support System Based on Swrl

Take advantage of established emergency pre-warning decision support system ontology to extract corresponding reasoning atoms. Atoms are used for setting up rules and realizing relevance between instances. The instance of record of weather disaster event "2.1 Heavy Snowy Weather with a Sharp Fall in Temperature in Qingdao" is related with the instance of pre-warning "issue Icy Road Yellow Pre-warning", which is used for taking as example for following illustration.

According to the case "2.10 Heavy Snowy Weather with a Sharp Fall in Temperature in Qingdao" which is issued by Qingdao Emergency Management office, the post disposal of the case is described as followed: Meteorological observatory of Qingdao issued forecast about snowy weather with temperature decline in 8th February. Influenced by cold air, it shows it will be heavy windy sleety with temperature decline in the next 3 days. It will be cloudy with little sleet in the daytime of next 2 days. It will be heavy snowy during the night in 10th and daytime in 11th with gradually temperature decline. The temperature is up to around 6° C. Meteorological observatory of Qingdao issued Icy Road Yellow Pre-warning in 10th February 4 pm.

Replace the 6° C of decline of nighttime temperature with average range of decline of nighttime of nighttime temperature (b is valued as the threshold) in the Icy Road Yellow Pre-warning events of lately 10 years. So the information could be abstracted in first-order predicate logic as follow ("Λ" is used as conjunction operator and "∨" is used as disjunction operator):

Cloudy in the daytime Λ (Little Snowy in the daytime ∨ Little Rainy in the daytime) Λ (Heavy Snowy in the nighttime ∨ (Snowy in the nighttime Λ the temperature decline range is [6, +∞])) ->Icy Road Yellow Pre-warning is issued.

And the first-order predicate logic sentence could be turned into a rule as follow:

daytimeWeather(?x) Λ hasWeatherSymptom(?x,cloudy) Λ ((hasWeather-Symptom(?x,snowy) Λ hasSnow(?x,little)) ∨ (hasWeatherSymptom(?x,rainy) Λ hasSnow(?x,little))) Λ nightWeather(?y) Λ (hasWeatherSympotom(?y,snowy) Λ (hasSnow(?y,heavy) ∨ (hasTemVar(?y,wide) Λ hasTemRange(?y,?z) Λ Swrlb: MoreThanOrEqual(?z,b))) -> hasYellowWarning(?x,?y).

As to the Object Property hasYellowWarning is declared as symmetric property in OWL. Therefore, the symmetric relationship "P (x, y) if and only if P (y, x)" is existed in logic relationship. Therefore, equal mapping relationship of daytime-Weather and nightWeather entities is automatically formed by OWL. And the corresponding rules are established. That means the conclusion "Icy Road Yellow Pre-warning is issued" also could be deduced by the circumstance of interchange of the condition of daytime Weather and night weather.

According to the established rules, axiom is established. Some rules could be established by extracting the conceptual factors and combining with the ontology definition. The judgment conditions of icy road pre-warning are extended by several extracted texts to gradually establish the part of rules in intelligent emergency pre-warning decision system.

59.4 Realization of Ontology and Rules in Emergency Pre-Warning Decision Support System

59.4.1 Emergency Pre-Warning Decision Support System Reasoning Rules Based on Swrl

The software Protégé is an OSS (Open Source Software) which is developed in Java by Stanford University [13]. The Protégé is used for knowledge acquisition and ontology compilation. The Protégé is mainly used for the establishment ontology in semantic web. And the Protégé is the crucial development tools of the ontology establishment.

This experiment is finished under the circumstance of 2.50 GHz CPU. Protégé V.3.5 is selected as the experimental tool of ontology establishment and OWL is selected as the descriptive language. The specific presentation of entities of ontology is shown in Fig. 59.4.

According to national emergency knowledge classification criteria in Fig. 59.2, emergency events are divided into the category of natural disasters, accidents disasters, public health events and social security events. And these categories are divided into two or more categories. Aiming at the construction of emergency pre-warning decision support system, six classes of basic resource are defined in

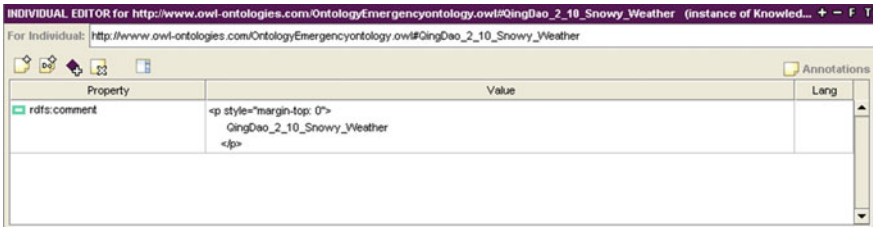


Fig. 59.4 Interface of ontology establishment in Protégé

ontology, which are Knowledge Record, Law Regulation, Technical Regulation, Emergency Common Sense, Knowledge Source and Emergency Type. Corresponding sub-classes is appended to basic classes. The root node of the whole ontology is started with owl:Thing in OWL. And every classes including custom classes are sub-classes of owl:Thing. For example, the natural disasters are set as a sub-class of emergency events. And the meteorological disasters are set to as a sub-class of natural disasters. The crucial code is shown as follows:>

```
<owl:Thing/>
<owl:Class rdf:ID="EmergencyType"/>
<owl:Class rdf:ID="Nature_Calamity">
        <rdfs:subClass Of rdf:resource="#EmergencyType"/>
</owl:Class>
<owl:Class rdf:ID="Meteorological_Disaster">
<rdfs:sub Class Of rdf:resource="# Nature_Calamity"/>
</owl:Cl ass>
```

The "2.10 Heavy Snowy Weather with a Sharp Fall in Temperature in Qingdao" which is described in Sect. 2.2.2 is taken as the example. The instance of Class Knowledge Record is established. And HasEmergency common sense, HasKnowldege _Record, HasLaws and Regulations, HasWeatherSymptom and other properties are defined in Object Property topObjectProperty which is the root node of Object Property. Corresponding sub-properties, domain and range are set to those properties. Take it for example; the Object Property HasSnow is set as a sub-class of Object Property HasWeatherSymptom. Its Domain is set to Symptom. And its range is set to Degress. The crucial code is shown as follows:

```
<owl:Class rdf:ID="Knowledge_Record"/>
<Knowledge_Record rdf:ID="Qingdao_2_10_snowy_Weather"/>
<owl:Object Property rdf:ID="HasSnow">
        <rdfs:sub Property Of rdf:resource="#HasWeather Symptom">
        <rdfs:domain rdf:resource="#Symptom"/>
        <rdfs:range rdf:resource="#Degree"/>
  </owl:Object Property>
```

At the same time, Data Properties tree is described and established. The description of object set is formed by relationship and restrain of properties and instances of classes.

59.4.2 Realization Tool of the Rules in Emergency Pre-Warning Decision Support System Swrl Editor

Swrl is a kind of rules expression language based on ontology [14]. Swrl Editor is run in the Protégé as a plug-in. It is an open source tool which is used for editing rules. And it is developed by Stanford University. Simply operable editor interface, highly relevance with OWL and interaction is provided by Swrl Editor. The integration mechanism of Swrl Editor and rule engineering is provided by Swrl Factory, which means the interoperability of the existing rules engineering API. The flexibility of creating, editing users and reading, writing Swrl rules is greatly improved via the establishment of mechanism. And the efficiency of accessing classes, properties and instances in OWL is also improved by this mechanism. The text box for creating rules in the inference of Swrl Editor is shown in Fig. 59.5.

59.4.3 Realization of Reasoning in Emergency Pre-Warning Decision Support System

Jena is the reasoning API aiming at RDF and OWL. It is developed by HP. And it can create, import and sustain RDF model. Jena Semantic is a reasoning system which is established by inspirational rules with CLISP in ontology field.

The Swrl rules are combined with the instances of ontology by Swrl Edition. Based on the established ontology of emergency pre-warning decision support system, Jena reasoning machine is selected to realize the data mining of

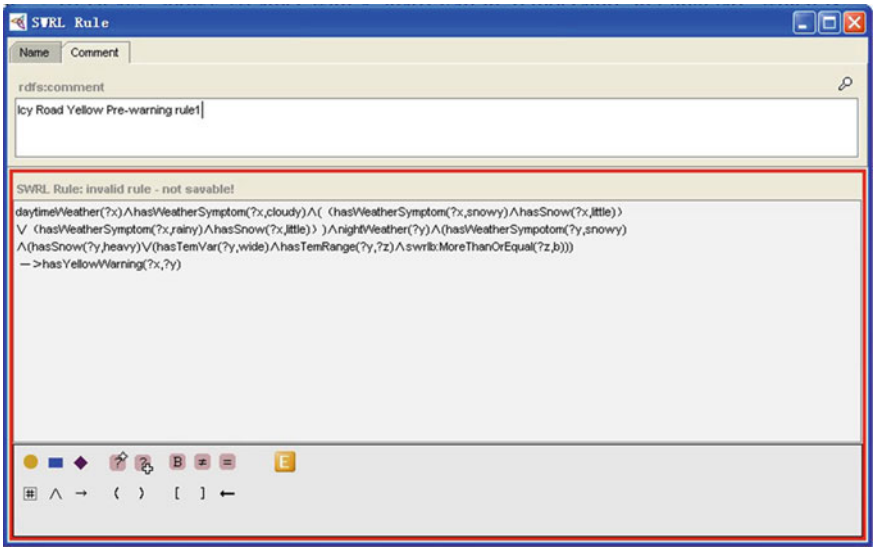


Fig. 59.5 The text box for creating rules in the inference of Swrl Editor

emergency pre-warning decision support system. And the crucial code is shown as follows:

```
/*Conceptual file Emergencyontology.owl is read in reasoning machine*/
Model schema=Model Loader.load Model("file:data/Emergencyontology.owl");
/*Instance document Emergencyontology.owl is read in reasoning machine*/
Model data=Model Loader.load Model("file:data/Emergencyontology.rdf");
/*OWL reasoning machine is created*/
Reasoner owl Reasoner = Reasoner Registry.getOWLReasoner();
/*The reasoning machine is bound to emergency pre-warning decision support
system inontology model*/
Reasonerwn Reasoner = owl Reasoner.bind Schema(wnontology);
/*The rules are added to existing reasoning rules set*/
String rules=read("file:data/rulefile.txt");
Reasoner reasoner=new Generic Rule Reasoner (Rule.parse Rules(rules));
```

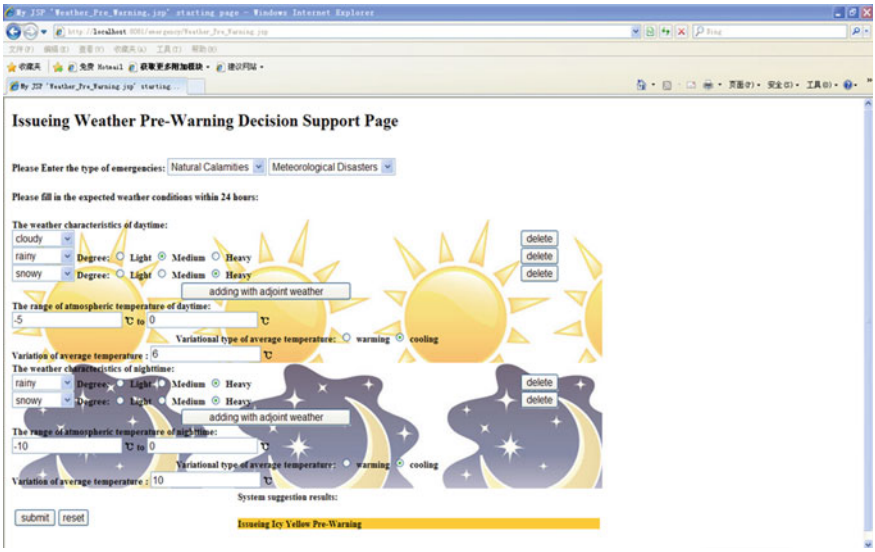


Fig. 59.6 The pre-warning decision result page of system

It is notable that Jena also is used as Java implementation. In our experiment, the jena.jar is directly imported into emergency pre-warning decision system. As to great compatibility of protégé and Jena inference engine to java, the reasoning of emergency pre-warning decision is realized by emergency pre-warning support system. Emergency pre-warning decision system is read into the system project which is compiled and run in the platform of myeclipse 8.5. And instance of Sect. 2.2.2 is taken as example. The purpose is the judgment of pre-warning type in the presupposed emergency scene. The page of issuing the suggestion of emergency pre-warning type is shown as Fig. 59.6.

It is obvious that the experiment is B/S structure form the Fig. 59.6. When a meteorological disaster occurs, the natural meteorological disaster is selected as the type of disaster. The system would jump to the issuing weather pre-warning decision support page. At the same time, the systematic background would be read into corresponding emergency ontology files (file format is *.owl) which includes the Swrl rules set according to the type of the above emergency type. The information of predicting weather conditions in 24 h which is provide by observatory as the judgment condition is needed to be input into the system. The reasoning result is the decision result of system proposal is shown. The inputting text box of forms in page is submitted to background and the reasoning result which is recalled by Ajax. It turns out that the result of experiment is promising and feasible.

59.5 Conclusions and the Future Works

All kinds of emergencies happen frequently. Developing scientific pre-warning mechanism has been more and more important in the process of emergency decision. Emergency pre-warning decision support system is mainly established by system ontology, Protégé OWL API, Swrl Editor and Jena reasoning engineering. In this paper, knowledge presentation and establishment of ontology of emergency pre-warning decision support system are discussed. The experiment realizes the semantic extraction of Swrl rules by generality of ontology. Because the cases in the system need more details, the reasoning of this system has limitations. The future study plan is adding more reasoning factors and making all Swrl rules complete to consummate effective emergency pre-warning decision support system.

References

1. Feng, Z., Yang, Q., Rao, G., Zhang, C., Li, G.: Representation and application of ontology based emergency plan for logistics. Appl. Res. Comput. **10**(5), 71–72 (2011)
2. Zhang, H.: Research on the extraction of dispose knowledge from text in emergency domain. Dalian university of Technology, Dalian (2010)
3. Qiu, J., Li, P., Wu, L., Wang, Y.: Research of emergency knowledge model based on problem. In: The 2009 IEEE/WIC/ACM International Conferences on Web Intelligence(WI'09) and Intelligent Agent Technology(IAT'09), pp. 325–328 (2009)
4. Xia, J.: Research on automatic reasoning techniques of semantic web in the context of ontology knowledge systems. Hefei University of Technology, Hefei (2004)
5. Ian, H., Ulrike, S.: Ontology Reasoning in the SHOQ Description Logic. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence, pp. 2–6 (2011)
6. Wang, C., Lan, H., Xie, H.: An integrated model of knowledge acquisition: empirical evidences in China. In: Proceedings of the International Conference on Information Management, Innovation Management and Industrial Engineering, pp. 335–338 (2008)
7. Zhong, X., Fu, H., Li, S., Huang, B.: Geometry knowledge acquisition and representation on ontology. Chin. J. Comput. **33**(1), 167–174 (2009)
8. Weber, R.O., Ashley, K.D.: Textual case-based reasoning. Knowl. Eng. Rev. **2**(7), 255–260 (2005)
9. Zhou, X., Cao, C.: Medical knowledge acquisition: an ontology based approach. Comput. Sci. **30**(10), 35–39 (2003)
10. Hetal, K.: The Protégé OWL plugin: An open development environment for semantic Web applications. Proceedings of the 3rd International Semantic Web Conference ISWC, Hiroshima. pp. 229–243 (2004)
11. Zen, Q., Cao, C.: Research on Mathematical knowledge acquisition and knowledge inheritance mechanism based on ontology. Microelectron. Comput. **9**(3), 19–27 (2003)
12. Fensel, D., Horrocks, I., van Harmelen, F., et al.: OIL: An ontology infrastructure for the semantic web. IEEE Intell. Syst. **13**, 38–45 (2001)
13. Zhang, M.A.: Generic knowledge-guided image segmentation and labeling system using fuzzy clustering algorithms. IEEE Trans. Syst. **32**(5), 571–582 (2002)
14. Rosina, O.W.: Textual case-based reasoning. Knowl. Eng. Rev. **7**(9), 255–260 (2005)

Chapter 60
Feature Reduction Using Locally Linear Embedding and Distance Metric Learning

Bo Yang, Ming Xiang and Liuwu Shi

Abstract Feature reduction is an important issue in pattern recognition. Lower feature dimensionality reduces the classifier complexity and enhances the generalization ability of classifiers. In this paper, researchers proposed a new method for feature dimensionality reduction based on Locally Linear Embedding (LLE) and Distance Metric Learning (DML). Researchers first adopt metric learning method to enhance the class separability, and map the original data to a new space. They use a transformation learned from the data via metric learning method, and then utilize the LLE method to generate an embedding from the transformed data to a lower dimensional manifold. Thus they achieving feature dimensionality reduction, where the final mapping for feature reduction is the composition of the above two transformations learned via DML and LLE method respectively. The method introduces the LLE method traditionally used in unsupervised tasks into the supervised learning domain via a proper and natural way. Experiment results clearly demonstrate the efficiency of the proposed feature reduction method in supervised learning tasks.

Keywords Feature reduction · Locally linear embedding · Distance metric learning

60.1 Introduction

Feature dimensionality reduction plays an important role in the domain of pattern recognition. Large feature dimension not only increases the complexity of classifier design, but also increases the risk of overfitting of the designed classifiers. Hence many techniques and methods have been designed to reduce the feature

B. Yang · M. Xiang (✉) · L. Shi
Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China
e-mail: mxiang@mail.xjtu.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_60,

dimension, which can roughly be classified as feature selection and feature extraction. Feature extraction includes linear and non linear methods. Traditional techniques such as Principal Component Analysis (PCA) [1], Multi-Dimensional Scaling [2], Linear Discriminant Analysis (LDA) [1], and etc. are linear dimensionality reduction methods, while some of the newly emerging methods such as Locally Linear Embedding (LLE) [3] and Isometric Maps (ISOMAP) [4] are nonlinear dimensionality reduction methods.

The idea of LLE method is that most real data lies on a low dimensional manifold embedded in a high dimensional space, and thus certain algorithms can be designed to map high dimensional data into a low dimensional space to reveal the underlying structure in the data, and thus achieving feature reduction. Experiment results in the literature have shown that LLE is an effective method for dimensionality reduction.

LLE method is usually used in unsupervised learning tasks, since it does not uses any information about the class label of the input data. To extend the LLE method to supervised learning tasks such as classifier design, a supervised LLE (SLLE) algorithm is proposed by Ridder et al. [6]. The SLLE method used class label information in the data when computing neighbors to achieve higher classification accuracy. However, when a new point is to be classified, the method proposed in [6] meets some difficulties since the class label of the point is unknown. To improve on this kind of SLLE algorithm, a new SLLE method based on Distance Metric Learning is proposed which overcome the shortcomings in [6], and experiment results shows that our method results in more effective feature reduction and higher classification accuracy.

This paper is organized as follows. In Sect. 60.2 the nonlinear dimensionality reduction algorithms, i.e. LLE and SLLE, are reviewed briefly. Section 60.3 introduces the principle of Distance Metric Learning. In Sect. 60.4 we present a new supervised Locally Linear Embedding method based on Distance Metric Learning, and the experimental results are presented in Sect. 60.5. Finally, some concluding remark is included in Sect. 60.6.

60.2 Reviews of LLE and SLLE

60.2.1 Locally Linear Embedding

LLE is a new method for feature dimensionality reduction that maps the original high dimensional data to a low dimensional space via the solution of certain constrained optimization problem. The basic idea of LLE for feature reduction is as follows [6]. Let $X = [X_1, X_2, \ldots, X_N] \in \mathrm{R}^{D\times N}$ be a data set of N points which are sampled from some underlying manifold whose intrinsic dimensionality is $d(d<D)$. Assuming the corresponding N points in the embedding space are denoted by $Y = [Y_1, Y_2, \ldots, Y_N] \in R^{d\times N}$, then the LLE method tries to compute Y such that nearby points in X corresponding nearby points in Y.

The procedure for computing Y from X is as follows. First, for each data point X_i, find a given number k of nearest neighbors according to Euclidean metric, and uses these k data points to approximate X_i via linear combination. The optimal weight (also called reconstruction weight) W_{ij} for the linear combination minimize the following quantity

$$\left\| X_i - \sum_{j=1}^{N} W_{ij} X_j \right\|^2 \tag{60.1}$$

and satisfy the following constraints:

$$\begin{cases} \sum_{j=1}^{N} W_{ij} = 1 & \text{if } X_j \in N_i(X_i) \\ W_{ij} = 0 & \text{if } X_j \notin N_i(X_i) \end{cases} \tag{60.2}$$

where $N_i(X_i)$ denotes the k nearest neighbors of point X_i. Note that here we need only solve for k variables, since for $X_j \notin N_i(X_i)$, we have $W_{ij} = 0$, and we need only solve for those W_{ij} where $X_j \in N_i(X_i)$. Clearly, the minimization problem is a constrained least squares problem, and can be easily solved using traditional optimization method. After we have computed the weigh W_{ij} for each data point X_i, we can obtain the matrix of reconstruction weights $W = [W_{ij}]_{N \times N}$.

Using the reconstruction weights matrix W, we can now compute the low-dimensional embeddings Y which best preserve the local geometry properties represented by the reconstruction weights W. Such a data set Y should minimizes the following quantity

$$\sum_{i=1}^{N} \left\| Y_i - \sum_{j=1}^{k} W_{ij} Y_j \right\|^2 \tag{60.3}$$

and satisfies the following two constraints:

$$\begin{cases} \sum_{i=1}^{N} Y_i = 0 \\ \frac{1}{N} \sum_{i=1}^{N} Y_i Y_i^T = I \end{cases} \tag{60.4}$$

where I is the $d \times d$ identity matrix. The optimization problem can be solved as follows. First, rewrite objective function as follows:

$$\min_{Y} \; tr(YAY^T) \tag{60.5}$$

where $A = (I - W)^T (I - W)$. Then computes the bottom $d + 1$ eigenvectors of A, where the corresponding eigenvalues are arranged in descending order. Exclude eigenvector whose eigenvalue is close to zero, then remaining d eigenvectors yield the final embedding Y.

60.2.2 Supervised Locally Linear Embedding

LLE is usually used in unsupervised learning tasks and uses no class label information. For supervised learning tasks, if the class label information were used, the recognition accuracy should be improved. The main purpose of SLLE is to extend the LLE method to supervised learning tasks, by using the label information contained in the data set [5–7]. A basic idea of current SLLE methods is to increase the distance between samples X_i and X_j from different classes, while keeping the distance unchanged for samples X_i and X_j from the same class. Let the data set be $X = [X_1, X_2, \ldots, X_N] \in R^{D\times N}$ and let

$$\Delta = \max d(X_i, X_j), i, j = 1, \ldots, N \tag{60.6}$$

where d is Euclidean distance in R^D. Then the new distance Δ_{ij} between X_i and X_j is defined to be

$$\Delta_{ij} = d(X_i, X_j) + \alpha\Delta\Lambda_{ij}, \quad \alpha \in [0, 1] \tag{60.7}$$

where $\Lambda_{ij} = 1$ if X_i and X_j belong to different classes, and $\Lambda_{ij} = 0$ otherwise. Note that this formulation takes class information into account, and the parameter α controls the amount to which the class information is used in calculating the new distance. When $\alpha = 0$, SLLE is equivalent to the original unsupervised LLE; when $\alpha = 1$, the result is the fully supervised LLE. Varying α between 0 and 1 gives a partially supervised LLE (called α-SLLE). In practice, the optimal value of α should be determined via a validation set. Hence this new distance on X is defined, we can then use it instead of the Euclidean distance to carried out LLE procedure outlined in the above section to reduce the data dimensionality.

60.3 Distance Metric Learning

The SLLE method discussed above in fact has some shortcomings that will be examined in detail in Sect. 60.4. To propose our new SLLE method that overcomes these shortcomings, we will rely heavily on Distance Metric Learning. The basic idea of Distance Metric Learning is to learn a new distance metric from the given data set X [8–12]. Experiment results have shown that a learned metric can significantly improve the performance in classification and clustering tasks. For supervised learning tasks, we have class labels for every data point in the data set X and a distance metric is learned as follows. For any pair of data points (X_i, X_j) of the same class label, we assign the pair to the set S, and for any pair (X_i, X_j) of different class labels, the pair is assigned to the set D. The sets S and D are in fact relations on X and if we consider two points from the same class as similar, and two points from different classes as dissimilar, then S is consisted of all pairs of

similar points, while D is consisted of all pairs of dissimilar points. Then the distance between two data points X_i and X_j is given by

$$d_M(X_i, X_j) = \left\|X_i - X_j\right\|_M = \sqrt{(X_i - X_j)^{\mathrm{T}} M (X_i - X_j)} \tag{60.8}$$

where the matrix M is learned from the data set X by solving the following constrained optimization problem [9]:

$$\max_M \sum_{(X_i, X_j) \in D} \left\|X_i - X_j\right\|_M \tag{60.9}$$

$$s.t. \quad \sum_{(X_i, X_j) \in S} \left\|X_i - X_j\right\|_M^2 \leq 1$$

$$M \succ 0$$

Note that the matrix M is semi-positive definite, so d_M is a valid distance metric [8]. Also note that the optimization problem is convex, so we can solve for M by semidefinite programming (SDP) or Gradient ascent method. The computational complexity for solving M using these method is usually very high for large data set, so a simplified method LMNN is proposed in [10] for solving M.

Note that M can be written as $M = L^T L$. So we can write (60.8) as

$$\begin{aligned} d_M(X_i, X_j) &= \sqrt{(X_i - X_j)^T M (X_i - X_j)} \\ &= \sqrt{(X_i - X_j)^T L^T L (X_i - X_j)} \\ &= \sqrt{(LX_i - LX_j)^T (LX_i - LX_j)} \end{aligned} \tag{60.10}$$

Let $Z_i = \mathrm{L}(X_i), i = 1, \ldots, N$, and let $Z = [Z_1, Z_2, , Z_N] \in R^{D \times N}$, then we have:

$$d_M(X_i, X_j) = d(Z_i, Z_j) \tag{60.11}$$

where d is the Euclidean distance on R^D. In the following, we will propose a method to combine the metric learning method with LLE, using both the matrix M learned from the data and the transformation L which can be obtained from M.

60.4 Supervised Locally Linear Embedding Based on Distance Metric Learning

The LLE method gives no direct mapping from the input space to the low dimensional embedding space. According to (60.3), the set Y of lower dimensional data points corresponding to the original set of high dimensional data points is obtained via the solution of an optimization problem. Thus, supposing we need to compute the output Y_0 corresponding to a new input X_0, we need in principle rerun

the entire LLE algorithm with the original data set augmented by X_0. This property of LLE raises certain difficulties for the current SLLE methods introduced in Sect. 60.3, when a new data point is to be classified. Let X be the original training set with label information from which the distance Δ_{ij} given in (60.7) is learned, and let X_0 denote the new data point to be classified. To classify X_0 using these SLLE methods, we first need to find the corresponding data point Y_0 in the embedding space, which entails the calculating the k nearest neighbors of X_0 according to the distance given in (60.7). However, since the new data point X_0 has no label information, we are not able to use this distance properly. This is the main drawback in the methods proposed in [13, 14]. To overcome this shortcoming of the current SLLE methods, we proposed a new SLLE method based on distance metric learning, hereafter called SLLE-DML.

The basic idea of our method is as follows. Given the data set $X = [X_1, X_2, \ldots, X_N] \in R^{D\times N}$ with label information, we first obtain the matrix M by solving the optimization problem (60.9), and then obtain the matrix L using the fact that $M = L^T L$. We then consider the matrix L as a linear transformation from R^D to R^D, and transform the original data set X to Z, where $Z = [Z_1, Z_2, \ldots, Z_N] \in R^{D\times N}$, $Z_i = L(X_i), i = 1, \ldots, N$. Then, we run the LLE algorithm on Z to obtain the corresponding points $Y = [Y_1, Y_2, \ldots, Y_N] \in R^{d\times N}$ and design a classifier using Y as the training set. For the new data point X_0 to be classified, we first transform X_0 to $Z_0 = L(X_0)$, and run the LLE algorithm on $Z' = [Z_0, Z_1, \ldots, Z_N]$ to calculate $Y' = [Y_0, Y_1, \ldots, Y_N]$, and then we can determine the class of X_0 by classify Y_0 using the designed classifier.

60.5 Experiments and Results

To verify the efficacy of the proposed feature reduction method, we test it using a 1NN classifier with three UCI data sets: Iris, wine and pendigits. The data sets descriptions are shown in Table 60.1. Each of the data sets is randomly split into a training set (80 %) and a testing set (20 %). For comparing the effectiveness of the proposed method, experiments are also performed using LLE as well as SLLE methods proposed in [3, 6] using 1NN classifier. Table 60.2 presents classification error rates with 3 different feature dimension reduction rate (25, 50 and 75 %). From the results, we can see that our method SLLE-DML obtains a higher classification accuracy than LLE and SLLE.

Table 60.1 The data set descriptions

Data set	Number		Dimension	Class
	Train	Test		
Iris	120	30	4	3
Wine	141	37	13	3
Pendigits	7494	3498	16	10

Table 60.2 The classification error rate of each method to the data sets with 1NN classifier

Data set	LLE			SLLE			SLLE-DML		
	75 %	50 %	25 %	75 %	50 %	25 %	75 %	50 %	25 %
Iris	30.00 %	20.00 %	10.00 %	6.67 %	3.33 %	3.33 %	3.33 %	3.33 %	3.33 %
Wine	26.31 %	23.68 %	10.53 %	26.31 %	13.16 %	10.53 %	18.42 %	13.16 %	5.26 %
Pendigits	7.24 %	5.31 %	4.89 %	1.98 %	1.81 %	1.74 %	2.21 %	1.69 %	1.03 %

60.6 Conclusion

A new feature reduction method is proposed in this paper. It combines distance metric learning and LLE learning to achieve efficient feature reduction for high classification performance. In method, a distance metric learning procedure is first applied to increase the class separability, and then LLE learning is implemented to reduce feature dimensionality. Experiment results using UCI data set shows that the method not only overcomes the shortcomings in the current SLLE learning methods, but also achieves higher classification accuracy.

Acknowledgments This research is supported by NSFC (NO. 60903123).

References

1. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd ed. Academic, New York (1991)
2. Borg, I., Groenen, P.: Modern Multidimensional Scaling. Springer, New York (1997)
3. Saul, L.K., Roweis, S.T.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**, 2323–2326 (2000)
4. Tenenbaum, J.B., Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**, 2319–2323 (2000)
5. de Ridder, D., Duin, R.P.W.: Locally linear embedding for classification. Technical Report PH-2002-01, Delft University of Technology, Delft, The Netherlands (2002)
6. de Ridder, D., Kouropteva, O., et al.: Supervised locally linear embedding. ICANN, pp. 333–341 (2003)
7. Kayo, O.: Locally linear embedding algorithm extensions and applications. Faculty of Technology, University of Oulu (2006)
8. Schultz, M., Joachims, T.: Learning a distance metric from relative comparisons. Adv. Neural Inf. Process. Syst. **16**, 41–48 (2004)
9. Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.: Distance metric learning, with application to clustering with side information. Adv. Neural Inf. Process. Syst. **15**, 505–512 (2003)
10. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res. **10**, 207–244 (2009)
11. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: Proceedings of International Conference on Machine Learning, Corvallis, Oregon, USA, pp. 209–216 (2007)
12. Ying, Y., Li, P.: Distance metric learning with eigenvalue optimization. J. Mach. Learn. Res. **23**, 1–26 (2012)

13. Zhang, S.: Enhanced supervised locally linear embedding. Pattern Recognit. Lett. **30**, 1208–1218 (2009)
14. Kouropteva, O., Okun, O., Pietikainen, M.: Supervised locally linear embedding algorithm for pattern recognition. Lect. Notes Comput. Sci. **2652**, 386–394 (2003)

Chapter 61
An Autonomous Rock Identification Method for Planetary Exploration

Chen Gui and Zuojin Li

Abstract One of the goals of planetary exploration is to cache rock samples for subsequent return to Earth in the future Mars Sample Return (MSR) mission. Rock segmentation is significant for the achievement of MSR mission and its scientific studies. The paper presents a new approach to detect and determine the locations of the rocks in the pair of images. This new method consists of two major processing procedures: Rock Rough Boundary Detection and Template Dilatation Edge Linking (TDEL). The first processing block rock rough detection has been designed to find a rock's closed rough contour using the OTSU algorithm (maximum between-class variance method) for background removal and using Canny algorithm for discontinuous contours of the rocks from the original stereo image pairs. TDEL is an algorithm of edge linking for obtaining accurate contours of the rocks in the image of multiple levels of a multi-scale image pyramid. Current work is in preparation for the eventual grasping of a rock using our developed technique. Researchers are currently testing the program code (OpenCV) by using image data acquired from the Aberystwyth University Pan-Cam Emulator (AUPE) instrument of the Trans-National Planetary Analogue Terrain Laboratory (PATLab). Experimentation results are presented and show the validity of the method which can effectively detect rocks in this paper.

Keywords Segmentation · Canny · Machine vision · Autonomous · Planetary exploration

C. Gui (✉) · Z. Li
College of Electrical and Information Engineering,
Chongqing University of Science and Technology,
Chongqing, China
e-mail: guic333@yahoo.com.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_61,

61.1 Introduction

The Mars Sample Return (MRS) mission plans to collect samples of Martian rock, soil and gas for returning to Earth and carrying out scientific analysis [1, 2]. In Mars rover missions ExoMars is the forthcoming ESA/Roscosmos 2016 and 2018 missions, which can be regarded as precursor missions to MSR. The first mission which is to carry a Trace Gas Orbiter and an Entry, Descent and Landing Demonstrator Module (EDM) will be launched and reach Mars in 2016. The second mission will carry a large capsule with a surface science platform and a rover to Mars in 2018.

In terms of rock segmentation research, several automated approaches and algorithms have been produced to find out rocks. For example, recently Castno proposed a method to detect the closed contours of rocks combining an edge detector with multi-resolution images [3]. The rock detection algorithm is just efficient while intensity sharp differences between soil and rocks are significant to reveal clearly linked boundaries. Methods employing a belief network based on a machine learning approach classify homogeneous regions from colour images [4]. Nevertheless, difficulty remains that a rock may have nonhomogeneous intensity and color, which transforms in terms of the illumination and geometry of the rock surface. Dunlop proposed an approach applying a normalized-cut strategy to fragment the original image into superpixels and then to merge them into rock regions [5]. However, several issues such as training set determination and boundary localization still remain. Thompson conducted a comparison for the performance of several rock detection algorithms [6]. A method using the texture-based rock segmentation and the edgeflow-based boundary refinement is proposed by Song [7]. The algorithm suffers from heavy computation. Pugh and Barnes also propose an approach for rock identification in the context of segmentation [8]. The difficulty of the method is merge threshold determination. Shang and Barnes have constructed image classifiers using Fuzzy-rough feature selection (FRFS) combined with Support Vector Machines (SVMs) [9]. The techniques have preferable classify result to the similar images.

In this study, a novel and uncomplicated method is proposed to achieve the segmentation of rocks in the pair of images. The experimentation results show that the segmentation results of the proposed approach are well consistent with human perception.

61.2 Approach

The general issue we attempt to address is the identification and matching of the scientifically interesting rocks (such as sedimentary, metamorphic and igneous rock) since these rocks are able to tell scientists about the history of geology and environment conditions on Mars. The process is broken down into a couple of

steps: Rock Rough Boundary Detection and Template Dilatation Edge Linking (TDEL). The first of these two steps, Rock Rough Contours Detection, is an image segmentation problem, and is a difficult task to extract rocks in an image, because current visual segmentation techniques poorly cope with the segmentation of rocks that ordinarily manifest different morphologies. Regularly it is very difficult to discriminate them from the background ground or soil. Here we address this challenge with a Canny and OTSU based contour segmentation method to extract the rough contours of the rocks in the images taken from the field of our Trans-Nation Planetary Analogue Terrain Laboratory (PATLab) [10]. The remaining step is getting the accurate closed contours of the rocks using TDEL algorithm. A generalization of our algorithm for rocks identification can be found in Fig. 61.1, and the following subsections describe the above steps in detail.

61.2.1 Target Identification

Target Identification is the first stage of the process of rock sample acquirement, which includes two steps: rough bock boundary detection and accurate rock boundary detection.

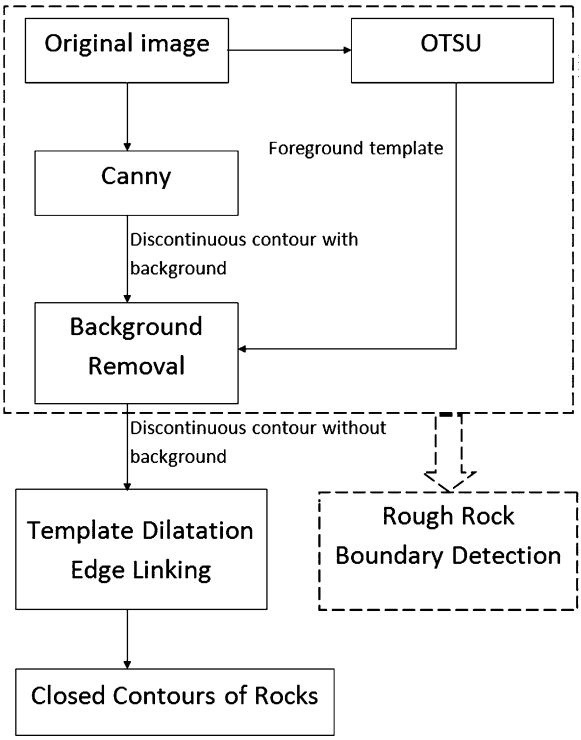


Fig. 61.1 Flowchart of matching keypoints procedure

61.2.1.1 Detection of Rough Rock Boundary

We assume the identified targets/rocks to be extracted from a stereo pair of initial navigation camera images are scientifically interesting. That said, target identification can be subdivided into a series of stages. First, it is desirable to employ the Canny algorithm to obtain discontinuous contours of the rocks from the original stereo image pairs. The Canny edge detector is an edge detection operator that uses a multi-stage algorithm to detect a wide range of edges in images. Figure 61.2 illustrates a stereo pair of original images. The Canny detector is used for detecting the right image edge. Background removal using the OTSU algorithm has been employed [11], Otsu's method is used to automatically perform histogram shape-based image thresholding or the reduction of a graylevel image to a binary image. The algorithm assumes that the image to be thresholded contains two classes of pixels or bi-modal histogram (e.g. foreground and background) then calculates the optimum threshold separating those two classes so that their combined spread (intra-class variance) is minimal. And a multi-scale method (breaking down into three levels including the original images) are utilized to derive a blob of each rock as an irregular bounding box which surrounds the identificated rock boundary on the top level image whose resolution is the lowest. Within this method, the first step is down-sampled twice to the Canny image, the result of this procedure is an image obtained whose size is quarter of the Canny image. The morphological method of image processing Erode and Dilate are then applied to gain the rough closed rock boundary, which is a coarse region in the top level image. The second step is to enlarge the coarse region once that will become an irregular bounding box encompassing the corresponding rock in the middle image obtained down-sampled by twice from the Canny image. The irregular bounding box is the desired rough rock boundary, which is red closed contour in Fig. 61.3.

Fig. 61.2 A stereo pair of images

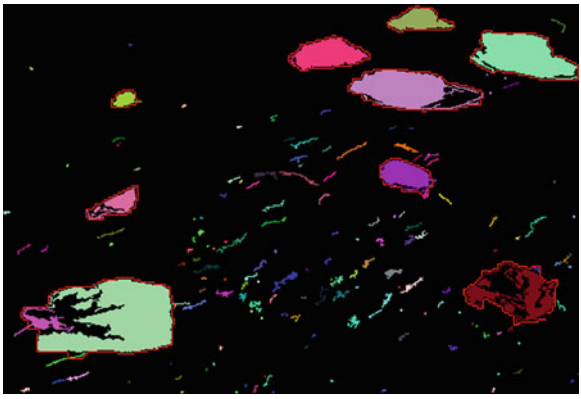


Fig. 61.3 The rocks included in the irregular bounding box. [*Red line* denotes the irregular bounding box]

61.2.1.2 Template Dilatation Edge Linking

A Template Dilatation Edge Linking (TDEL) algorithm is then used for the closed contour of each rock in the middle and bottom level images. The complete algorithm of TDEL is described as follows:

1. Fill the image (e.g. Fig. 61.3) with a black colour except for the regions of the irregular bounding box (Fig. 61.4).
2. Choose a random pixel as the start point on the irregular bounding box.
3. Build a m × m template based on the start pixel (m = 3) (Fig. 61.5).
4. Judgement: Are there other colour pixels in the template? If Yes, (a) If these pixels are the same colour, then sequentially connect using a straight line, (b) If there are different colours in these pixels, then connect the nearest two pixels of the different colour using a straight line. Then to (6); else, repeat. Until all pixels visited.

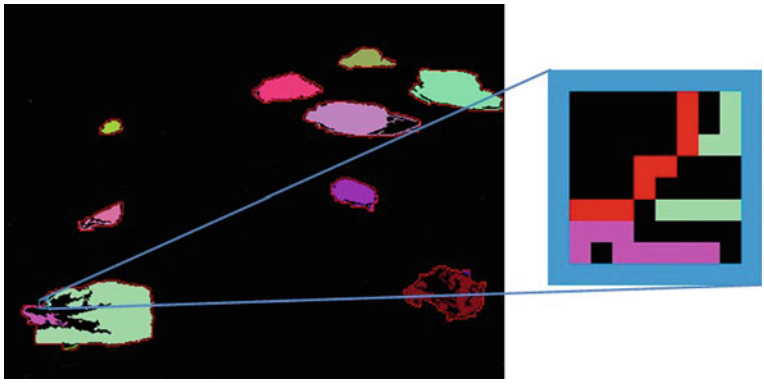


Fig. 61.4 Patch of the image for subsequently explaining the TDEL algorithm. In the patch *red pixels* form a rough rock boundary, the other *colour pixels* are the real rock boundary which would be connected

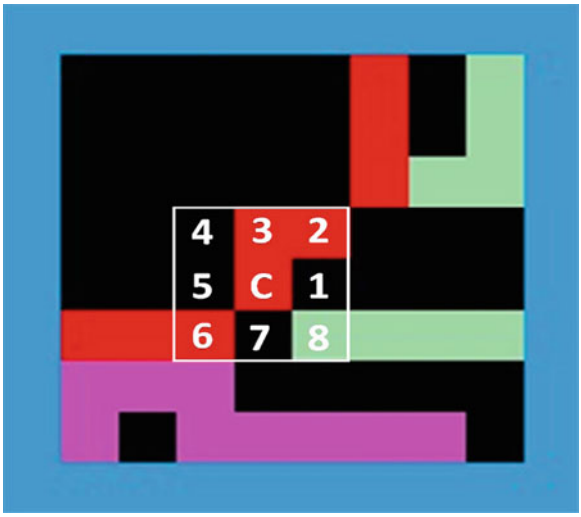


Fig. 61.5 3 × 3 template. 'C' pixel is the center of the template on the rough rock boundary. Here '8' pixel is just different colour except for *red* and *black* colour, so there is not connection performed

5. Dilate the current template (m = m + 2), then to (4) (Fig. 61.6).

6. Choose next neighbour pixel point relative to the current pixel point on the irregular bounding box.

7. If the pixel point is the end of all sequential traversal pixels on the irregular bounding box, then the algorithm end; or else, then to (3).

TDEL is an approach to accurately find edge fragments, and sequentially trace them into a closed contour.

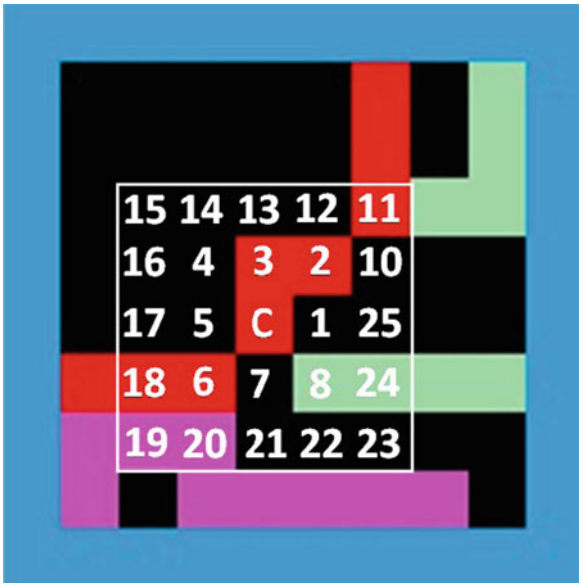


Fig. 61.6 5 × 5 template. The template is dilatation of the above 3 × 3 template. Here '8', '24', '19' and '20' pixels are different colour except for red and black colour, and the colour of '8' and '24' is different from the colour of '19' and '20'. Therefore, we separately calculate the distances of between '8' and '19', '8' and '20', '24' and '19', and '24' and '20'. The nearest distance is from '8' to '20' in these results, so the two pixels are connected using a *straight line*

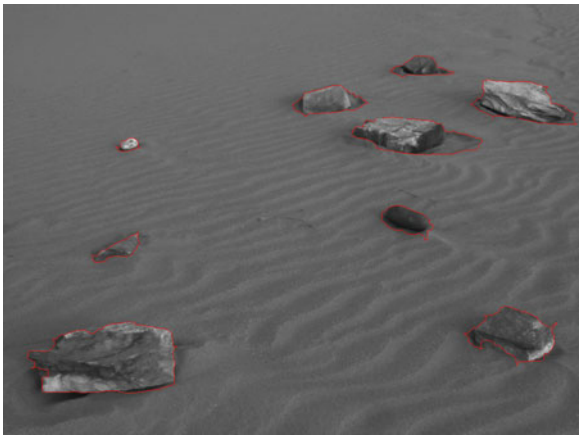


Fig. 61.7 *Red* outlines show each rock's detection results

61.3 Experiment

We are currently testing our program code (OpenCV) using image data acquired from the Aberystwyth University PanCam Emulator (AUPE) [12, 13] instrument of our Trans-Nation Planetary Analogue Terrain Laboratory (PATLab), and we have been able to demonstrate satisfactory rock identification results through achieving the above methods and steps (Fig. 61.7).

61.4 Conclusion

This paper describes researchers' current work of autonomous rock extraction and identification on rock targets for planetary exploration. The accurate closed contours of the rocks are able to be obtained from the image by using this method. Based on the current work, researchers can match arbitrary points on the body of a rock in a pair of images, which are obtained with two wide angle multispectral cameras (WACs) in Trans-National Planetary Analogue Terrain Laboratory (PATLab). It is possible to derive size, shape, and the 3D location of rock through combining these points with appropriate camera parameters in the future works.

Acknowledgments This work is supported in part by Natural Science Foundation Project of CQ CSTC (No. cstc2011jjA40030).

References

1. iMARS Working Group. Preliminary Planning for an International Mars Sample Return Mission, June 2008
2. Space Studies Board, National Research Council. Vision and Voyages for Planetary Science in the Decade 2013–2022. National Academics Press, pp. 6–21 (2011)
3. Castno, A., et al.: Intensity-based rock detection for acquiring onboard rover science. In: Proceedings of the 34th Lunar and Planetary Science Conference (2004)
4. Thompson, D.R., et al.: Automatic detection and classification of geological features of interest. In: Proceedings of IEEE Aerospace Conference (2005)
5. Dunlop, H., et al.: Multi-scale features for detection and segmentation of rocks in mars images. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7 (2007)
6. Thompson, D.R., Castno, R.: A performance comparison of rock detection algorithms for autonomous planetary geology. In: Proceedings of IEEE Aerospace Conference (2007)
7. Song, Y.H.: Automated rock segmentation for mars exploration rover imagery. In: 39th Lunar and Planetary Science Conference, pp. 2043–2051 (2008)
8. Pugh, S., Barnes, D., et al.: Automatic pointing and image capture (APIC) for exmars type mission. In: International Symposium on Artificial Intelligence, Robotic and Automation in Space (i-SAIRAS). August/Sept 2010
9. Shang, C., Barnes, D.: Classification of Mars McMurdo panorama images using machine learning techniques. In: Proceedings of IJCAI Workshop on AI in Space: Intelligence Beyond Planet Earth (2011)
10. Barnes, D., et al.: The Europlanet RI TransNational Access planetary analogue terrain laboratory. In: European Planetary Science Congress, vol. 3 EPSC2008-A-00550 (2008)
11. Otsu, N: A threshold selection method from gray-level histograms. IEEE Trans. Sys. man Cyber. **9**(1), 62–66 (1979)
12. Pugh, S., Barnes, D., et al.: AUPE: A Pancam Emulator for the ExoMars mission. i-SAIRAS 2012 (2012)
13. Barnes, D., Pugh, S., et al.: Multi-Spectral vision processing for the ExoMars 2018 mission. In: 11th Symposium on Advanced Space Technologies in Robotic and Automation (ASTRA 2012), vol. ES-TEC, The Netherlands, pp. 12–14 (2011)

Chapter 62
Recognition of CD4 Cell Images Based on SVM with an Improved Parameter

Yinfeng Liu

Abstract A Recognition method of CD4 cell images and the principle of support vector machines are studied in this paper. The recognition method based on SVM is proposed to solve Microscopic image recognition of CD4 cell with a small sample size and less prior knowledge. An improved parallel grid search method is used to choose the parameters of OSU_SVM classifier, so the computation time is significantly reduced. Cross validation is employed to test the performance of SVM classification. With this new method, the recognition rate of CD4 cell image reaches to 95 % which higher than that of conventional methods. The tests show that the OSU_SVM approach is better than the conventional methods such as Fisher classifier, BP neural networks and LS-SVM.

Keywords CD4 cell · Microscopic image recognition · Support vector machine (SVM)

62.1 Introduction

CD4 + T lymphocyte (hereinafter referred to as CD4 cells) is an important cell in immune system. The quantity of CD4 cells is an important indicator of immune function. Recognition of CD4 Cell Images with image processing technology will be used extensively.

There are many methods concerned with classification [1], such as statistical pattern classification, classification trees and neural networks and so on, but these methods require priori knowledge and certain number of samples. Support vector machine shows the unique advantages and good prospects in solving the small

Y. Liu (✉)
School of Informatics, Linyi University, Linyi, Shandong, China
e-mail: lyfly1977@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_62,

sample, nonlinear and high-dimensional identification problems. In this paper, SVM is applied to classify the cells to resolve the problem which priori knowledge of the circumstances is absent.

62.2 The Basic Principle of SVM

62.2.1 Classification Mechanism of SVM

SVM is a new pattern recognition method based on the statistical learning theory, which is developed in the 1992–1995 [2–4]. SVM is the concrete realization from VC dimension and structural risk minimization principle [1, 5]. It is proposed originally from the optimal classification surface of the linearly separable case.

The case that the two kinds of samples are linearly separable is shown in Fig. 62.1. The solid points and the hollow points are two types of training samples, respectively. H is a classification line to correctly separate the two types of samples. H1 and H2 are the two lines which pass the two type samples. They parallel to the classified line and have the shortest distance to the classified line. The distance between H1 and H2 is named as the class interval (margin). The class interval requires not only the optimal classification boundary separating the two types of error-free samples, but also maximal distance. The former guarantees the ERM, and the latter ensures the smallest confidence interval. Extending to higher dimensional space, the optimal separating lines become the optimal classification surface.

The training samples located on the H1, H2 are called as the support vectors. They are marked with large circles, as shown in Fig. 62.1. Optimal classification surface problem can be expressed as constrained optimization problems. The optimal classification function can be written as follows:

$$\begin{aligned} g(x) &= \operatorname{sgn}\Big(\sum_{i=1}^{n} \alpha_i^* y_i (x_i \cdot x) + b^*\Big) \\ &= \operatorname{sgn}\Big(\sum_{x_i \in \mathrm{SV}} \alpha_i^* y_i (x_i \cdot x) + b^*\Big), \end{aligned} \tag{62.1}$$

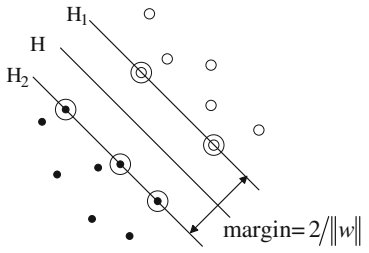


Fig. 62.1 Optimal separating line

where, x is the sample to be classified, SV is the support vector set, b^* is the classification threshold. α_i^* is the Lagrange coefficient, and y is the class label.

When the two samples are linearly inseparable, loose items and punishment factor can be introduced to control the degree of punishment. The right or wrong sub-sample role will achieves a trade-off between the wrong points and the complexity.

For the non-linear classification problem, SVM will turn the input space into a high-dimensional space by a non-linear mapping $k(x)$. Hence, the non-linear problem will turn into a linear problem, and then the optimal separating hyperplane will be calculated in a new high-dimensional space. As the optimization functions and classification functions involve only the inner product between the samples $(x_i \cdot x)$, the transformed higher dimensional space is also just the inner product $(k\ (x_i) \cdot k(x))$. According to the functional theory, if the kernel function $(k\ (x_i) \cdot k(x))$ satisfies with Mercer condition, it corresponds to a transform space of inner product $K\ (x_i, x) = (k(x_i) \cdot k(x))$. The common kernel functions include linear kernel, polynomial kernel and radial basis kernel function. The use of appropriate kernel function can be an alternative to non-linear mapping of a high-dimensional space, which will achieve a linear classification after non-linear transformation. The corresponding classification discriminant function can be obtained as follows:

$$\begin{aligned} g(x) &= \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i^* y_i K(x_i \cdot x) + b^*\right) \\ &= \mathrm{sgn}\left(\sum_{i \in \mathrm{SV}} \alpha_i^* y_i K(x_i \cdot x) + b^*\right) \end{aligned} \tag{62.2}$$

62.2.2 The Training Algorithms of SVM

The training algorithms of SVM have three major categories [6]. The first is quadratic programming algorithm; SVM can be attributed to a quadratic programming problem which includes penalty function solution and the simplex method. The second category is the decomposition algorithm. When the training samples increase, the quadratic programming algorithm will face dimension disaster. Some researchers proposed SVM decomposition training algorithm to deal with the large-scale training set [7, 8], such as the SVM-Light algorithm proposed by Joachims, Platt's SMO algorithm [9]. The third category is the incremental algorithm. S. N. Ahmed first proposed incremental SVM training algorithm. Each time only a small group of the training samples can be dealed with by quadratic programming algorithm are selected. When the samples leave the support vectors and abandon the non-support vectors, the new samples will be mixed into the training samples. According to the references and the samples of CD4 images, OSU_SVM classifier based on LibSVM and LIB-SVM is chosen in this study.

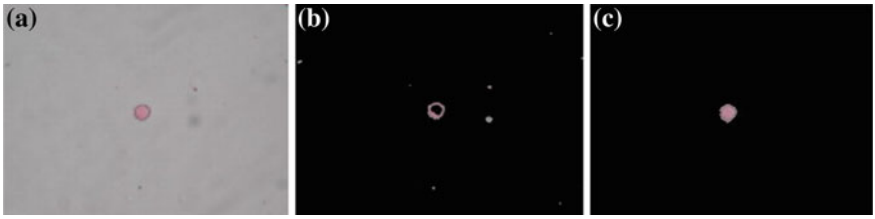


Fig. 62.2 Image preprocessing. **a** Original, **b** KSW entropy method, **c** Image after filling

62.3 Image Recognition Process

62.3.1 Image Acquisition and Pre-Processing

The collection system of CD4 cell microscopic image consists of microscope, CCD camera, image capture card and computer. The collected images are BMP with 24-bit RGB format. Microscopic image pre-processing includes sharpening, image segmentation, clearing small particles and noise removal, filling empty hole after split. The results are shown in Fig. 62.2. According to the characteristics of the image, the results of the KSW entropy method, OTSU (maximum between-class difference) method and the minimum error method are compared [10–12]. Finally we choose the KSW entropy method is selected for segmentation.

62.3.2 Acquisition and Selection of Image Feature

According to the characteristics of CD4 cells including structural integrity, smooth edges, round or shape class circular, single-core, internal lighter in color, cherry red, darker outside, no gray or black impurities in the cell, good light, the 54 features about shape, color, optical density, texture from the target areas are acquired and analyzed.

The original feature dimensions are high, so correlation inevitably exists among the different features, which will increase the complexity and computation of training. The distance among the classes is selected as a separable criterion. The features are selected by increasing L minus R method. At last, several different sets of features are chosen as the input of the classifier.

62.3.3 Classifier Selection and Experimental Methods

The OSU_SVM 3.0 Matlab toolbox [13] developed by Ma and Zhao from Ohio State University is used as a simulation tool. It is based on LibSVM [14] by Chang

and Lin. The toolbox uses the working set selection algorithm of the modified SMO and SVM-Light, which can be applied to areas such as regression and classification estimates.

In the process of the training function SVM works, the parameters can be determined, which includes the type of kernel function and the correlation coefficient, penalty factor C, SVM types and some corresponding weights. In the experiments, c-SVC is selected as SVM type, radial basis function $K(\boldsymbol{x}, \boldsymbol{x_i}) = exp(-\text{Gamma} \times \|\boldsymbol{x} - \boldsymbol{x_i}\|^2)$ is used as the kernel function. Some other random weights can be set randomly, or use the default.

The penalty factor C and the coefficient Gamma of kernel function can be determined by verifying recognition rate of the test samples. In order to preventing overlearning, cross-validation is used in the training process [14]. All the samples are divided into four parts. We choose three parts as the training sample set, the features of the training sample set are normalized. Hence the maximum value of the same features is 1, and the minimum is 0. The normalized parameters of the training samples are used to deal with the other in the test samples. The mean recognition rate of four different test sample sets is selected as the correct recognition rate which corresponding to C and Gamma values.

The range of C and Gamma are very large. The parallel grid search method [14] will spend more time by searching carefully; otherwise lower accuracy by searching roughly, which is difficult to obtain good results. In this paper, the parallel grid search method has been improved, which is divided into coarse search and fine search. Firstly, coarse search is used to obtain the coefficient. Taking into account the breadth of the search range, the value of C is $[2^{-15}, 2^{-14},\ldots,2^{0},\ldots,2^{14}, 2^{15}]$ respectively, and the value of Gamma is $[2^{-20}, 2^{-19},\ldots,2^{0},\ldots, 2^{9}, 2^{10}]$ respectively. When looking for the best combination, set C = a, Gamma = b (if many combinations are the best option, then select the smaller group), narrow the range of search, then the value of C is $[2^{-5}a, 2^{-29/6}a, \ldots, 2^{0}a, \ldots, 2^{29/6}a, 2^{5}a]$ respectively again, and the value of Gamma is $[2^{-5}b, 2^{-29/6}b, \ldots, 2^{0}b, \ldots, 2^{29/6}b, 2^{5}b]$ respectively. And then the fine search will start. The values of C and Gamma corresponding to the highest recognition rate will be obtained. With the mentioned above, the improved parallel grid search method ensures the accuracy and the speed. Time is significantly reduced.

The relationship between the classification recognition rate and the coefficient of C and Gamma is shown in Fig. 62.3, where the number of features is 6.

62.3.4 Results

In the experiment, the 238 samples will be divided into four groups randomly. According to the selected number of features and the above methods, the best values of C and Gamma corresponding to the number of different features are determined by size. These 238 samples are divided into odd part and even part. The odd part normalized is used as the training sample, which will determine the

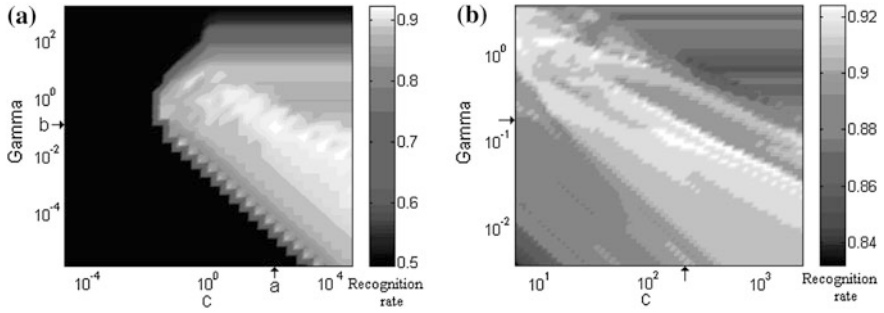


Fig. 62.3 The relationship between the classification recognition rate and the coefficient of C and Gamma. **a** Coarse search, **b** Definite search

values of C and Gamma and train SVM classifier. The even part normalized by the same group coefficient will be tested. The remaining 76 samples after similar treatment will be used as the validation samples. The recognition rate of the test and the validation is shown in Fig. 62.4.

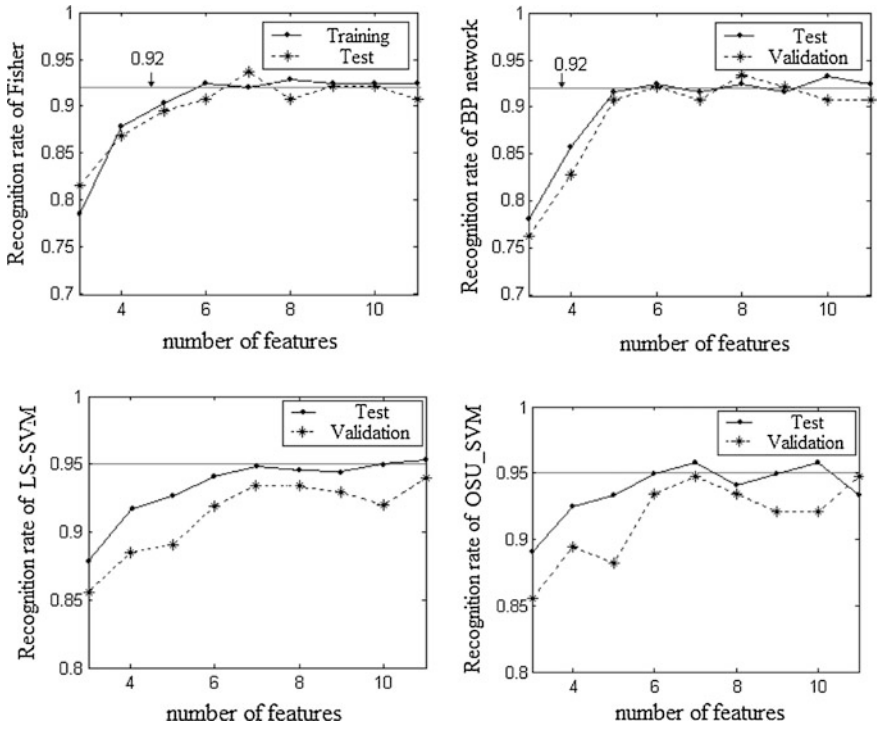


Fig. 62.4 The relationship between the recognition rate of different classifiers and the number of features

When the number of feature grows, the recognition rate of OSU_SVM classifier increases grows, too. It reaches to 95 % approximately when the number of feature is 7. The recognition rate is unstable, but still relatively high when the number of feature increases again.

Using the same samples, the classification experiments are tested by Fisher linear classifier, BP neural network and LS-SVM classifier [15]. The recognition rate of Fisher linear classifier is about 90 % when the number of feature is 7. The recognition rate of BP neural network is instability, up to 91 %. The recognition rate of LS-SVM classifier is about 94 %. All the results of the three classifiers are lower than those of OSU_SVM. Therefore the OSU_SVM classifiers have better classification effect than the Fisher linear classifier, BP neural network and LS-SVM classifier.

62.4 Conclusion

Under the condition of the small number and the big dimension of training samples, it is difficult to recognize them. This article uses the OSU_SVM classifier to identify categories of the CD4 cell microscopic images. Compared with the traditional theory-based classification method, cross-validation and improvement of parallel grid search method based on radial basis function SVM in CD4 cell recognition effect has a higher reliability and superiority. Without too much prior knowledge and skills, SVM (especially OSU_SVM based on LIBSVM) has the advantage of less control parameters. Considering the points depending on the training samples, the complexity of algorithm and dimension-independent, OSU_SVM is more suitable for CD4 cell image recognition classification.

References

1. Sergios, T.: Pattern Recognition [M]. China Machine Press, Beijing (2009)
2. Vapnik, V.N.: The Nature of Statistical Learning Theory [M]. Springer, New York (1995)
3. Vapnik, V.N.: An overview of statistical learning theory [J]. IEEE Trans. Neural Netw. **10**(5), 988–999 (1999)
4. Harman, G., Kulkarni, S.: Statistical Learning Theory and Induction [M]. Encycl. Sci. Learn. **19**, 3186–3188 (2012)
5. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition [J]. Data Min. Knowl. Disc. **2**(2), 121–167 (1998)
6. Moore, G., Bergeron, C., Bennett, K.P.: Model selection for primal SVM [J]. Mach. Learn. **85**(1–2), 175–208 (2011)
7. Paihsuen, C., Rongen, F., Chihjen, L.: A study on SMO-type decomposition methods for support vector machines [C]. IEEE Trans. Neural Netw. **17**, 893–908 (2006)
8. Lin, C.J.: On the convergence of the decomposition method for support vector machines [J]. IEEE Trans. Neural Netw. **12**, 1288–1298 (2001)

9. Platt, J.C.: Fast training of SVMs using sequential minimal optimization. In: Schölkopf, B., Burges, C.J.C., Smola, A.J. (eds.) Advances in Kernel Methods—Support Vector Learning, pp. 185–208. MIT Press, Cambridge [M] (1998)
10. Kapur, J.N., Sahoo, P.K., Wong, A.K.C.: A new method for Gray-level picture thresholding using the entropy of the histogram [J]. Comput. Vis. Graph. Image Process. **29**(2), 273–285 (1985)
11. Otsu, N: A threshold selection method from grey level histogram [J]. IEEE Trans. Syst. Man Cybern. **9**, 62–66 (1979)
12. Kittler, J., Illingworth, J.: Minimum error thresholding [J]. Patten Recognit. **19**(1), 41–47 (1986)
13. Ma, J., Zhao, Y., Ahalt, S.: Matlab toolbox OSU-SVM 3.0 [EB/OL]. http://www.eleceng.ohio-state.edu/maj/osu-svm
14. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 27:1–27:27 (2011)
15. Rui, Z., Wenjian, W., Yichen, M.: Least square transduction support vector machine [J]. Neural Process. Lett. **29**, 133–142 (2009)

Chapter 63
A Method of Automatic Regression Test Scope Selection Using Features Digraph

Yifan Li and Jun Guo

Abstract Regression Testing is an extremely efficient technique used to ensure the quality of changed program. For feature-based systems, the traditional approaches rely on engineer's empirical practice. Therefore, the regression scope will vary from person to person. To improve those unstable methods, the authors presented an algorithm which could automatically select the retest features. The method was based on the building of weighted features digraph and the calculation of dependence values. Meanwhile, the authors proposed some "selection rules" as criteria for these retested features. Besides, a tool "FeaNetwork" had been built to evaluate the performance of our algorithm, and the result was better than empirical methods.

Keywords Regression testing · Regression test selection · Features digraph · Dependence weight

63.1 Introduction

During the cycle of modern software project development and maintenance, it is convinced that the regression testing always play a significant role in keeping product's quality. Its implementation can contribute to ensure the other parts of software work as expected while one part had been changed. Some large and complicated systems are composed of hundreds of features. For each one, a test case suit is executed to validate its function, thus the problem is how to determine which features should be re-tested after one feature's source code have been modified.

Y. Li · J. Guo (✉)
Computer Center, East China Normal University, Shanghai, China
e-mail: jguo@cc.ecnu.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_63,

There are various techniques have been introduced on the study of regression test scope selection, e.g. in the work of Engström et al. [1], they made a systematic summary on this problem's solutions, they put forward that these methods can be distinguished by different application ranges, such as "code changes" [2–5], "specification changes" [6–9], "change impact analysis" [10], specific applications [11], "regression test for GUIs and test automation" [12], and "test process enhancement" [13]. However, most of them focused on the program's source code change or white box test [14].

In this paper, we proposed a method to solve this problem for feature-based systems. First step is finding a way to denote the relationships among all features. A digraph is designed to solve this problem: (1) each node in it represents one feature, (2) the pointed node by an arrow suggests that it is affected by the other end node, (3) the weight value of link indicate arrow-end's dependence degree on transmitter-end. Figure 63.1 shows part of this digraph, in this figure, 0.9 means that feature 2 is heavily depended on feature 1, in terms of the Weight range 0.0–1.0

Second step is to discover those hidden dependence relationship and calculate their values. For the initial features digraph, those confirmable weight values between two nodes need to be set manually, meanwhile, the vast majority of others are still unknown. An algorithm is designed to handle the generations of weight values from hiding relationships. The detailed calculation will be described in Sects. 63.2 and 63.3.

Then, a selection rule will be adopted to find those required regression testing features from improved features digraph. In the paper, we decide to select the weight >0.7 features, which are highly dependent on changed feature, for regression test.

In the rest of this paper, Sect. 63.5 draws a brief summary of our work. We also have implemented our algorithm as a tool and showed its performance in Sect. 63.4.

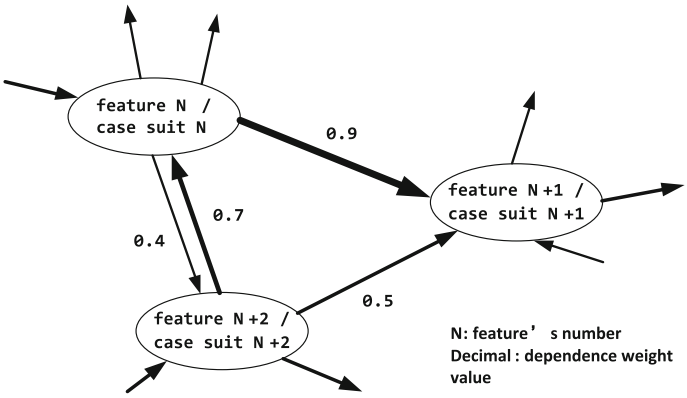


Fig. 63.1 A digraph with weight-values indicates the relationships between one feature and the others

63.2 Materials

63.2.1 Case Library Management

In this paper, the following table is used to manage a big test case library of a program. Each feature of it owns a case suit as a subset of library (Table 63.1).

Index number 1, 2,…, n is generated randomly, they will be used to represent all the features or case suits in rest sections.

63.2.2 Dependence Weight

The weights are given to characterize those dependence degrees between two connected features, or in other words, to indicate the level of influence of start–end feature on arrowhead-end feature.

Weight values are restricted between 0.0 and 1.0.

- 0.0 implies the two features are independent, or have small enough effect on each other, that can be omitted. In digraph, there is no direct linked line to connect them.
- 1.0 is seldom set in features digraph, since it means the arrowhead feature has absolute dependence on the other end. This condition is not-so-common actually.

For initializing a features digraph, some approximate but enough credible weight values will be set up manually based on personal experience.

63.2.3 Data Structure

An n-by-n matrix is established to store the whole features digraph with weight values, that is

$$A = \left[a_{i,j}\right]_{n\times n} \tag{63.1}$$

Table 63.1 A Sample of cross-reference table for features and case suits

Index number	System feature	Case suit
1	Feature xxx	Case #21; case #25; …
2	Feature yyy	Case #13; case #58; …
…	…	…
n	Feature zzz	Case ##; case ##; …

where n is the number of program's features, ai, j is a weight value which describes the dependence from feature i to j.

It will be set 0.0 before manual initialization or automated calculation (see details in Sect. 63.3).

63.2.4 Regression Test Selection Rule

After the construction of matrix A, next step is to decide which features should be involved in regression test scope as the source code of feature i has been changed.

In considering the meaning of $a_{i,j}$, the features subset **S** is selected where each $a_{i,j} > \mathbf{SR}$ for $j \in \mathbf{S}$. **SR** is a predefined value and it can vary by following different workload levels, and its changing, for a specified system, will be indicated using diagram of curve in Sect. 63.5.

63.3 Methods

In Sect. 63.3, we present the details of our method on automated regression test scope selection. The workflow is given in Fig. 63.2.

Then, the core method of calculating two features' dependence value is described. Through this, our final goal is to establish a matrix A^x, save all the dependence values between each two features with the x-length path. Every A^x is executed and all of them are combined to help make retest scope selection, subject to the loop-termination criteria $x > \mathbf{MPL}$. **MPL** is also a predefined value to set the calculation precision, which is also a stable value in a specified system. More details can be found in example section of this paper.

63.3.1 Dependence Values Calculation

The initial weight values of a Features Digraph without links, or, in other words, a Matrix **A** with n^2 0.0 elements, are set manually. After that, we can obtain a sparse digraph or matrix with some nonzero values, and they are the foundation of follow-up execution. Therefore, it is absolutely necessary to use discretion in determining which value should be set and how much the value is. We offer two suggestions for this task:

- To set the element as *Null* if the dependence is unknown.
- To set the element as *0.9* instead of 1.0 in most cases.

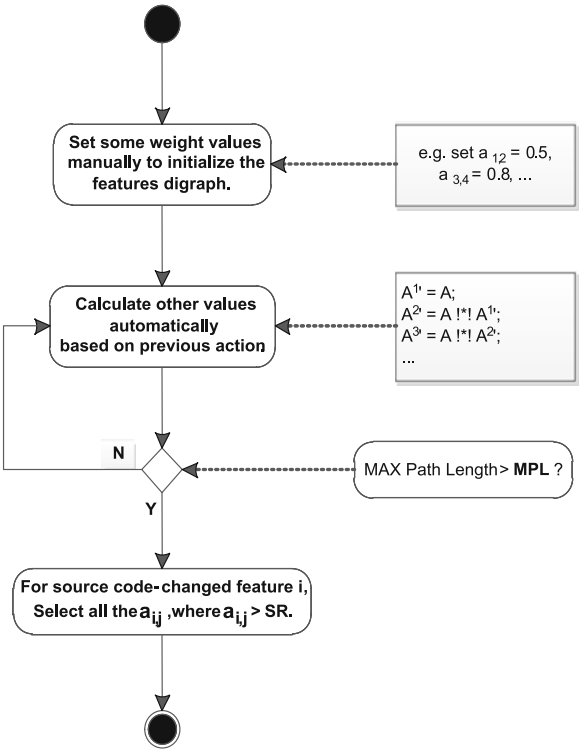


Fig. 63.2 The workflow of our method for automated regression test scope selection

According to the initial Matrix A, we can build an incomplete but trusted Features Digraph. The next aim is to explore other hidden dependence weight values, as these elements in matrix are set 0.0 due to the limitation of human's experience, then to revise and complete the digraph.

In order to solve this question, a fragment of Features Digraph is segmented out to simplify the analysis process. We noticed that the links between two features can be treated as some paths of different lengths. It is shown in Fig. 63.3.

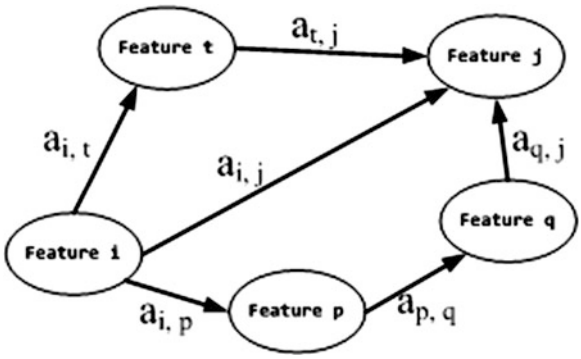


Fig. 63.3 A fragment of features digraph, based on the initial matrix A

For figuring out the dependence value from "Feature i" to "Feature j", using the preset weight values from initial Matrix A and it can be calculated as

$$a'_{i,j} = 1 - \prod MPLx = 1(1 - axi, j) \tag{63.2}$$

where $a'_{i,j}$ is the final result of calculation that dependence from "Feature i" to "Feature j". $a^{x}_{i,j}$ is the dependence value of all the x-length paths between them, the computation method will be introduced in following section.

In particular, $a^{1}_{i,j}$ is defined in initial Matrix A, equal to $a_{i,\ j}$ as shown in Fig. 63.3.

63.3.2 Dependence Value Calculation of x-Length Path

In Features Digraph, multiple x-length paths may exist to make connections from "Feature i" to "Feature j", the calculation formula of x-length path dependence value is defined as below for each path

$$a^{x1}_{i,j} = 1 - \left(1 - a^{x1}_{i,1}\right)\left(1 - a^{x1}_{1,2}\right)\left(1 - a^{x1}_{2,3}\right)\ldots\left(1 - a^{x1}_{x-1,j}\right) \tag{63.3}$$

where 1, 2, …, x−1 is the sequence of all features, except i and j, included in this x-length path "x1", and meanwhile, no more features lie in the path of "Feature n" and "Feature n + 1'', $1 \leq n \leq x - 2$.

For each x-length path "xa" from "Feature i" to "Feature j", a corresponding dependence value $a^{xa}_{i,j}$ is present. Hence the dependence value of all x-length path from "i" to "j" can be computed using the following expression

$$axi, j = 1 - \prod k\, a = 1(1 - axai, j) \tag{63.4}$$

where k is the number of all x-length paths from "Feature i" to "Feature j".

63.3.3 Computation of Matrix

In last paragraphs, the method of dependence calculation by a simple math deduction has been discussed, based on a simple Features Digraph fragment. After that we will introduce the way to obtain dependence values by computing features digraph matrix $A^{x\prime}$, which contains every dependence values from "Feature i" to "Feature j" with the x-length path. It can be defined as

$$Ax' = [axi, j]n \times n \tag{63.5}$$

where x is the path length and n is the number of features.

Especially, $A^{1'}$ is equivalent to A, the initial Features Digraph Matrix.

For $A^{x\prime}$, we can calculate it using the following method

$$A^{x\prime} = A! * !A^{(x-1)\prime} \tag{63.6}$$

where "!*!" is a pre-defined operation principle, such as matrix multiplication. The calculation is given as below

$$A! * !B = \begin{pmatrix} (A! * !B)_{11} & \cdots & (A! * !B)_{1n} \\ \vdots & \ddots & \vdots \\ (A! * !B)_{n1} & \cdots & (A! * !B)_{nn} \end{pmatrix} \tag{63.7}$$

where both A and B are n-by-n matrix, the element $(A!*!B)_{nn}$ is defined as

$$(A! * !B)_{ij} = 1 - \prod n\ k = 1\left(1 - A_{ik}B_{kj}\right) \tag{63.8}$$

According to the Eq. (63.2), the final result matrix A′ consists of all dependence values, each one of which derives from the combination of various length paths, as below

$$A' = \left(a'_{i,j}\right)_{n\times n} = 1 - \prod MPL\,x = 1(1 - A^{x\prime}) \tag{63.9}$$

In Sect. 63.2 of this paper, we introduced the regression test selection rule, and it will be used to determine which feature should be retested as the final matrix A′ has been established.

63.4 Example

We implemented a tool, *FeaNetwork,* to evaluate the performance of our algorithm. FeaNetwork is a command-line program built by Python programming language and all the matrix data is stored in a comma-separated values (CSV) file, which can facilitate the following manipulation and calculation. Also, this file format is easy to be accessed by other applications such as Windows Office Excel, and can be transmitted between database and program.

To demonstrate our method, a feature-based system "basic input output system (BIOS) for Intel servers" was used in the experiment. It is comprised of 89 features and 1207 corresponding test cases. We modified one feature "Memory Slot A1 Status" before the experiment for regression test scope selection. In Fig. 63.4, the comparison of two methods' performance is depicted.

In this experiment, since all the selected features from empirical method would also be chosen by our method and even more untapped ones were neglected due to the limitation of personal intuition, it is always better than, or at least as good as, empirical method.

From the graph, it can be seen that the performance of our algorithm rises with the decrease of SR.

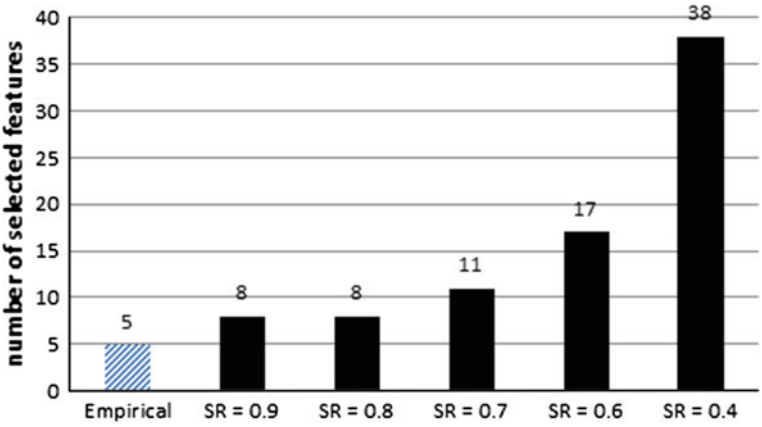


Fig. 63.4 The number of selected features of two methods, where MPL = 10

In first part of Sect. 63.4.1, we took a further discussion on the numbers of failed cases obtained from retests of these selected features by using two methods, it indicated that our algorithm is definitely more effective on bug detection.

63.4.1 Settings of SR and MPL

According the above results, we retested those selected features from two methods, empirical method and FeaNetwork.

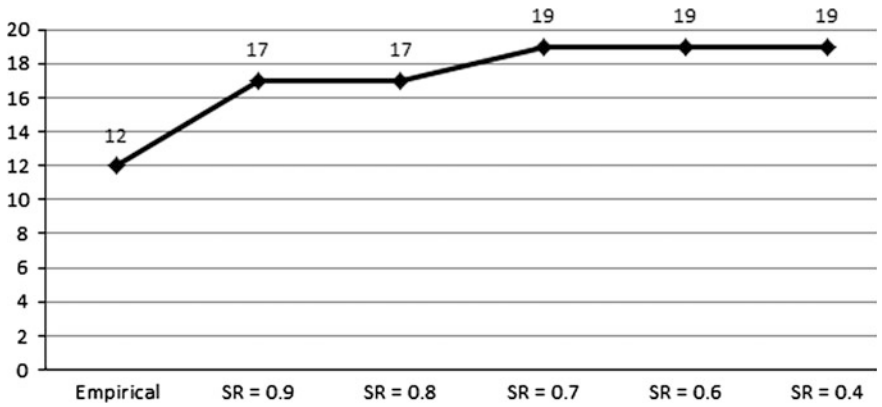


Fig. 63.5 The number of failed test cases during regression tests of two methods, where MPL = 10

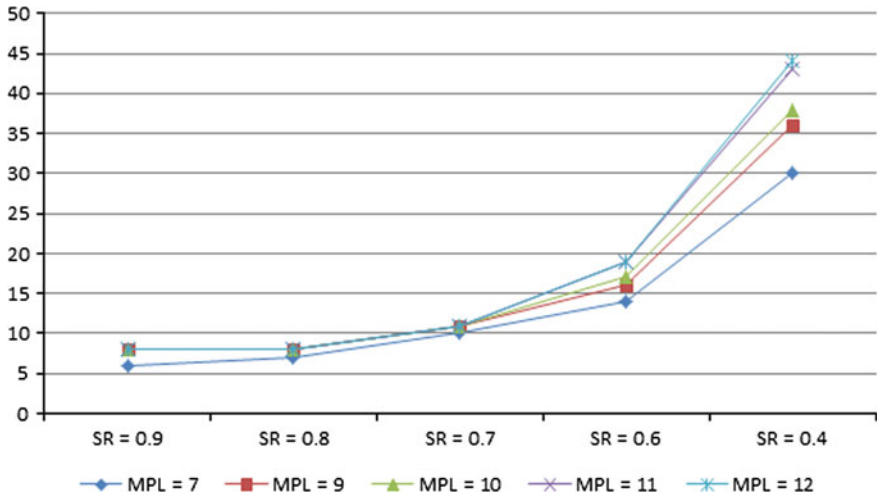


Fig. 63.6 The number of selected features for various MPL

In Fig. 63.5, the trend curve shows that the performance of FeaNetwork will not raise where the SR get 0.7. Therefore we can set SR as 0.7 in the specified experiment. It need to compute a new SR if the tool is applied for another system.

For MPL, we can obtained its value from the below calculations, (Fig. 63.6)

In above graph, it can be figured out that the performance stop rising where MPL = 9 when the SR is set 0.7. Consequently, we can predefine MPL as 9 for the specified experiment, and it needs to be re-computed for other systems.

63.5 Conclusion

In this paper, the authors have presented a new approach for automating regression test selection. It's distinctly different from early techniques on this problem because the method is used for feature-based systems and focused on the inherent dependencies among those features. Meanwhile, the authors proposed a model "Features Digraph" to describe all the features' relationship and designed an algorithm to compute their values.

The established feature digraph is also reusable, even if some feature's source codes are modified. It only needs to be updated as the architecture of system changed, e.g. additions and deletions of features.

Acknowledgments This work is supported by the National Natural Science Foundation of China via the Grant No. 60903092.

References

1. Engström, E., Runeson, P., Skoglund, M.: A systematic review on regression test selection techniques. Inf. Softw. Technol. **52**, 14–30 (2010)
2. Y.-F. Chen, D. S. Rosenblum, K.-P. Vo.: Test tube: a system for selective regression testing. In: Proceedings of the International Conference on Software Engineering, IEEE, Los Alamitos, CA, USA, pp. 211–220 (1994)
3. Gupta, R., Harrold, M.J., Soffa, M.L.: Program Slicing-Based Regression Testing Techniques. Softw. Test. Verification Reliab. **6**, 83–111 (1996)
4. Rothermel, G., Harrold, M.J.: A safe, efficient regression test selection technique. In: ACM Transactions on Software Engineering and Methodology (TOSEM), vol. 6, pp. 173-210 (1997)
5. Wu, Y., Chen, M.-H., Kao, H.M.: Regression testing on object-oriented programs, presented at the International Symposium on Software Reliability Engineering (ISSRE 1999), Boca Raton, FL, USA (1999)
6. Mao, C., Lu, Y.: Regression testing for component-based software systems by enhancing change information. In: Proceedings of the 12th Asia-Pacific Software Engineering Conference (2005)
7. Orso, A., Harrold, M.J., Rosenblum, D., Rothermel, G., Soffa, M. L., Do, H.: Using component metacontent to support the regression testing of component-based software. In: Proceedings of IEEE International Conference on Software Maintenance (ICSM 2001), pp. 716–725 (2001)
8. Sajeev, A.S.M., Wibowo, B.: Regression test selection based on version changes of components. In: Software Engineering Conference, 2003, Tenth Asia-Pacific, pp. 78–85 (2003)
9. Yanping, C., Robert, L.P., Sims, D.P.: Specification-based regression test selection with risk analysis, presented at the 2002 Conference of the Centre for Advanced Studies on Collaborative research (2002)
10. Ren, X.X., Shah, F., Tip, F., Ryder, B.G., Chesley, O.: Chianti: A tool for change impact analysis of Java programs. ACM Sigplan Notices **39**, 432–448 (2004)
11. Haftmann, F., Kossmann, D., Lo, E.: A framework for efficient regression tests on database applications. Vldb J. **16**, 145–164 (2007)
12. Memon, A.M.: Using tasks to automate regression testing of GUIs. In: IASTED International Conference on Artificial Intelligence and Applications (AIA 2004) (2004)
13. Klosch, R.R., Glaser, P.W., Truschnegg, R.J.: A testing approach for large system portfolios in industrial environments. J. Syst. Softw. **62**, 11–20 (2002)
14. Tillmann, N., de Halleux, J.: Pex–White Box Test generation for. NET. Test and Proof. **4966**, 134–153 (2008)

Chapter 64
The Development of the Wireless Vehicle Detector Data Receiving and Processing Host System

Hongyu Li and Yuan Tian

Abstract Vehicle detector is the key technology of Intelligent Transportation System (ITS), and the data receiving and processing host is the important part of wireless vehicle detector. This essay focuses on developing the wireless vehicle detector data receiving and processing host system based on MPS430F1161 ultra-low power microcontroller and nRF905 wireless remote control module. The system is superior to the traditional ring type coil vehicle detection system on performance, real-time, cost, life circle, and maintenance and so on.

Keywords Vehicle detector · Data processing host · MSP430F1611

64.1 Instruction

In China, various vehicles detection equipments have been largely applied to projects [1], which have achieved certain detection effect, but there also exists some inadequacies. For example, infrared detectors cannot resolve the various heat sources and the impact of the work environment, and the anti-noise ability is weak; video detector depends highly on the quality of the background area, which affects detection effect, and has large amount of data operation, high cost [2]. Geomagnetic wireless vehicle detector has been widely used due to its characteristics of free from the influence of environmental conditions, low cost and easy to maintain. Geomagnetic wireless vehicle detector consists of data receiving and processing

H. Li (✉) · Y. Tian
Chang'an University, No.126 South, Erhuannan Road, Xi'an, China
e-mail: lhongyu525@126.com

Y. Tian
e-mail: tianyuan@st.chd.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_64,

host, data measurement host and signal host. Data receiving and processing part accepts data wirelessly and processes data; Data measurement host is used in conjunction with signal host, detects and measures the speed and length of the passing vehicle, converts it into the occupied time which is required by the vehicle to pass the corresponding length detection coil, then sends the data to the processing host; Signal host includes data interface conversion unit and detector input, wireless vehicle detector results are converted from digital to analog coil output by conversion unit.

The data receiving and processing host is the core of the wireless vehicle detector, and its development has important theoretical and practical significance to improve the reliability and stability of vehicle detection products. This paper presents a design plan of data receiving and processing host, which accepts data wirelessly through the data receiver, analyzes data, sets up functional settings and status display by the microcontroller, transfers the analyzed data to the signal host through the interface conversion unit.

64.2 Principle of Work

The data receiving and processing host of this system mainly collects and analyzes data of wireless sensor, then transmits to the signal through the data interface conversion. The receiving and processing host is divided by its function into several parts, including the wireless receiver, host of data processing, interface converters, power supply, status display, function setting and chassis, as shown in Fig. 64.1.

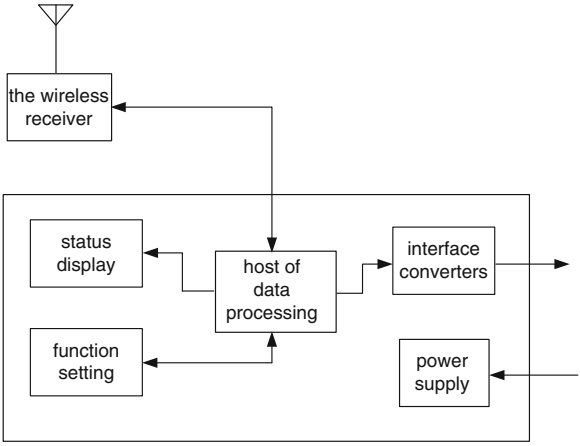


Fig. 64.1 Data receiving and processing host composition diagram

64.3 Whole Design of Receiving and Processing Host

64.3.1 System Design

The system is based on MSP430F1161, and has simple structure. Power is mainly to provide a reliable power supply for the entire circuit; The data receiving host receives data wirelessly from the wireless sensor, uses nRF905 to complete wireless transceiver; Function setting part can set some parameters of the microcontroller; Status display part displays the results of data analysis by LED; The interface converter transmits the analysis of the data and the results to the signal, by RS-485 communication transmission.

64.3.2 Design of Major Modular

64.3.2.1 CPU

The vehicle detector data processing host system uses MSP430F1161 MCU as core part of the controller. MSP430F1161 is an efficient microcontroller with 256 K +256 k bytes of flash, has unique advantage in the low power consumption of battery-powered applications, which operating voltage is of 1.8–3.6 V, power consumption of normal work is about 200 μA, while it can reach 2 μA even 0.1 μA under low-power mode. The MCU has abundance inside resources with 2 KB internal ARM, six programmable input and output ports, two 16-bit timer/counter, two 8-bit parallel ports having the interrupt feature: P1 and P2; four 8-bit parallel port: P3, P4, P5 and P6 [3].

Interface circuit of MCU is very simple, implement interfaces with other circuits by using general I/O ports respectively, picks the wireless receiver data through P6 port, then sets up function through the P4 port, and displays the state of the circuit communication through P5 port.

64.3.2.2 Circuit Design of Wireless Receiver

The wireless receiver mainly adopts nRF905 chip, the nRF905 chip integrates power management, crystal oscillators, low noise amplifiers, frequency synthesizer power amplifier modules. Manchester encoding/decoding is done by the on-chip hardware without use Manchester encoding data, and low power consumption, very convenient to use [4]. The circuit schematic diagram is shown in Fig. 64.2; the antenna portion of the circuit adopts a 50 Ω single-ended antenna according to the needs of the design. The nRF905 transfers data through the SPI interface and microcontroller, receives and sends wireless data by ShockBurstTM transceiver mode; it is reliable and easy to use.

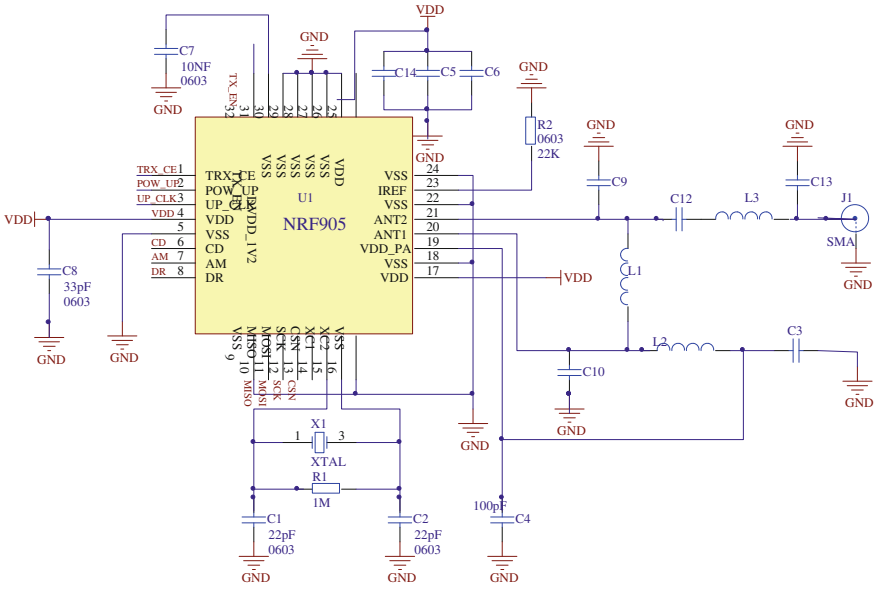


Fig. 64.2 Wireless receiver circuit

64.3.2.3 Power Circuit

Power part of the hardware system employs the LM317 chip due to the voltage supply range of MSP430F1611 microcontroller is 1.8–3.6 V. LM317 is a chip buck, since the entire system uses 3.3 V supply, power supply to the system is achieved by LM317 chip supplying voltage conversion from 220 to 3.3 V. The Power supply circuit is shown in Fig. 64.3.

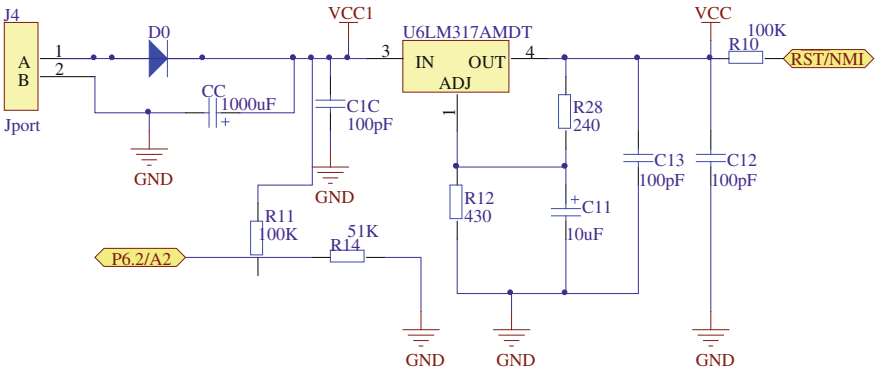


Fig. 64.3 Power supply circuit

64.3.2.4 Status Display and Function Settings

The choice of status display or function setting is made via DIP switches. Function setting can set the input level of corresponding pins which represent the sensitivity, work mode and output mode of the microcontroller, inquiry the key by the microcontroller, thus complete the corresponding handler. Status display part displays the communication status by LED.

64.3.2.5 Communication Circuit

Schematics of serial communication circuit based on SP3485 chip is shown in Fig. 64.4. It is connected to the MCU through terminals J10 and the MSP430F1611 achieve to control by P2 port.

64.4 Programming

System program mainly includes system initialization, wireless receiver module test and major loop control part. The loop control part consists of data processing update module and UART communication module.

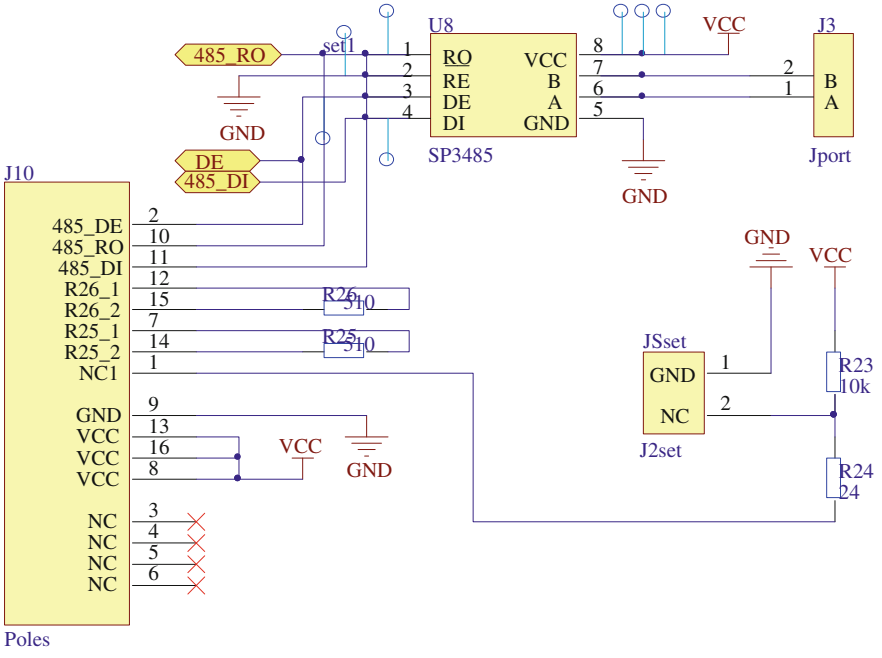


Fig. 64.4 Schematics of communication circuit

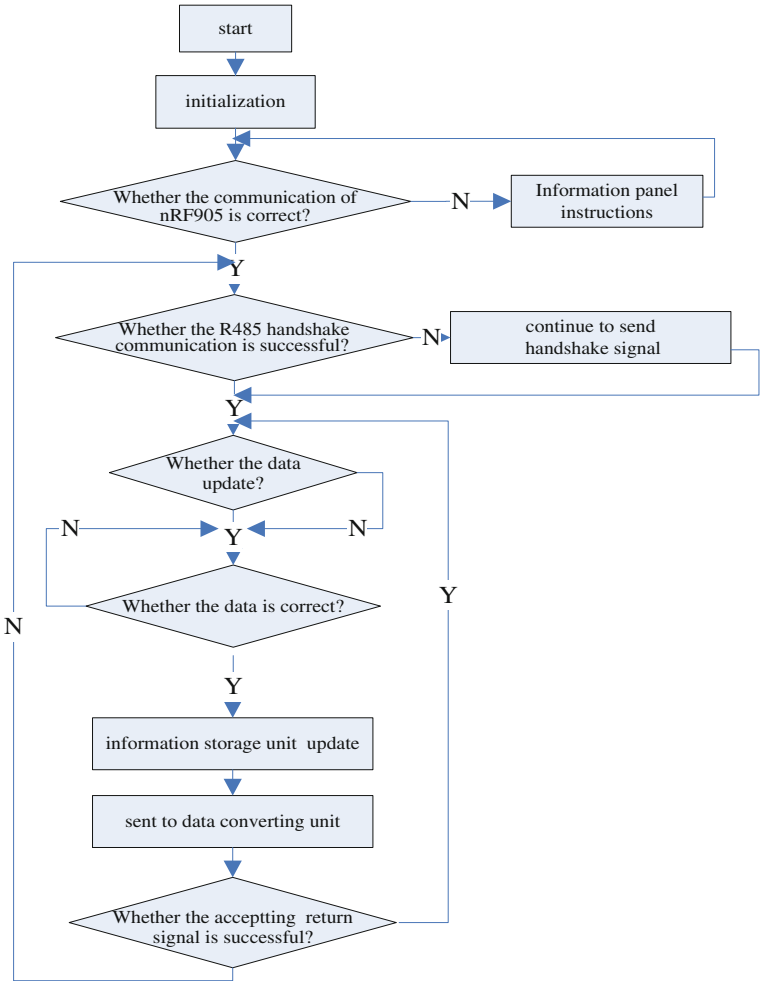


Fig. 64.5 Main program flow

Wireless transceiver module will be initialized once the system activated, if communication of nRF905 is correct, systems enter the loop control part and carry on R485 handshake communication, determines whether the new data is received by data reception flag DR (P2.5) after the handshake succeed, extracts the received data information from nRF905 and makes data validation, if the correct data is received, information storage unit will be updated, then the data be sent to data converting unit through R485, and changes the state of the panel display lamp; If the R485 handshake fail, continues to send handshake signal until it succeed. The state flag of the panel lights will be changed supposing that the wireless communication module test fail. The main program flow is shown in Fig. 64.5.

Table 64.1 Related technical indicators

Traffic flow accuracy	>98 %
Velocity precision	>95 %
Detection radius	< 1.5 m
Battery life	5–8 years
Communication distance	50 m
Operating temperature	−40 to +80 °C

64.5 Experimental Result

The communication distance between data processing host and signal host is 50 meters, and there is no wired connection. Compared with traditional vehicle detector based on C8051f121 [5], test results show that this system is real-time. And it has features of reliable performance, low power consumption and long battery life. Related technical indicators are shown in Table 64.1.

64.6 Conclusion

This paper designs and implements wireless communication based on MSP430F1611 and nRF905. The design obeys the simple and practical design principle, and makes full use of the software and hardware resources of the microcontroller. It not only strives to achieve simple structure to make it easy to control cost, but also attempts to minimize the power consumption.

References

1. Le, L., Ni, P.: Design of vehicle detection wireless node design based on the magneto resisitive sensor. Ind. Control. Comput. 83–84 (2009) (In chinese)
2. Amine, H., Robert, K., Pravin, V.: Wireless magnetic sensors for traffic surveillance. Transp. Res. 13 (2008)
3. Qin L.: MSP430 SCM development and application typical examples, pp. 2–3. China Power Press, Beijing (2005) (In chinese)
4. Jia, Q., Wang, D., Zhang, Z.: Design of wireless data transmission system based on nRF905. Int. Electron. Elements. 29–31 (2008)
5. Junhua, Z.: Design and application of intelligent vehicle detector based on C8051F121. Micro Comp. Inf. 129–130 (2009) (In chinese)

Chapter 65
The Feature Extraction of Rolling Bearing Fault Based on Wavelet Packet—Empirical Mode Decomposition and Kurtosis Rule

Cheng Wen and Chuande Zhou

Abstract The feature extraction method of rolling bearing fault was presented based on the combination of wavelet packet-EMD (empirical mode decomposition) and kurtosis rule. Its first step is to reduce the signal noise by the wavelet packet, and then do EMD decomposition. Based on the characteristic that the kurtosis is very sensitive to impact the biggest IMF component of kurtosis is selected to do Hilbert envelope demodulation. As a result, the fault feature information of rolling bearing was obtained. The implementation process of this method was analyzed by simulation signal and the method was successfully applied in inner race of rolling bearing fault diagnosis.

Keywords Wavelet packet · EMD · Kurtosis coefficient · Rolling bearing

65.1 Introduction

The rolling bearing with local defect would cause pulse impact at runtime so the produced vibration signal has a modulation and non-stationary characteristics. Envelope demodulation and wavelet packet decomposition are the commonly used method in the rolling bearing fault diagnosis [1]. Wavelet packet decomposition has very strong localization analysis ability and can decompose signal into several frequency bands in relatively frequency range. Although wavelet packet decomposition results exist the energy overlapping problem between each frequency band [2], but the wavelet packet can do signal denoising so as to improve the signal-to-noise ratio of the signal. Empirical mode decomposition (EMD) is a new method put forward by American scholar Huang [3], which is suitable for

C. Wen (✉) · C. Zhou
College of Mechanical and Power Engineering, Chongqing University of Science and Technology, Chongqing, China
e-mail: wch2002@126.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_65,

nonlinear and non-stationary signal analysis. It can decompose the complex signal into a series of intrinsic mode function components (IMF) and reveal a complex signal time–frequency characteristic, especially suitable for rolling bearing fault diagnosis analysis. The collected signal from the field has seriously influenced the EMD decomposition quality because of strong noise interference. Therefore, the signal processing by wavelet packet analysis is very necessary before EMD decomposition [4]. Kurtosis coefficient is a dimensionless parameter which can reflect the degree of signal peak effectively [5] and it is very sensitive to the impact. The kurtosis coefficient is bigger the impact composition proportion in signal is more. The representative feature of rolling bearing fault is cyclical impact characteristics. Therefore, the signal processing method based on the combination of wavelet packet-EMD and kurtosis rule was proposed in this paper. This method can efficiently extract the fault characteristic information in rolling bearing fault diagnosis process.

65.2 Wavelet Packet—EMD Decomposition

Rolling bearing fault, especially the earlier fault, would cause bearing element high frequency inherent frequency components and show strong modulation phenomenon. Wavelet packet analysis can decompose the low frequency and high frequency components at the same time, which can make the high frequency band signal have good time domain resolution and greater flexibility in time-frequency analysis [6]. The random multi-scale decomposition characteristics of wavelet packet analysis provide a more elaborate analysis method for signal and it also has the higher frequency resolution. Wavelet packet used in signal denoising can effectively improve the signal-to-noise ratio and the treating process includes wavelet packet decomposition and wavelet packet decomposition reconstruction. Wavelet function and threshold value selection is closely related to the quality of wavelet packet signal denoising.

EMD is based on the time scale features of data itself for signal decomposition, not need to set any advance base function, and has very obvious advantages in the treatment of non-stationary and nonlinear signal [7]. EMD decomposition is a tranquilization processing process and it can decompose the original signal into the sum of limited IMFs. Each IMF component must meet two conditions: (1) the number of the maximum, the minimum is equal to the number of zero-crossing or differ at most 1 from the number of zero-crossing; (2) the upper and lower envelope lines are locally symmetric about time axis.

Suppose a signal is x(t), which is decomposed by EMD and can be expressed by the sum of the IMF component and the residual amount, as shown as Eq. (65.1):

$$x(t) = \sum_{i=1}^{n} c_i(t) + r(t) \quad (i = 1, 2, \cdots, n) \tag{65.1}$$

where, ci(t) is the IMF component, rn(t) is the residual amount (tendency item).

All the IMF components obtained by EMD decomposition are base band signals. These signals are in different band range and include the different information characteristics. The IMF components with obvious fault features were analyzed by Hilbert envelope demodulation [8], and then the fault feature information of equipments can be extracted. Hilbert transform method can Ref. [9–11].

If the signals contain lots of background noise and directly do EMD decomposition without processing a large amount of modal aliasing phenomenon would appear to impact signal analysis process. Therefore, if the signals do EMD decomposition after denoised by using wavelet packet, the EMD decomposition quality would be improved and the accuracy of signal analysis would be improved.

65.3 Comprehensive Analysis of Wavelet Packet-EMD and Kurtosis Rule

65.3.1 Kurtosis Rule

Kurtosis coefficient C_q is a dimensionless characteristic parameter, which is independent of speed, size, and load. Kurtosis coefficient is very sensitive on the impact signal so it is suitable for the fault diagnosis of surface damage, especially the fault diagnosis of early rolling bearing. The mathematical description of the kurtosis coefficient is as shown as Eq. (65.2):

$$C_q = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\sigma^4} \tag{65.2}$$

where, μ is the mean value of signal x(t); σ is the standard deviation of signal x(t).

When the rolling bearing is trouble-free the vibration signals approximately obey the normal distribution and the kurtosis coefficient C_q is about equal to 3. With the emergence of the fault and development impact vibration makes the signals amplitude obviously deviate from the normal distribution. At the same time, kurtosis coefficient is getting higher too. The larger the kurtosis coefficient is, that is to say the greater impact composition proportion of vibration signal is, the more far rolling bearing deviating from its normal state is, the more serious the fault is. The initial failure of rolling bearing is often shown as local spot corrosion of the inner race, outer ring or rolling body because of its working mechanism. The pitting parts and the contacted other components of the bearing would produce impact effect and with the bearing operation a cyclical impact vibration would be produced. Therefore, kurtosis coefficient could detect the impact characteristics of vibration signal and reflect the rolling bearing fault feature information.

65.3.2 Analysis Method of Wavelet Packet-EMD and Kurtosis Rule

The signal analysis process based on wavelet packet-EMD and kurtosis rule is as shown as Fig. 65.1. Analyzed signal $x(t)$ was done EMD decomposition after wavelet packet denoising (including wavelet packet decomposition and wavelet packet reconstruction) and then the kurtosis coefficient of the obtained IMF components by EMD decomposition were calculated. After IMF components were sort based on kurtosis rule the IMF with largest kurtosis coefficient was selected as the main IMF. The main IMF component containing the most obvious fault information was selected to do Hilbert envelope demodulation and then to extract rolling bearing fault feature information.

65.3.3 Signal Simulation of Rolling Bearing

According to the structure characteristics of rolling bearing, the vibration response of single degree of freedom model under a periodic pulse force can be used to simulate rolling bearing vibration. In the small damping conditions, the vibration signal of rolling bearing can be expressed as Eq. (65.3):

$$x_1(t) = \left(\sum_{n=1}^{N-1} D\delta(t - nT)\right) * \left(Ae^{-\xi 2\pi f_n t}\cos(2\pi f_d t - \theta)\right) + N(t) \tag{65.3}$$

where, $\delta(t)$ is the unit pulse force, T is impact period, D is pulse strength, ξ is the damping ratio, f_n is the natural frequency, f_d is damped natural frequency, $N(t)$ is interference signal.

Add sine signal $x_2(t) = 0.02\ sin(40\,\pi t)$ into Eq. (65.3), analysis signal become to $x(t) = x_1(t) + x_2(t)$. When the sampling frequency of signal is 1000 Hz, the time interval of impact composition is 0.04 s, sampling length is 2000 points and $f_d = 100$Hz, the time domain waveform of the simulation signal $x(t)$ is as shown as Fig. 65.2a. If the simulation signal was done EMD decomposition directly the decomposition results, the top five modal components (IMF1–IMF5), were as

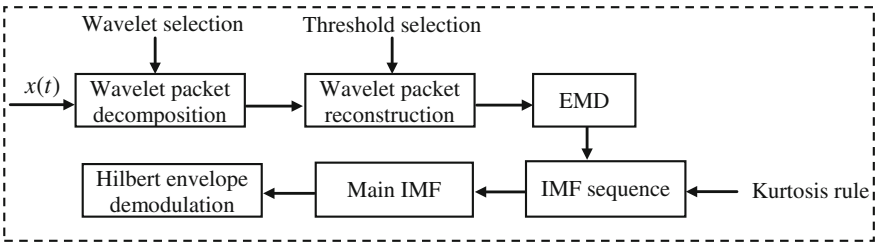


Fig. 65.1 The flow diagram of wavelet packet-EMD and kurtosis rule

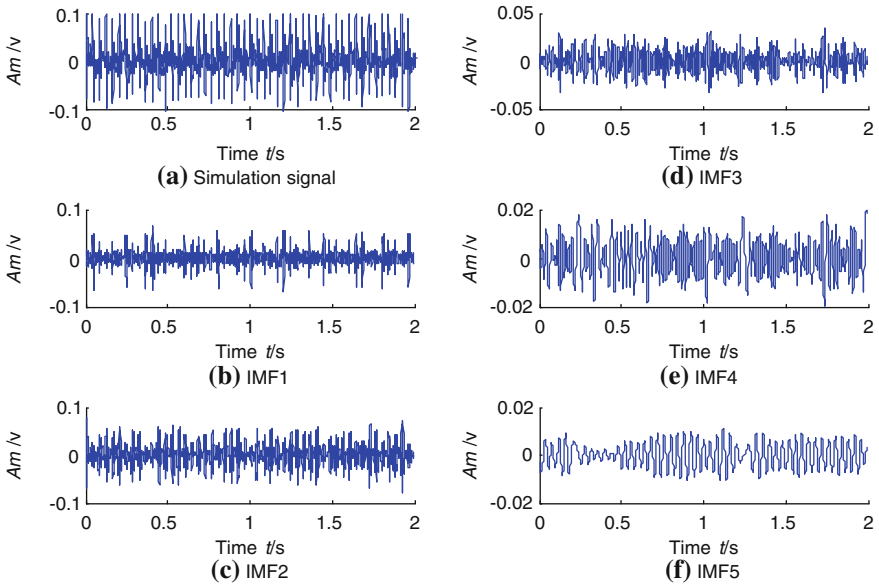


Fig. 65.2 The results of simulation signal decomposed by directly EMD

shown in Fig. 65.2b, c, d, e and f. Figure 65.2 shows that IMF1, IMF2, IMF3 and IMF4 are all high frequency signals, which include the impact composition and noise, and the impact signal has no decomposition to some modal component. The signal in Fig. 65.2f is the sine signal. IMF5 appears obvious modal aliasing phenomenon because of the boundary error accumulation caused by noise and pulse interference, which influences the analysis accuracy of signal.

To reduce the EMD decomposition modal aliasing the simulation signal was done wavelet packet noise reduction and the results were shown in Fig. 65.3b. The EMD decomposition results were shown in Fig. 65.3c, d, e, f and g. The signal in Fig. 65.3g was the sine signal, compared with Fig. 65.2f, its modal aliasing was greatly reduced. Therefore, it is very necessary to reduce noise interference to enhance EMD decomposition quality.

The kurtosis coefficients of modal components were calculated from Fig. 65.3c, d, e, f and g and the results were 5.8600, 4.5792, 1.9070, 1.8245 and 1.6326. The kurtosis coefficient of IMF1 component is the largest, which shows that the frequency band contains rich impact compositions. The envelope spectrum obtained from IMF1 Hilbert envelope demodulation was shown in Fig. 65.3h. The 24.9 Hz frequency component is highlighted in the envelope spectrum and this is consistent with periodic pulse impact (interval 0.04 s) frequency 25 Hz. It reveals the impact composition of simulation signal and provides a basis for rolling bearing fault diagnosis.

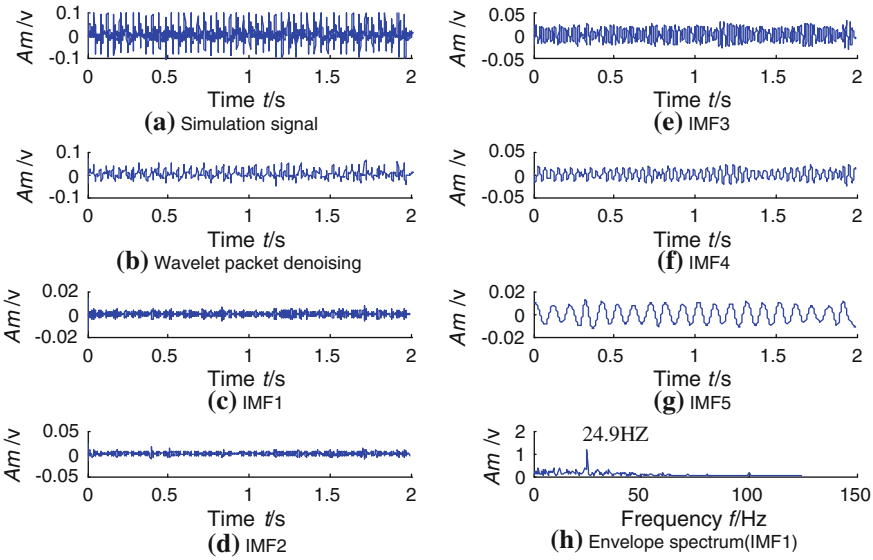


Fig. 65.3 The results of simulation signal processed by wavelet packet-EMD and kurtosis rule

65.4 Case of Rolling Bearing Fault Diagnosis

The experimental data in the case adopt the vibration acceleration signal of induction motor rolling bearing which were provided by Case Western Reserve University. Keep sampling frequency f_s is equal to 12 kHz and the speed n is equal to 1750 r/min, vibration signals of inner race fault of the bearing is measured respectively to obtain the bearing characteristic frequency by calculation. The inner race blade passing frequency is 144.28 Hz.

Wavelet function has a big effect on wavelet packet decomposition quality. Symlets wavelet improved from Daubechies wavelet with approximate symmetry is especially suitable for wavelet packet denoising of rolling bearing signal and the analysis results by Symlets wavelet are better than by the other wavelet function [12]. Based on above analysis Sym8 wavelet was chosen. When the data length N was equal to 2048, the measured signal was done 3 layer wavelet packet decomposition, the denoise signal was done EMD decomposition and kurtosis coefficient calculation and the modal component with obvious impact composition was done envelope demodulation.

Figure 65.4 shows that the failure analysis process of inner race of the bearing. The signal in Fig. 65.4a is the original signal and in Fig. 65.4b is the signal by wavelet packet denoising. The signal in Fig. 65.4b is done EMD decomposition and the top five modal components are obtained, as shown as Fig. 65.4c, d, e, f and g.

The kurtosis coefficient of IMF1–IMF5 was calculated and the results were shown in Table 65.1. The kurtosis index of IMF2 is the largest so it can reflect the vibration characteristics of rolling bearing. The IMF2 is supposed to the main IMF

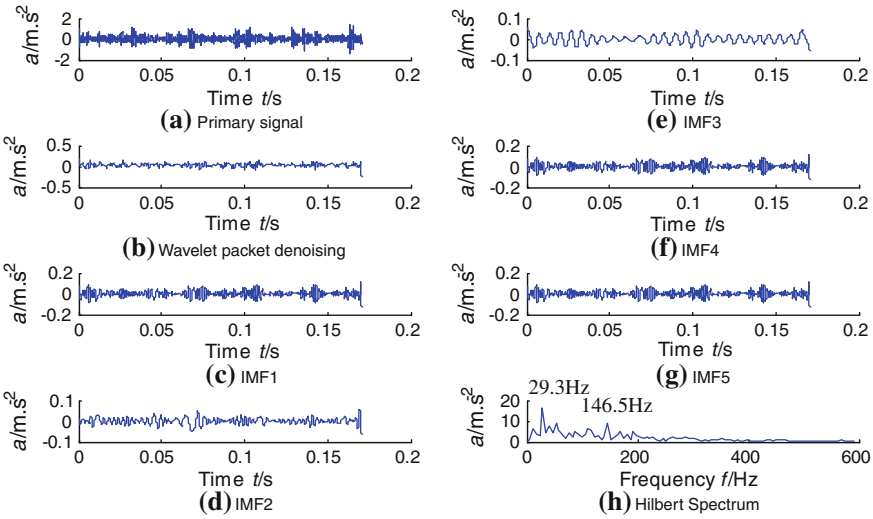


Fig. 65.4 The results of inner race fault analyzed by wavelet packet-EMD and kurtosis rule

Table 65.1 The kurtosis coefficient of IMF1–IMF5 in Fig. 65.4

IMF	IMF1	IMF2	IMF3	IMF4	IMF5
C_q	3.8035	3.9970	2.3165	2.0859	2.5169

component and the envelope spectrum of IMF2 was got by Hilbert envelope demodulation, as shown as in Fig. 65.4h. In the envelope spectrum the 146.5 Hz frequency component appeared obviously. This is similar to 144.28 Hz, the inner race frequency of bearing obtained by theoretical calculation. But the turn frequency is 29.3 Hz, therefore, it can be judged that the inner race have obvious damage.

65.5 Conclusion

Wavelet packet analysis greatly improves the signal analysis accuracy by multiple level band partition and can be used to signal denoising processing.

Empirical mode decomposition (EMD) can obtain a series of intrinsic mode function (IMF) represented signal characteristic time scale. But the noise would affect EMD decomposition quality; therefore, EMD decomposition is very necessary to do signal denoising.

The signal analysis method based on Wavelet packet—EMD and leptokurtosis rule has used the wavelet packet denoising, EMD decomposition and leptokurtosis maximum principle to extract rolling bearing impact characteristics. Numerical

simulation and bearing fault diagnosis case show that this method can improve the accuracy of fault diagnosis. At the same time, it provides a new method for roller bearing fault feature extraction and has certain application prospect.

Acknowledgments The work is supported by Chongqing Scientific and Technological Program (No. CSTC2012 gg-yyjs70012).

References

1. Hui, Z., Shujuan, W., Qingsen, Z.: Research on fault diagnosis of rolling elements bearing based on wavelet packets transform. J. Vib. Shock **23**(4), 127–130 (2004)
2. Lu, S., Xiaojun, Z., Wenbin, Z.: Fault diagnosis of rolling element bearing based on morphological filter and grey incidence. J. Vib. Shock **28**(11), 17–20 (2009)
3. Huang, N.E., Shen, Z., Long, S.R.: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc. R. Soc. Lond. Ser. **454**, 903–905 (1998)
4. Baoping, T., Yonghua, J., Xiangchun, Z.: Feature extraction method of rolling bearing fault based on singular value decomposition-morphology filter and empirical mode decomposition. J. Mech. Eng. **46**(5), 37–42 (2010)
5. Aijun, H., Wanli, M., Guiji, T.: Rolling bearing fault feature extraction method based on ensemble empirical mode decomposition and kurtosis criterion. Proc. CSEE **32** (11), 106–111 (2012)
6. Xiaofeng, L., Shuren, Q., Lin, B.: Wavelet packet analysis—based empirical mode decomposition and its application. China Mech. Eng. **18**(10), 1201–1204 (2007)
7. Yan, R., GAO, R.: Hilbert-Huang transform based vibration signal analysis for machine health monitoring. IEEE Trans. Instrum. Meas. **55**(6), 2320–2329 (2006)
8. Lei, Y.: Machinery fault diagnosis based on improved Hilbert-Huang transform. J. Mech. Eng. **47**(5), 71–77 (2011)
9. Loh, C.H., Wu, T.C., Huang, N.E.: Application of the empirical mode decomposition-Hilbert spectrum method to identify near-fault ground-motion characteristics and structural responses. Bull. Seismol. Soc. Am. **91**(5), 1339–1357 (2001)
10. Huang, N.E., Shen, Z., Long, S.R.: A new view of nonlinear water waves: the Hilbert spectrum. Annu. Rev. Fluid Mech. **31**(1), 417–457 (1999)
11. Boashash, B.: Estimating and interpreting the instantaneous frequency of signal. IEEE Trans. Signal Process. **80**(4), 520–568 (1992)
12. Chen, L., Guizeng, W., Hao, Y.: Gap conditions monitoring for main bearing based on noise measurements. J. Vib. Shock **22**(3), 33–36 (2003)

Chapter 66
Robot Perception Based on Different Computational Intelligence Techniques

Nacereddine Djelal and Nadia Saadia

Abstract In this paper we tend to investigate an optimal model through the application of three well known techniques of computational intelligence; neural net work, fuzzy logic and Support Vector Regression (SVR). An attempt has been made to get a stochastic model of the contact force of a robot with its environment. The model is provided by an identification process based on supervised learning. The utility of this model lies in a high quality perception and allowing us to implement an intelligent control of industrial tasks performed by a robot. The objective behind using the computational intelligence techniques is to improve the performance of the industrial robot's tasks such as: welding, painting, pieces insertion, etc. The findings of the investigation have showed that the SVR is the most suitable technique for achieving a high accurate model.

Keywords Neural network · Fuzzy logic · Support vector regression · Identification · Contact force modeling

66.1 Introduction

In recent years, the industrial robots become a perquisite in a lot of industrial applications which need high accuracy performance. In fact, we aim in this work at improving the quality of robot's perception by using different computational intelligence techniques in order to model the contact force of the robot to be implemented in an intelligent control law.

N. Djelal (✉) · N. Saadia
Laboratory of Robotics, Parallelism and Electroenergetics, University of Sciences and Technology Houari Boumediene, BP 32 El-Alia 16111 Bab-Ezzouar Algiers, Algeria
e-mail: ndjelal@usthb.dz

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_66,

In the last few decades, many communities of researchers were interested in improving the performance of industrial robot in different tasks; in our previous works [1–4], we proposed to control a robot by the feed forward neural network.

The outline of this paper is organized as follows: first, we present the paradigm of the supervised learning. Second, we explain in detail the three proposed computational intelligence techniques i.e. the neural network used in the modeling, the fuzzy logic technique and the SVR. After that, the aforementioned techniques are to be implemented so as to construct the desired models. In so doing, we tend to present the dataset of learning which is presented as the pair input/output where the shed torque of the joint motor represents "input" and the force values as "output". Then, we begin the identification process during which we choose the most appropriate model through regularization by cross validation. Next, each technique is to be evaluated by the measure of the Mean Magnitude of Relative Error (MMRE) and Standard Deviation Magnitude of Relative Error (StdMRE). Finally, the findings and conclusions are discussed.

66.2 Supervised Learning for Identification

This section describes the mechanism of supervised learning which is commonly used to identify or model the system that is used in recognition or intelligent controller. The Fig. 66.1 illustrates the Paradigm of supervised learning, where the training data are used as input for the Supervisor (Fuzzy Rules for example in the case of the fuzzy logic) and the Learning Machine. The learning algorithm must adjust the parameters of the Learning Machine to minimize the error.

Where: x represents the vector of the shed torque of the joint motor

y represents the vector of the force values and $\hat{y}$is the predicted value of y

The learning machine which is capable of implementing a set of functions $f(x, w)$, $w \in A$, where A is a set of parameters; and the weights and bias in the neural network for example [5].

We distinguish three techniques based on supervised learning which are the following:

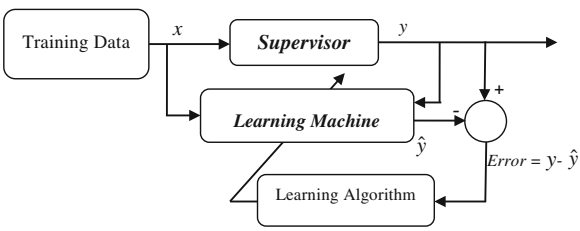


Fig. 66.1 Paradigm of supervised learning

66.2.1 Modeling Based on Neural Network

This section is devoted for the presentation of the neural network structure and how it is used in order to model the contact force. It is worth mentioning that we have to follow these steps:

Firstly, we suggest architecture for the neural network i.e. the number of the neurons which constitutes the input layer, the hidden layer and the output layer. The activation function must be determined in this phase. In the second, the training of the neural network necessitates a training data and a training algorithm; the Levenberg–Marquardt algorithm in our case is chosen. This method consists of optimizing the step of training to increase the speed of convergence of the global criteria [6].

66.2.1.1 Proposed Architecture

The proposed architecture of neural network uses three layer neural networks with a same activation function; nonlinear sigmoid function [4] (Eq. 66.6);

$$f(x) = \frac{1}{2} X_g \frac{1 - \exp(-4x/X_g)}{1 + \exp(-4x/X_g)} \tag{66.1}$$

where Xg is the parameter which determines the sigmoid function shape.

This network is composed of one input layer of six neurons, one hidden layer of twenty neurons and one output layer of six neurons.

66.2.1.2 Learning Process of the Neural Network

This process of learning provides us with the parameters of the neural network i.e. weights and the bias. Before starting this process we must fix the criterion or the cost function that describes the relationship between the target and measured values. In our dataset which is composed of a set of input/output, we have relied on a supervisor to learn the neural network:

The Levenberg–Marquardt algorithm uses this approximation (Eq. 66.2) to construct the Hessian matrix as follows [6]:

$$w_{k+1} = w_k - \left[J^T J + \mu I\right]^{-1} J^T e_f \tag{66.2}$$

where:

w_k: is a vector of current weights and biases,

Where, the scalar μ is equal to zero (similar to Newton's method). When μ is large, this becomes gradient descent with a small step size. Then Newton's method is faster and more accurate.

Thus, μ is decreased after each successful step (low performance function) and it is augmented only when a tentative step increases the performance function.

66.2.2 Modeling Based on Fuzzy Logic

In this section we aim to apply fuzzy logic to identify the model of the contact force through the training dataset. The shed torque of the joint motor *x(t)* is used as the input whereas the force values *y(t)* as the output. So, we can extract the dynamic process model to establish a model using 10 candidate inputs to the fuzzy logic: *y(t−1)*, *y(t−2)*, *y(t−3)*, *y(t−4)*, *x(t−1)*, *x(t−2)*, *x(t−3)*, *x(t−4)*, *x(t−5)*, and *x(t−6)* as historical data [7].

66.2.3 SVR Based Modeling

The adoption of a new estimation technique that is based on support vector machines seems to be powerful to get a model mapping between the input/output set. Suppose the training data $\{(x_1, y_1), \ldots\ldots, (x_\ell, y_\ell)\} \subset \chi \times \Re$, where χ is the space of the input patterns [8]. In ε-SV regression is proposed by Vapnik in the Ref. [5]. The aim is to find the function *f(x)* that has the big ε deviation from the actually obtained targets y_i for the whole training data which accept errors less than ε.

For simplicity grounds, we use a linear function f given by this equation:

$$f(x) = \langle w, x \rangle + b \text{ with } w \in \chi, b \in \Re \tag{66.3}$$

where; $\langle \cdot, \cdot \rangle$ is the dot product in χ.

To ensure the flatness in the case of (66.3) we have to minimize the norm$\|w\|^2 = \langle w, w \rangle$. So, we can formulate this problem as a convex optimization problem as follows:

$$\begin{aligned} &\text{minimize } \frac{1}{2}\|w\|^2 \\ &\text{subject to } \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon \end{cases} \end{aligned} \tag{66.4}$$

Like the "soft margin" of the Support Vector Machine, loss function [9] which was used in SV machines developed by Cortes and Vapnik (1995). The introduction of slack variables ξ_i, ξ_i^* to adapt with otherwise infeasible constraints of the optimization problem we obtain:

$$\text{minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}(\xi_i + \xi_i^*)$$
$$\text{subject to } \begin{cases} y_i - \langle w, x_i\rangle - b \le \varepsilon + \xi_i \\ \langle w, x_i\rangle + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \tag{66.5}$$

where: C > 0 is constant and ξ_i, ξ_i^* are slack variables

The idea of the soft margin is to use the slack variables and ε-insensitive loss function (Eq. 66.6) to get a linear SVM as shown in Fig. 66.2 [8].

$$|\xi|_\varepsilon := \begin{cases} 0 \text{ if } |\xi| \le \varepsilon \\ |\xi| - \varepsilon \text{ otherwise.} \end{cases} \tag{66.6}$$

Figure 66.2 presents graphically the soft margin for a linear SVM. Only the points which appear outside present the shaded region contribute to the cost insofar, as the deviations are penalized in a linear method. It turns out that in most cases of the optimization problem can be solved more easily in its dual formulation [8].

66.2.3.1 Preprocessing by Kernels

Making the SVR algorithm nonlinear entails the use of kernels [8]. This can be simply achieved by the preprocessing of the training patterns x_i by a map Φ: $w \rightarrow F$ into feature space F.

We refer to the use of the Quadratic features in $\Re^2$. So, we consider the map Φ: $\Re^2 \rightarrow \Re^3$ given by this equation :

$$\Phi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2) \tag{66.7}$$

where this function aims to project the features into another representative space.

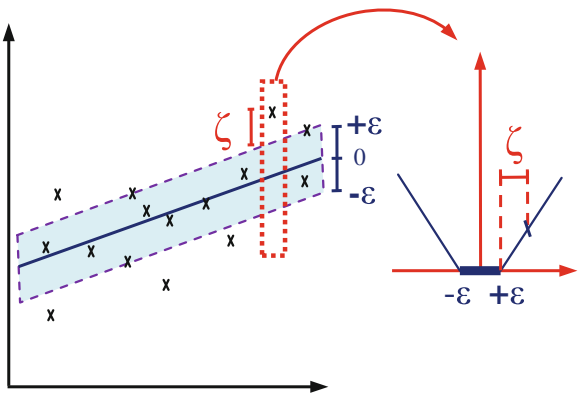


Fig. 66.2 The soft margin loss setting for a linear SVM

66.3 Constructions of Models

The modeling process via a paradigm of learning needs a training data. In order not to encounter under fitting and over fitting problems, a technique of cross validation is introduced. In the last, we must evaluate the obtained models of the contact force.

66.3.1 Dataset of Learning

The adopted dataset of learning in this work is collected during an industrial robot achieving industrial task (such as the assembly i.e. insertion of piece in hole). So, the dataset of learning which is provided as the pair input/output in which the shed torques of the joint motors represent the "input" while the values of force represent the "output".

66.3.2 Regularization Based on Cross Validation

The underlying idea of this regularization lies in the way in which a model is selected from a set of models. As an example, we consider choosing the order of a polynomial. The polynomial model quality depends proportionally on its order. So, the best model is the one that has a sufficient fitting, and hence the training error lessens [10]:

66.3.3 Models Evaluation

So as to assess and contrast the performance of models provided by different computational techniques we have proposed to calculate the measures of the MMRE and the StdMRE. These two criteria are widely used in evaluating the model's quality [11] and [12].

$$MMRE = \frac{1}{n}\sum_{i=1}^{n} MRE_i \tag{66.8}$$

where: MRE_i is a normalized measure of the difference between the actual output values of the model (y_i) and estimated value ($\hat{y}_i$) given by [12]:

$$MRE_i = \frac{|y_i - \hat{y}_i|}{y_i} \tag{66.9}$$

Table 66.1 Models evaluation

Model	SVR	Fuzzy logic	Neural network
MMRE	0.720	1.810	4.26
StdMRE	0.690	0.420	6.48

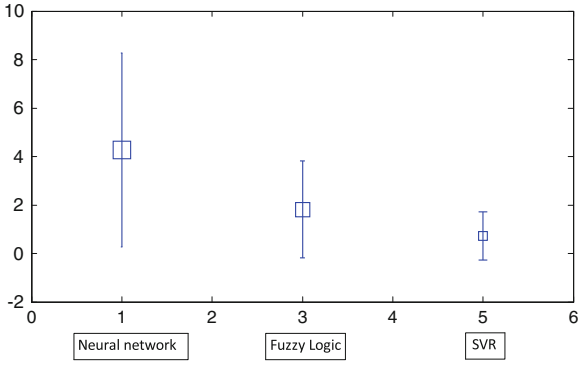


Fig. 66.3 Boxplots of evaluation criteria

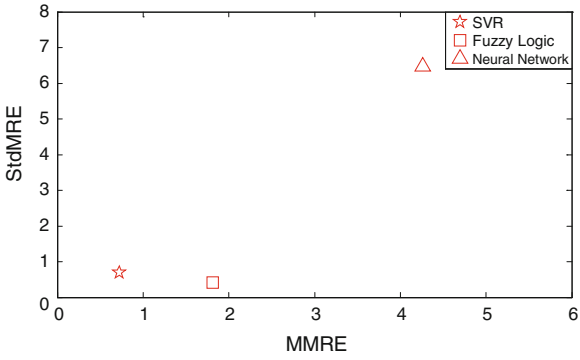


Fig. 66.4 MMRE versus StdMRE evaluation criteria

The use of StdMRE is needed because it has a less sensitivity to the extreme values in comparison with MMRE [12].

66.4 Results and Discussions

The findings of the current study showed that the SVR is the most appropriate technique for getting the best model of contact force. This claim is proved by the values of MMRE and StdMRE (Table 66.1) which appear to be the lowest values.

Moreover, the fuzzy logic technique seems to be better than the neural network one.

As illustrated in Fig. 66.3, the boxplot allows us to visualize the accuracy of each model. We have observed through the figure that the SVR has the highest level of accuracy because of its narrowest box and the smallest whiskers.

In Fig. 66.4 we can see that the coordinates (MMRE, StdMRE) of the SVR are near to zero which means that the model of contact force has the lowest error value.

66.5 Conclusion

In conclusion, it is quite safe to say that three techniques of computational intelligence were implemented. The proposed paradigm of supervised learning was used in the process of identification and it was validated. After analyzing the results, we come to the conclusion that the contact force model provided by SVR technique is proved to be the best one compared with the other two techniques i.e. fuzzy logic and neuron network. Further research is needed to enhance the findings of this work by the use of a contact force model in controlling of an industrial robot to achieve intelligent tasks.

References

1. Saadia, N., Amirat, Y., Pontnaut, J., Ramdane-Cherif, A.: Neural adaptive force control for compliant Robots. 6th international work-conference on artificial and natural neural networks, IWANN 2001 Granada, Proceedings, Part II, pp. 436–443. Springer (2001)
2. Saadia, N., Amirat, Y., Pontnau, J., M'Sirdi, N.K.: Neural hybrid control of manipulators, stability analysis. Robotica **19**, 41–51 (2001)
3. Saadia, N., Amirat, Y., Pontnau, J., Ramdane-Cherif, A.: Hybrid force/position control of nonlinear systems using neural networks. In: Section ECPD international conference on advanced robotics intelligent automation and active systems, pp. 68–73. Vienna, Austria (1996)
4. Touati, Y., Amirat, Y., Saadia, N., Ali-Chérif, A.: A neural network-based approach for an assembly cell control. J. Appl. Soft. Comput. 1335–13438, Elsevier (2008)
5. Vapnik, V.N.: The nature of statistical learning theory. Springer, New York (1995)
6. Demuth, B.H., Martin, M.B., Hagan, M.T.: Neural Network Design. PWS Publishing, Boston (1996)
7. Roger Jang, J.S.: Neuro-Fuzzy and Soft Computing A Computational Approach to Learning and Machine Intelligence. Prentice Hall, Upper Saddle River (1997)
8. Smola, A.J., Schölkope, B.: A tutorial on support vector regression. Statistics and Computing, Kluwer Academic Publishers, Dordrecht (2004)
9. Bennett, K.P., Mangasarian, O.L.: Robust linear programming discrimination of two linearly inseparable sets. Optim. Methods Softw. **1**, 23–34 (1992)
10. Witten, I.H., Frank, E.: Data Mining Practical Machine Learning Tools and Tech niques. Elsevier, Amsterdam (2005)
11. Conte, S., Dunsmore, H., Shen, V.: Software engineeringmetrics andmodels. Benjamin/Cummings, Menlo Park (1986)
12. Elish, M.O.: A comparative study of fault density prediction in aspect-oriented systems using MLP, RBF, KNN, RT, DENFIS and SVR models. Artif. Intell. Rev. doi:10.1007/s10462-012-9348-9(2012)

Chapter 67
Study on Technique for GPS IF Signal Simulation

Huaijian Li, Jun Dai, Wenguang Li and Li Liu

Abstract In order to solve the difficulty of simulating the satellite signal with the complexity of the structure of hardware signal source and the variability of the environment, an IF signal model architecture is used in this paper. By producing navigation message and observational data with the mathematical simulation system offline, these data are modulated by the pseudo-random code and the IF carrier to produce the IF analog signal. Then the IF analog signal becomes the IF digital signal after further sampled and Quantify. The GPS IF signal is produced by the use of Matlab simulation and then is further located though the receiver. Simulation result shows that receiver positioning result is basically the same as observational data which is generated by the mathematical simulation system, proving that the scheme is of good reason. The IF signal model can effectively solve the disadvantage of hardware simulation equipment and thus a class of problems including High input and Low output which hardware simulation attributes are solved.

Keywords GPS · The pseudo-random code · Sampling · Quantify

67.1 Introduction

Starting from Texas Instruments company's earliest development of the GPS signal simulator in the 1977 [1], accompanied by the establishment of the GPS satellite constellation and the emergence of other satellite navigation systems, the development of the satellite signal simulation technology is also changing.

H. Li (✉) · J. Dai · W. Li · L. Liu
School of Aerospace Engineering, Beijing Institute of Technology, Beijing, China
e-mail: daijun502@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_67,

Any research in the field of satellite navigation must first be established in a true and reliable data base, and then be extended to the application. But true satellite signal is high-frequency RF signal and is mixed in a changing environment noise. If you want to simulate by hardware method, you need to consider the high-frequency signal propagation, delay, multipath interference and many other influential factors which resulting in hardware analog devices becoming complex [2].

Analog satellite navigation signal is required for the development and validation of the navigation receiver, compared to the direct use of a real satellite signal, analog satellite signal can provide accurate controllable, reproducible simulation environment and the non-normal or undesirable test condition, so that the receiver R&D efficiency is greatly improved.

GPS satellite signal simulation is a key technology in the current GPS research area, which provides real simulation for test of high-dynamic receiver. GPS measurement system can also be used for demonstration program which is widely used in GPS receiver debugging, signal processing and positioning algorithm research. So satellite navigation signal simulation technology has important theoretical and practical significance.

67.2 The GPS IF Signal Model

Real GPS signal generating process is as follows: in the GPS satellite, the 50 bps GPS navigation message data is modulated with 1.023 MHz pseudo-random code C/A code to generate a GPS baseband signal and baseband signal is modulated with by L1 (1575.42 MHz) RF the carrier, and then the GPS RF signal is produced. Due to the signal delay of transmission, when the GPS signal is transmitted from the satellite to the receiver, the receiver receives the RF signal which is transmitted by the satellite with a transmission delay (such as troposphere delay, ionosphere delay, earth rotation effect, relativistic effect). Meanwhile, due to relative motion between the satellite and the receiver, the GPS signal will generate the Doppler shift so that the signal frequency which the receiver receives will change.

In the IF signal software simulation system, we use the mathematical simulation system which produces satellite navigation message data and observational data offline. By the settlement of the user trajectory and the establishment of the GPS satellite orbit model simulation, simulation data including navigation message and observation data is created. Navigation message format is 300 bit, which is accounted for 40 bytes, remaining 20 bit empty. Observation data is created according to the settlement of user location the satellite trajectory model.

Reading simulation data file in the navigation data and the observed data, GPS signal source can calculate the emission time of the satellite signal and the receiver state when the signal reached which includes the visible satellites, the Doppler shift, the signal amplitude, the phase of the pseudo-code at launching moment, the code phase in transmission time and the carrier phase in the emission time. After

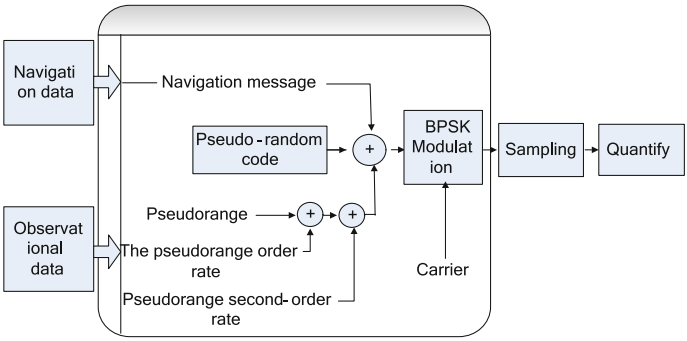


Fig. 67.1 IF signal generation model flow chart

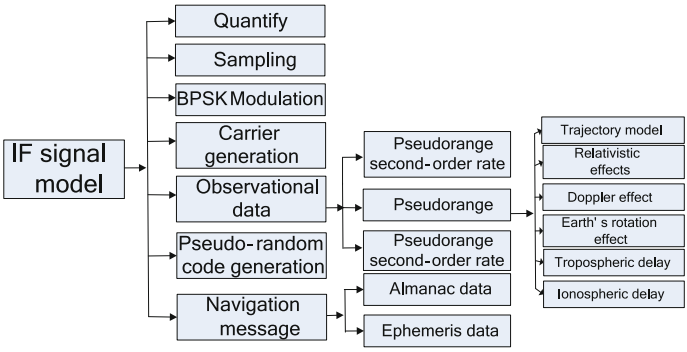


Fig. 67.2 IF signal generation framework figure

the information is modulated by radio frequency to be the IF analog signal, the analog IF signal after further is sampled, quantized to become a digital IF signal, as shown in Figs. 67.1 and 67.2.

67.3 Key Technical Analysis

67.3.1 Pseudo-Random Code Generation

The GPS C/A code is a Gold code of length 1023, and the code rate is 1.023 MHz. The repetition period of the pseudo-random sequence is 1 ms, including two 10-bit shift register $G1(X)$ and $G2(X)$, and the initial state is all 1. C/A Code is XOR result between after delay $G2(X)$ Output sequence and $G1(X)$ Direct output sequence, $G2(X)$ PN code delay effect is obtained by the choice of the position of the two taps. Two taps of each satellite is not the same. There are many ways to

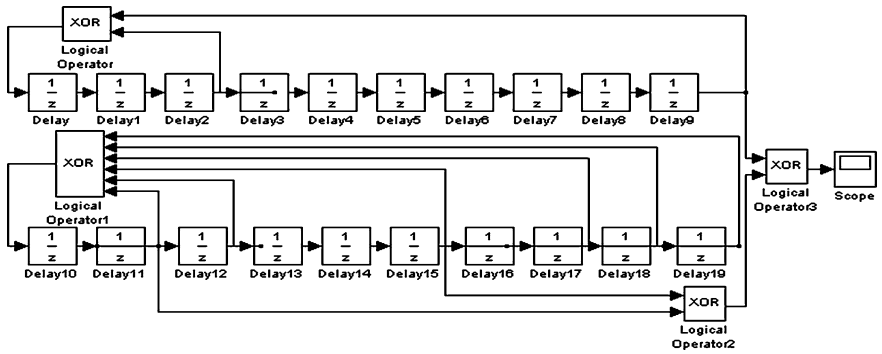


Fig. 67.3 No. 3 star simulation module figure

generate C/A code by Matlab, we use Matlab/simulink simulation. We take Star 3 for example: the simulation module of the 3rd Star C/A code is as shown in Fig. 67.3.

67.3.2 Spread Spectrum Modulation

When GPS satellite navigation message is transmitted to the user, spread spectrum modulation technique is used. The navigation message is about to send in a low rate into the composite code spread transmission, thus extending the range of the spectrum of the signal, reducing the signal power density, enhancing the data transmission of immunity and confidentiality.

D code navigation message is the user basic information, which is used for navigating and positioning. It mainly contents satellite ephemeris, satellite clock correction parameters, ionosphere delay correction parameters, the working status of the satellite information and the C/A code conversion to capture the P code information, an outline of all the satellite ephemeris [3].

Spread spectrum modulation is the process where the C/A code modulates the data codes, i.e. C/A $\oplus$ D. the C/A code rate is 1.023 MHz, however D code rate is 50 Hz, so D code in the 20 cycle of C/A code is kept constant, therefore the 20 C/A code load signal. When the load signal is valid, the data is loaded in the C/A code sequence. However, it is needed to identify the initial phase of the C/A code and D code, i.e. it is obtained according to the calculation of the observation data.

67.3.3 Sampling

The Nyquist sampling theorem shows that in order to prevent signal aliasing in the sampling process, sampling frequency f_s must be greater than twice the highest

frequency. However, when Bandpass signal is sampled, the sampling frequency may be less than twice the highest frequency of the signal, but it must be greater than twice the signal bandwidth.

When GPS front-end is designing, the choice of sampling frequency don't only need to satisfy the sampling theorem, but also must meet another important condition, that is, the sampling frequency is not an integer multiple of the C/A code of the code rate (1.023 MHz), otherwise, it will result in the sampling should be synchronized with the C/A code rate, i.e. the sampling time will always get the same sampling data [4].

In practice, due to the presence of C/A code Doppler frequency shift, the sampling frequency is not only not be the integer multiple of the rate of the C/A code, but also can't be 1.023 MHz plus or minus the Doppler frequency C/A code resulting the integer multiple of the frequency value. Actual choice of the sampling frequency is not close to integer multiple of the value of this frequency, otherwise the sample data will be relatively close to, the need of using a relatively long data can be detected to the time difference of the sampled data which is unnecessary [4].

67.3.4 Quantify

The analog signal is sampled to be a discrete signal, but since it is still an analog signal, this sampling signal must be quantized only to be a digital signal. The basic principle is comparison with a threshold value which determines the output of the corresponding data bit which is 0 or 1. Increasing the number of quantization bits helps to reduce the quantization variance, thereby reducing the loss of receiver sensitivity is caused by the quantization variance [5]. 1 bit quantization variance loss is about 3.5 dB, 2bit quantization variance loss is about 1.2 dB, 3 bit quantization variance loss is about 0.6 dB. Most receivers rarely used more than 4 bit quantization, because it is substantially that quantization variance loss no longer continues to increase because more 4 bit quantization can't help to reduce the quantization loss [6].

We refer to References <Design and Validation of an Accurate GPS Signal and Receiver Truth Model for Comparing Advanced Receiver Processing Techniques> [3] in Matlab quantify program:

The design process is as following:

1. The calculation of the variance and the average of all signals;
2. The initial value of the threshold value is the variance*(68.3/70), where the variance is the variance of all the group data;
3. The signal values after quantization decide whether to adopt the quantization result or whether to reset threshold value for a new attempt to quantify.

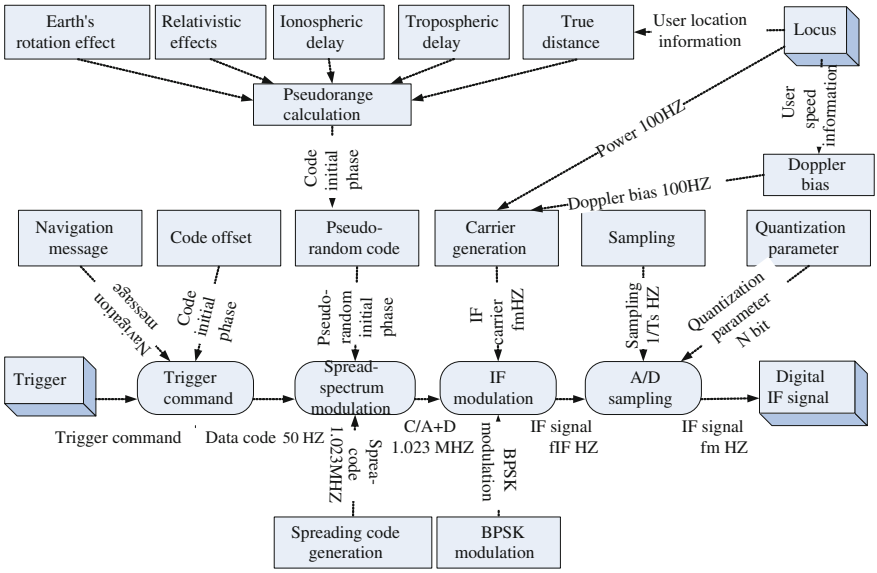


Fig. 67.4 IF signal to produce a data flow diagram

67.4 IF Signal Generator

The process that IF signal is generated is shown in Fig. 67.4. As can be seen from Fig. 67.4, Navigation message is generated by the spread spectrum modulation, IF modulation frequency analog signal is:

$$s_{IF}(t) = A_{IF}(t)C(t-\tau(t))D(t-\tau(t))\cos[2\pi(f_{IF}+fd(t))t+\varphi_0] \tag{67.1}$$

where:

$A_{IF}(t)$	The amplitude of the signal;
$\tau(t)$	The time delay of the signal;
$C(t-\tau(t))$	The pseudo-random code;
$D(t-\tau(t))$	The navigation data;
φ_0	The initial phase of the IF carrier;
f_{IF}	The frequency of the IF carrier;
fd	The Doppler shift;

After the ADC sampling, the signal can be expressed is:

$$\begin{aligned} x_{IF}(nT_s) = A_{IF}(nT_s)C(nT_s-\tau(nT_s))D(nT_s-\tau(nT_s)) \\ \cos[2\pi(f_{IF}+fd(nTs))nT_s+\varphi_0] \end{aligned} \tag{67.2}$$

n	The serial number;
T_s	T sampling period;

67.5 Simulation and Analysis

Set the user receiver position longitude 110°, latitude 39°, the velocity is zero. 2006-01-01 0:30:0 broadcast ephemeris data the calculated visible satellite asterisk 3, 13, 15, 16, 19, 21, 23, 25, 27. Fetch the IF signal frequency for 4.123968 MHz, the sampling frequency is 16.367667 MHz, and quantization bit is 2 bit. The reliable software receiver is used to test the simulation result. The result is obtained after the IF loop processing, as shown in Table 67.1:

For IF processing loop only own eight channels, only existing eight data visible stars. The loop processing input the navigation positioning settlement program, the result is shown in Table 67.2.

As can be seen from Table 67.2, the location positioning accuracy is 1 m, speed accuracy is 0.001 m/s. The error result shows that the positing result though soft signal source and soft receiver is same as the data which is generated by mathematical simulation system, producing that the IF signal simulation framework model is reasonable and the simulation method is correct and feasible.

Table 67.1 A state when the signal reaches the receiver

PRN	Doppler frequency (Hz)	Signal propagation time (ms)
3	740.4884	67.07823
13	704.3876	72.76416
15	1207.347	79.38541
16	−1828.22	71.71761
21	−235.027	82.2703
23	−1656.13	71.01113
25	−2859.58	78.63494
27	3109.852	83.37185

Table 67.2 A state when the signal reaches the receiver

Location	X direction (m)	Y direction (m)	Z direction (m)
Test value	−1697557.9277	4664002.0746	3992317.0228
Theoretical value	−1697558.8273	4664001.5751	3992317.8035
Error	−0.8996	0.4995	−0.7807
Speed	X direction (m/s)	Y direction (m/s)	Z direction (m/s)
Test value	0.0007	−0.0005	0.0003
Theoretical value	0	0	0
Error	0.0007	−0.0005	0.0003

67.6 Conclusion

Based on real GPS signal generator model, an overall framework of the IF GPS signal generation model is put forward in this paper. And then the main key technologies in the intermediate frequency signal generation process is described in detail, including the completion of the pseudo-random code C/A code generation and the sampling frequency and the determine of quantization bits. At last the GPS IF signal is acquired by simulating Matlab. From the analysis of the result, the good rationality of the model is proved and the difficulty with the complexity of the structure of hardware signal source and the inoperability of upgrade is solved.

Acknowledgments Project supported by the state key program of National Natural Science of China(Grant NO. 61173077); Project supported by the National High Technology Research and Development Program of China (863 Program) (Grant NO. 2011AA120505).

References

1. Candy, D.W.: NAVSTAR/GPS SIMULATOR. IEEE Proceedings of the National Aerospace and Electronics Conference, 1977, pp. 323–329 (1977)
2. Jianping, Y., Jianjun, L., XiaoKui, Y.: Satellite Navigation Principles and Application. China Astronautic Publishing House, Beijing (2010)
3. Brown, A., Gerein, N., Taylor, K.: Modeling and simulations of GPS using software signal gerneration and digital signal reconstruction. Proceedings of ION NTM (2000)
4. Gang, X.: GPS Principle and Receiver Design. Electronic Industry Press, Beijing (2011)
5. Oppenheim's, A., Schafer, R., Buck, J.: Discrete-Time Signal Processing: The Discrete-Time Signal Processing. In: Liu, S., Huang, J., (eds.) translated version 2.: Xi'an Jiaotong University Press, Xi'an (2010)
6. Parkinson, B., Spilker, J., Axelrad, P., Enge, P.: Global Positioning System:Theory and Application. American Institute of Aeronautics and Astronautics (1996)

Chapter 68
New Method of Image Denoising Based on Fractional Wavelet Transform

Peiguang Wang, Yan Yan and Hua Tian

Abstract Nowadays, there are many mature image denoising methods, such as linear filtering and nonlinear filtering. In order to improve the denoising effect, a novel signal denoising method based on fractional wavelet transform (FRWT) is proposed in this paper. It combines the advantages of the fractional Fourier transform (FRFT) and the wavelet transforms (WT). By the simulation experiment, the optimal fractional order of FRWT is obtained with an iterative algorithm according to the PSNR of output signals. This method takes output peak signal to noise ratio (PSNR) and information entropy (IE) as the denoising evaluation index. The results of experiment show that the novel methods could effectively remove noise, and maintain information quantity maximally at the same time by adjusting the fractional order p and wavelet scale.

Keywords Fractional wavelet transform · Peak signal to noise ratio · Information entropy

68.1 Introduction

According to the physical properties of the image, such as the statistical distribution and the spectrum law, a lot of novel image denoising methods have been emerging. The purpose of the image denoising is distinguishing between noise and original image. How to retain the image information maximally becomes the key of the process of reducing the noise of the image [1]. Traditional denoising method concludes linear filtering algorithm and nonlinear filtering algorithm, liking the mean filter and median filter. But the classical methods have inherent defect, that it

P. Wang (✉) · Y. Yan · H. Tian
College of Electronic and Information Engineering, Hebei University, Baoding, China
e-mail: pgwang@mail.hbu.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_68,

can not maintain the image information quantity, so the image becomes blurred. The combination of wavelet transforms and median filtering algorithm can ensure the information quantity in a certain extent, but due to poor applicability, the advantage is not obvious.

In recent years, fractional wavelet transform (FRWT) becomes a new research direction as the effective transformation of fractional Fourier transform (FRFT) and wavelet transform (WT). D. Mendlovic and Z. Zalevsky gave the definition, decomposition and reconstruction formula of fractional wavelet transform in 1997 [2]. C. Linfei and Z. Daomu realized image encryption by using fractional wavelet transform theory in 2005 [3], the computer simulation algorithm of image processing becomes possible. In this paper, the fractional wavelet transform theory is applied to reduce the noise of image. The results of experiment show that the novel methods could effectively remove noise, and maintain information quantity maximally at the same time. Therefore, the overall effect of the proposed method is far superior to the traditional method.

68.2 Mathematical Definition of Fractional Wavelet Transform

The FRWT may be formulated as follows [4]:

$$W^{(p)}(a,b) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} B_p(x,x')f(x')h^*_{ab}(x)dxdx' \tag{68.1}$$

where $h^*_{ab}(x)$is the wavelet basis function, as follows :

$$h^*_{ab}(x) = \frac{1}{\sqrt{a}}h(\frac{x-b}{a}) \tag{68.2}$$

$B_p(x,x^{'})$ is the kernel function, as follows:

$$B_p(x,x') = \sqrt{2}\exp[-\pi(x^2+x'^2)] \times \sum_{n=0}^{\infty}\frac{i^{-pn}}{2^n n!}H_n(\sqrt{2\pi}x)H(\sqrt{2\pi}x') \tag{68.3}$$

where H_n is an n order Hermite polynomial, or,

$$B_P(x,x') = \frac{\exp\left[-i\left(\frac{\pi \mathrm{sgn}(\sin\phi)}{4} - \frac{\phi}{2}\right)\right]}{|\sin\phi|^{1/2}\exp\left[i\pi\frac{x^2+x'^2}{\tan\phi} - 2i\pi\frac{xx'}{\sin\phi}\right]} \tag{68.4}$$

The one dimensional signal reconstruction formula is as follows:

$$f(x) = \frac{1}{c}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{a^3}W^{(p)}(a,b)B_{-p}(x,x') \times h(\frac{x'-b}{a})dadbdx' \tag{68.5}$$

The value range of order p is (0, 1], When p = 1, the whole formula is converted into a conventional wavelet transform (WT). For the two dimensional signal f(x, y),

$$W(a_{mn},\tilde{b}) = \int\int\int\int B_{p1,\,p2}(x,y;x',y')f(x,y) \times h^{*}_{a_{mn}b}(x',y')dxdydx'dy' \tag{68.6}$$

The two dimensional reconstruction formula:

$$f(x,y) = \frac{1}{c}\int\int F\left[\sum_m\sum_n\int\int\frac{1}{a_m a_n}W(a_{mn},\tilde{b})H(a_m u,a_n v)\exp(-j2\pi u b_{x'},-j2\pi v b_{y'})db_{x'}db_{y'}\right] \times (x',y')B_{-p_1,-p_2}(x,y;x',y')dx'dy' \tag{68.7}$$

The order of the two dimensional transformation is p = [p1, p2], when p1 = p2 = 1, it is converted into the conventional wavelet transform (WT).

As the mathematical definition, the algorithm is linear transformation. If the input is two linear superposition and mutually independent signals, the fractional wavelet transform will be the respective transformation of signals [5]. This is essentially a new time–frequency type, using the single variable to represent the time–frequency information of input signal. The transformation is suitable for processing non-stationary signals, and the fractional order p can be adjusted freely. The overall effect is far superior to the traditional time–frequency method.

68.3 Algorithm and Flowchart

Fractional wavelet transform (FRWT) algorithm is realized according to the mathematical definition of fractional wavelet transform, the fractional Fourier transform (FRFT) algorithm and wavelet transform (WT) algorithm [6]. If the input signal is a two dimensional image, then the order p of the entire transformation becomes [p1, p2] form. Respectively, it can be realized a two dimensional transformation by the row and column directions. The flowchart of denoising progress using the FRWT algorithm is shown in Fig. 68.1.

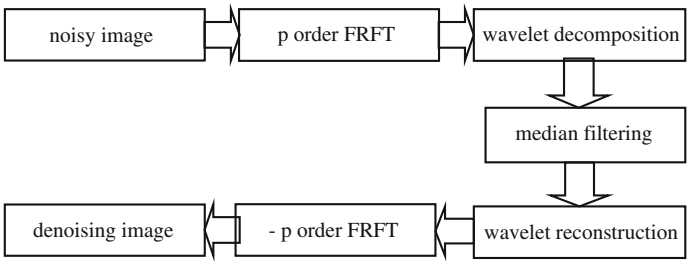


Fig. 68.1 Denoising algorithm flowchart

68.4 The Selection of FRWT Basis Functions and Levels

FRWT algorithm of the novel method needs to select the appropriate basis functions and levels. Peak signal to noise ratio (PSNR) is the primary denoising evaluation index. The formula is as follows:

$$PSNR = 10 \times \log\left(\frac{255^2}{MSE}\right) \tag{68.8}$$

where MSE is the mean square error,

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\|I(i,j) - K(i,j)\|^2 \tag{68.9}$$

The original image (a) selects a classic 256 × 256 experimental image, noisy image (b) adds the Gauss noise, $\sigma = 0.01$ (Fig. 68.2).

As we can see from Table 68.1, when the basis function is db5, the level is 3, the PSNR is highest, and thereby the basis functions and levels are determined.

(a) **(b)**

Fig. 68.2 The original image and noisy image. **a** Original image. **b** Noisy image

Table 68.1 PSNR of different basis functions and levels

PSNR	db3	db4	db5	sym3	sym4	sym5
2	46.7123	46.7217	46.7209	46.7230	46.6988	46.7021
3	46.7286	46.7354	46.7540	46.7351	46.7358	46.7413
4	46.7395	46.7303	46.7368	46.7336	46.7314	46.7414

68.5 The Selection of FRWT Order

Fractional wavelet transform has an advantage that exists the adjustable order p, matrix form p = [p1, p2], p1 is horizontal axis processing variable, p2 is longitudinal axis of the processing variable. In order to ensure the image cross on coordinate axis, p1 = p2, the range is (0, 1).

Fractional wavelet transforms basis function and level is db5 and 3. We investigate the denoising evaluation index, including PSNR and IE, through the order p gradually being increased from 0.1 to 0.9. The peak signal to noise ratio (PSNR) is the primary index, the value is higher, and the noise in the output image is lower. Shown in Table 68.2, p = 0.3 and 0.6, PSNR is relatively higher, in particular, p = 0.3, PSNR = 46.7548, PSNR and IE is the highest.

68.6 Computer Simulation

Objectively, by a series of parameter settings and adjustments, the PSNR completely can reach the maximum. However, as the output image, the readability should be considered, which reflects the maintain information quantity as information entropy (IE) [7].

Shown in Table 68.3, to select the appropriate parameters, the PSNR of two methods are basically the same (46.7540 and 46.7548), the novel method (FRWT) of image information entropy IE (6.7268) is greater than the traditional method as median filter (6.7187).

Table 68.2 Denoising evaluation index of FRWT different orders

order p	PSNR	IE
0.1	46.7389	6.7276
0.2	46.7439	6.7166
0.3	46.7548	6.7268
0.4	46.7319	6.7221
0.5	46.7383	6.7261
0.6	46.7497	6.7249
0.7	46.7254	6.7265
0.8	46.7378	6.7250
0.9	46.7279	6.7246

Table 68.3 Median Filter, mean filter and FRWT

Type	PSNR	IE
Median Filter	46.7540	6.7187
FRWT	46.7548	6.7268

Fig. 68.3 The experimental images. **a** Median filter. **b** FRWT

Figure 68.3, the amount of noise is low in (a) and (b), but the novel method maintains information quantity maximally by comparing information entropy (IE). Denoising overall effect of the novel method is superior to the traditional method.

68.7 Conclusion

In this paper, a novel image denoising method is proposed and coining the fractional wavelet transform (FRWT). This method removes the image noise effectively and maintains information quantity maximally by using the fractional Fourier transform (FRFT) and wavelet transform (WT). Compared with the conventional wavelet transform, the proposed method demonstrates this ability to provide a higher peak signal to noise ratio (PSNR). For the noisy image, firstly, it performs fractional Fourier transform and wavelet decomposition. Then, the transformed signal is filtered with the median filter. Finally, the output is reconstructed with the wavelet reconstruction and fractional Fourier inverse transform. To adjust the fractional order p and wavelet scale, the novel methods could effectively remove noise, and maintain information quantity maximally at the same time. So the overall effect of denoising is superior to the traditional method. More work should be done on the optimization step of FRWT, when the type and strength of noise is different.

References

1. Jiecheng, X., Dali, Z., Wenli, X.: Overview on wavelet image denoising. J. Image Graph. **7**(3), 209–217 (2002)
2. Mendlovic, D., Zalevsky, Z., Mas, D.: Fractional wavelet transform. Appl. Opt. **36**(20), 4801–4806 (1997)

3. Linfei, C., Daomu, Z.: Optical image encryption based on fractional wavelet transform. Opt. Commun. **254**(2005), 361–367 (2005)
4. Lin, Y.: Wavelet-fractional fourier transforms. Chin. Phys. B **17**, 170–179 (2008)
5. Ying, H.: The fractional wave packet transform. Multidimension. Syst. Signal Process. **4**(9), 399–402 (1998)
6. Yuqing, H., Youren, W., Hui, L., et al.: New signal denoising method based on fractional wavelet packet transform in time-frequency domain. Chin. J. Sci. Instrum. **32**(7), 1534–1539 (2011)
7. Hongjin, Y., Xueying, Z., Xiaogang, H.: Medical image denoising based on wavelet transform and median filtering. J. Taiyuan Univ. Technol. **36**(5), 4–7 (2005)

Chapter 69
Semantic Representation of Role and Task Based Access Control

Guang Hong, Weibing Bai and Shuai Zhang

Abstract Proper representation of Role and Task in access control mechanism can be a solution for privacy invasion problem. In this paper, authors have designed the Role and Task based access control (RTBAC) model and developed the XML schema for representing the schema of the model. Basic conceptions and entities of RTBAC model include user, role, permission, privilege, task, dependency, application data, data object, and operation. The relationships among entities include user/role assignment (RU), role/privilege assignment (RP), task/role assignment (TR), and task/permission assignment (TP) etc. This model supports object privacy since it introduces a new constraint called Role and Task between subject and object. It supports more constraints on object's policy than current Role-based Access Control Model does.

Keywords XML · Role and task based access control · Role-based access control

69.1 Introduction

Recently, many enterprises are growing toward a heterogeneous, distributed environment. This has motivated many enterprises to adopt their computing services for efficient resource utilization, scalability and flexibility. The Extensible Markup Language (XML) is a good solution for supporting the policy enforcement in a heterogeneous, distributed environment. In addition, as wireless networking has become more common, ubiquitous computing begins to receive increasing attention as Internet's next paradigm [1]. Invisible and ubiquitous computing aims at defining environments where human beings can interact in an intuitive way with

G. Hong (✉) · W. Bai · S. Zhang
Department 5 of Wuhan Mechanical Technology College, Wuhan, China
e-mail: xinye5804@sohu.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_69,

surrounding objects [2]. The ubiquitous service should consider more frequent movement than current Internet services since the user can use various services anytime, anywhere. For these services, ubiquitous system needs to control a lot of information, which it leads to lots of privacy invasion problems.

Even though several security mechanisms are suggested for user privacy, none is yet de facto standard model.

69.2 Research on Access Control Techniques

The main purpose of access control is to restrict the use of resources. Only authorized user can access certain resources. To meet the access control requirements in enterprise environment, some aspects must be taken into account:

1. Realization of access control for sensitive information.
   There are many objects such as function modules, application menus; static or dynamic data and documents need access control.
2. Realization of access control for the workflow.
   The data in workflow are characterized by fluidity and time-restriction. They have different access control requirements under different circumstances or states. So the access control should be restricted at proper moments.
3. Solution to problem of position hierarchy of inner users.

The inheritance relationships among positions are involved both in access control of sensitive information and workflow information.

At present, the studies of access control concentrate on role-based access control (RBAC) and task-based authorization control (TBAC) [3–6].

The idea of RBAC is to connect the privilege with role, so the authorization can be realized by assigning the roles to certain users. The user's privileges are the union of roles' privileges which are assigned to the user. Relationships among roles should be restricted. The actual correspondence between users and their privileges may influenced by those relationships. RBAC is an ideal access control model for the administration of users, but there also many deficiencies because of the expansion of modern enterprise and the increasing number of system users. These increased potential risk and complexity in management and maintenance. By the requirements of access control in enterprise environment, RBAC has these disadvantages:

1. The model doesn't take into account the contextual environment.
   When the states of object changed, it may need other protection. When the states of user changed, it may have other restrictions.
2. It is a static access control model.

Because of the large amount of workflow information in enterprise environment, it is difficult to realize access control in practice.

TBAC can deals with workflow access control perfectly. With the task-orientated, it sets up security model and realizes security mechanism from view of task. It can dynamically manage privileges according to the states of tasks. Therefore, TBAC is suitable for distributed computing, information processing activities with multiple points of access, decision making in workflow and distributed process, and transaction management system. One shortcoming of TBAC is that it cannot deal with access control of non-workflow. In practical application, there are many users in enterprise. It is not suitable for using the trustee-set to manage them. There are also many hierarchy relationships among users. TBAC also cannot handle them perfectly.

With the development of network technologies, single security technique cannot guarantee the security of PMI system. The combination of other security techniques is the trend of access control development. It needs to develop a new model to meet the access control requirements for PMI in enterprise environment.

69.3 Role and Task Based Access Control

To meet the access control requirements in enterprise environment, this paper introduced a Role & Task Based Access Control Model (RTBAC). The aim of RTBAC is to realize a role and task based access control in enterprise environment, so as to protect the workflow application data and sensitive information, and suitable for enterprise application. The idea of selecting role and task as its basic elements is based on such a conception: the roles are assigned to users; the user may acquire the tasks according to the roles they represented; in processing the tasks, users are admitted to have the privileges of accessing objects required by tasks; at the same time, the user may get privileges of accessing static information of enterprise.

69.3.1 Role and Task (RT)

There are many common examples where access decisions must include other factors such as user attributes, object attributes, user tasks to other entities. The tasks among entities associated with an access decision are often very important. Roles of RBAC can be used to represent tasks. However, using roles to express tasks may be inefficient and/or counter intuitive. When roles cannot be used to represent tasks, it is common to program access decision logic directly into an application. Therefore, the task component should be included in the access control model.

Role and Task (RT) based access control model is the relation between Subject Entity (SE) and Object Entity (OE). For example, the RT may be accessing, browsing, read, write and so on. This RT is determined by OE and administrator.

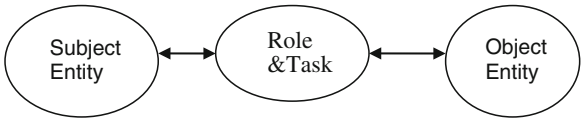


Fig. 69.1 The relation among subject entity, object entity and role and task

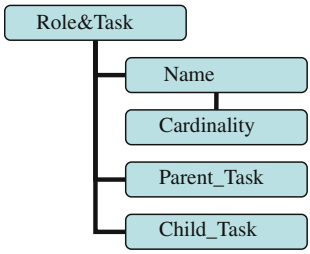


Fig. 69.2 XML schema of role and task hierarchy

SE is subjects such as users and devices that have a right to use some services in ubiquitous computing environment. OE is objects such as users and devices that are targets of service. Figure 69.1 shows the relation among subject entity, object entity and RT.

Definition1 (*Role and Task*) Role and Task *rt* is represented through a symbolic formalism and can be expressed as $rt = \{rt1, rt2 \ldots rti\}$ where *i* is integer.

69.3.2 XML Schema of Role and Task Hierarchy

Figure 69.2 show XML schema to represent RT hierarchy in RTBAC model.

69.3.3 RTBAC Model

RTBAC model is shown in Fig. 69.3.

Basic conceptions and entities of RTBAC include user, role, permission, privilege, task, dependency, application data, data object, and operation. The relationships among entities include user/role assignment (RU), role/privilege assignment (RP), task/role assignment (TR), and task/permission assignment (TP) etc.

User is the actor for task instance.
Role is the abstract descriptions for a certain user.
Permission is the executive right to operate the application data.
Privilege is the concrete right to operate the data object. It is divided into dynamic part and static part.
Task is a logic unit in the task flow.

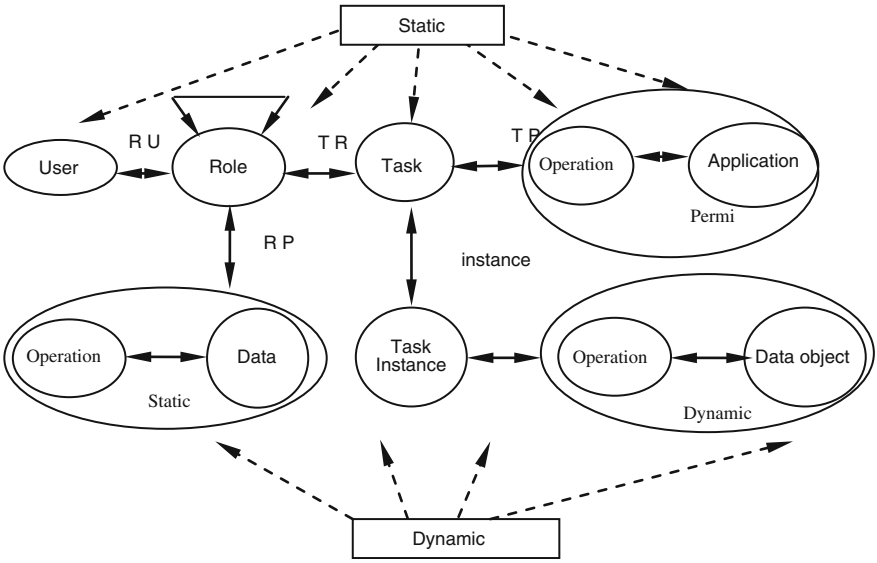


Fig. 69.3 Role and task based access control model

Application Data is the data used in the task flow.
Data Object is the concrete data, for instance a certain document or subscription.
Operation is the action of the user to the object, for instance read, write print etc.

For a concrete case, User administration module manages the registered users' basic information. Role definition module defines the possible roles in the system. Privilege definition module defines the operation privilege of objects. User/role assignment module attributes the defined privileges to the roles. Role/privilege assignment module realizes the privilege assignment. Task definition module segment tasks in process, defines the smallest security process unit, i.e. the authorization steps, assigns and binds them to due privilege and gives them corresponding permissions, and defines corresponding.

User set, role set, task role, operation set application data set, permission set respectively express as USERS, ROLES, TASKS, OPS, APDATA, PRMS. Role and task respectively express as r and t.

Relations among all the entity are as follows:

1. RU$\subseteq$ ROLES $\times$ USERS, users(r) = {u$\in$USERS$\wedge$ (u,r)$\in$RU};
2. RP$\subseteq$ROLES $\times$ PRMS,users(r) = {u$\in$USERS$\wedge$(r,p)$\in$RP};
3. TR$\subseteq$ TASKS $\times$ ROLES,roles (t) = {r$\in$ROLES$\wedge$ (t,r)$\in$TR};
4. PRMS = 2 (APDATA $\times$ OPS);
5. TP$\subseteq$TASKS $\times$ PRMS, permissions (t) = {p$\in$PRMS$\wedge$ (p,t)$\in$TP}.

In the process of access control, in order to avoid the risk of commercial deceit, the users, roles, permissions and tasks be better be restricted. Here come up four types of conflict entities: conflicting permissions, conflicting roles, conflicting

tasks, and conflicting users. These conflicts can be described by authorization constraint rules. Authorization constraint rules include static constraint rules and dynamic constraint rules.

69.3.3.1 Static Constraint Rules of RTBAC

Static constraint rules of RTBAC include:

1. One user cannot belong to conflicting roles;
2. Conflicting users cannot belong to conflicting roles;
3. One task cannot be processed by conflicting roles;
4. One task cannot assign conflicting permissions;
5. The number of users related to roles cannot exceed the role's cardinality.

69.3.3.2 Dynamic Constraint Rules of RTBAC

Dynamic constraint rules of RTBAC include:

1. Users can only execute the assigned task instances;
2. One user cannot executes the conflicting task instances in one process instance;
3. Conflicting users cannot execute conflicting task instances in one process instance;
4. Tasks in one binding task set must be executed by one user in one process instance.

RTBAC supports the two important security principles:

1. Least privilege principle
   In RTBAC model, the users have no connection with privileges before the task is instanced, only after that, are the concerning privileges stimulated. After the tasks are completed, the connection would be cancelled.
2. Duty separation principle.

In RTBAC model, the dynamic duty separation is realized by means of conflicting task set and binding task set. Additionally, RTBAC obeys the privilege abstract principle. The operations of application data should not be simply limited to read or write operations, all the processes can be abstracted as operations.

There are many key techniques that should be solved if RTBAC is to be applied in enterprise environment successfully. Considering practical application of RTBAC in enterprise environment, and to deal with the inheritance relation of roles among different ranks of positions, this paper divided the privileges into role privileges and private privileges, and borrowed some thoughts of EHRBAC which administrate the role privileges and private privileges through ordinary inheritance and extended inheritance; and introduced the object-role set and multi-division method to categorize the access control objects.

To meet the access control requirements in enterprise environment, this paper introduced a Role & Task Based Access Control Model (RTBAC). The aim of RTBAC is to realize a role & task based access control in enterprise environment, so as to protect the workflow application data and sensitive information, and suitable for enterprise application. The idea of selecting role and task as its basic elements is based on such a conception: the roles are assigned to users; the user may acquire the tasks according to the roles they represented; in processing the tasks, users are admitted to have the privileges of accessing objects required by tasks; at the same time, the user may get privileges of accessing static information of enterprise.

69.4 Conclusion

In this paper, authors designed the Role and Task based access control (RTBAC) model and developed the XML schema for representing the schema of the model. This model supports object privacy since it introduces a new constraint called Role and Task between subject and object. It supports more constraints on object's policy than current Role-based Access Control Model does. This model solves the privacy invasion problem, which is more frequently involved in ubiquitous computing environment than current Internet services. This model defines new concepts such as Role and Task and the strength value of Role and Task. The Role and Task is relation between subject and object. The strength value of Role and Task is degree of strength relation between subject and object. It supports more constraints on object's policy than current RBAC. It is able to represent complex and specific policy.

References

1. Stajanoand, F., Anderson, R.: The resurrecting duckling security issues for ubiquitous computing. IEEE Security and Privacy, pp. 22–26 (2002)
2. Bussard, L., Roudier, Y., Molva, R.: Untraceable secret credentials: Trust establishment with privacy. The Second IEEE Annual Conference on Pervasive Computing and Communications Workshops, pp. 122–126 (2004)
3. Lu, H., Xia, T.: The research of role tree-based access control mode. J. Donghua Univ. (Eng. Ed.) **27**(2), 274–276 (2010)
4. Chae, S.-H., Kim, W.: Semantic representation of RTBAC: Relationship-based access control model. APWeb/WAIM 2007 Ws, LNCS 4537, pp. 554–563 (2007)
5. Wang, L., Xie, X.-Y., Yang, Y.-Z., Zhang, A.Y.: An authorization management model based on RTBAC. Guizhou Sci. **27**(3), 51–53 (2009)
6. Huang, G., Sun, L.-S.: An access control framework for reflective middleware. J. Comput. Sci. Technol. **23**(6), 895–904 (2008)

Chapter 70
New Immersive Display System Based on Single Projector and Curved Surface Reflector

Xiao-qing Yin, Ya-zhou Yang, Zhi-hui Xiong, Yu Liu and Mao-jun Zhang

Abstract An immersive display system based on single projector and curved surface reflector is presented in this paper. In this system, the light illuminated by one projector is reflected by a curved surface reflector to a cylindrical rear projection screen. This system implements uniform enlargement of the projected image on both the horizontal and vertical direction and displays virtual scene of large continuous field of view. With the help of the curved surface reflector, the image distortion caused by curvature of the screen can be eliminated without using complicated image transformation algorithms. Additionally, seamless image can be obtained by using single projector. The projection experiment proves that this system can achieve satisfactory immersive display quality.

Keywords Immersive display system · Curved surface reflector · Single projector · Rear projection

70.1 Introduction

The development of immersive display system has been accelerated by many applications including 3D virtual navigation, teleconferencing and training simulation. People staying in the immersive display system can perceive the surrounding environment as they did in the real world [1, 2].

There has been plenty of research on immersive display systems [3–7]. A summary of immersive display technology is presented by Huang et al. [3]. The large scale display wall designed by Princeton University uses multi-screen

X. Yin (✉) · Y. Yang · Z. Xiong · Y. Liu · M. Zhang
College of Information System and Management,
National University of Defense Technology,
Changsha, China
e-mail: happyyxq2012@sina.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_70,

and multi-channel parallel output technology and it achieves certain immersive feeling [4]. However, participants can only look on the planar picture shown on the display wall other than stay in the virtual environment, in which the immersive feeling is limited. i-Cone^{TM} creates an extended workplace for participants by conical screen geometry, optimizing projector placement and curved screen display with a large continuous field of view [5]. However, the movement of the participants is limited because their shadows may fall on the screen. Additionally, the distortion correction of this system is difficult to accomplish. Some surrounding display systems based on projectors and reflectors can achieve certain immersive display effect [6, 7]. In multi-projector display systems, it often takes a long time to place the projectors in the right locations and there are many image mosaics problems including image registration and color calibration [8–10]. Approaches to the problems involve complicated algorithms and high-class graphics hardware. Seamless image can be obtained in this system by using single projector, which can avoid the problems of multi-projector display systems.

A new immersive display system based on single projector and curved surface reflector is presented in this paper in order to overcome the disadvantages of traditional immersive display systems. This system implements uniform enlargement of the projected image on both the horizontal and vertical direction and displays virtual scene of large continuous field of view. Moreover, the participants can acquire more movement freedom by means of rear projection.

70.2 System Construction

This system includes four modules: projector curved surface reflector, cylindrical rear projection screen and computer, as shown in Fig. 70.1. Participants are on one side of the screen and are surrounded by it while the projector and the reflector are on the other side. The light illuminated by the projector is reflected by the curved surface reflector to the screen. The optic axis of the projector is perpendicular to the ground. Image forms on the cylindrical rear projection screen and participants observe the image behind the screen.

The screen is a cylindrical section whose radius is R and the total field of view is fv (degree). The width of the screen is $w_s = R \times fv \times \pi/180$ and the height is h_s. Thus the size of the projection picture on the screen is $w_s \times h_s$. This screen can be manufactured by pasting a large piece of rear projection curtain onto a cylindrical steel bracket. The horizontal distance between the projector lens and center of the screen is d_{ls}. The field angle of the projector is γ and the length-width ratio is 4:3. The curved surface reflector is placed above the projector.

The principle procedure of immersive displaying is as follows:

(1) As the width-height ratio of the image on the screen is $r_{wh} = w_s/h_s$, part of the projected image with a length-height ratio of r_{wh} should be selected from the original image, which is shown in Fig. 70.2. This part of image is the wide-

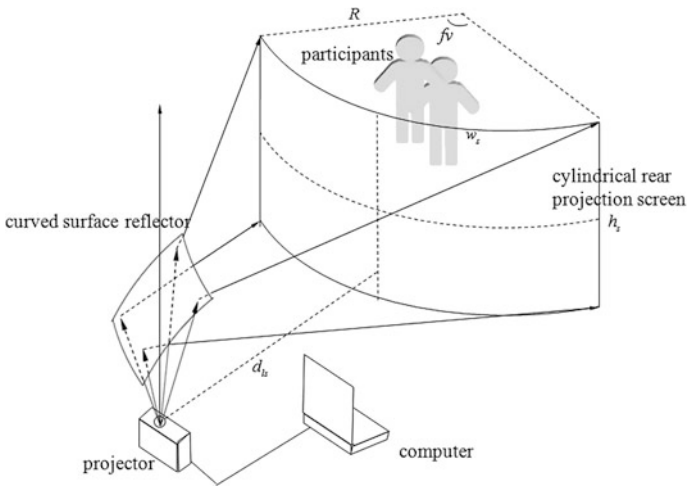


Fig. 70.1 Structure of this system

Fig. 70.2 Part of the projected image is selected

angle scene to be displayed. In order to improve the immersive feeling effect, the objects appearing in this image are supposed to be in the same size as they are in reality when projected on the screen. A rectangular part of size $w_s \times h_s$ (in reality) should be selected from the original image, according to the actual size of the objects.

(2) This part of image should be shrunk uniformly on the horizontal direction, making the length-width ratio be 4:3 before it is projected (Fig. 70.3). It is because the projector can only project image with the length-width ratio of 4:3.

(3) The image is projected by the projector and the light from the projector is reflected by the curved surface reflector. By properly designing the curved surface reflector, the projected image can be uniformly enlarged on both the horizontal and the vertical direction. Then the participants can observe the

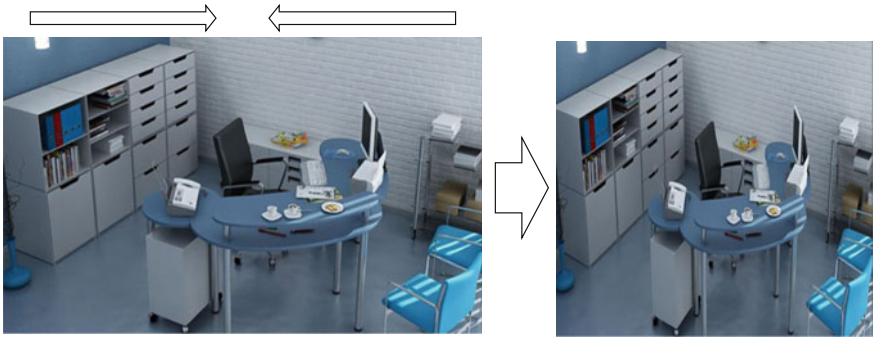


Fig. 70.3 Uniform shrink of image on the horizontal direction and making length-width ratio be 4:3

image formed on the cylindrical rear projection screen, which is shown in Fig. 70.7.

The key to implementing this immersive display system is designing the curved surface reflector, which will be analyzed in the next section.

70.3 Designing the Curved Surface Reflector

In actual application, immersive display systems are often used to display real-time video image. As image transformation algorithms can be quite complicated and may affect the display quality of the video. It is necessary to simplify the algorithms in order to reduce the computing time. The algorithms can be simplified and implemented by transforming part of image transformation function to the shape of curved surface reflector. By properly designing the curved surface reflector, the projected image can be uniformly enlarged on both the horizontal and the vertical direction. The coordinate system can be built according to Fig. 70.4.

In Fig. 70.4, the pinhole of the projector's lens is O, which is at the origin of the coordinate system. The optical axis is the y-axis. All the incoming rays go through point O. Incoming ray OP is reflected by point $P(x, y, z)$ on the reflector and the reflected ray goes through point $S(x_s, y_s, z_s)$ which is on the screen. The projection picture on the cylindrical rear projection screen is denoted as P_c and the center of P_c is $O_c(x_0, y_0, z_0)$. The value of x_0 is set to be 0, which means that the shape of the projection picture is symmetric to z-axis. According to the symmetrical characteristic of the projection picture and the ray casting, the surface of the reflector should be symmetric to y-axis.

Assume that there is an imaginary planar projection screen above point O, which is perpendicular to the z-axis. Half of the imaginary screen and the incoming ray are illustrated in Fig. 70.5. The intersection point of the screen and the z-axis is $C(x_c, y_c, z_c)$. The distance from point O to the imaginary planar

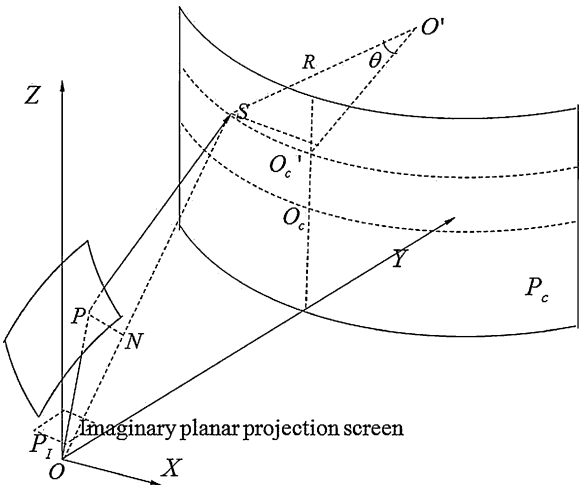


Fig. 70.4 Building coordinate system

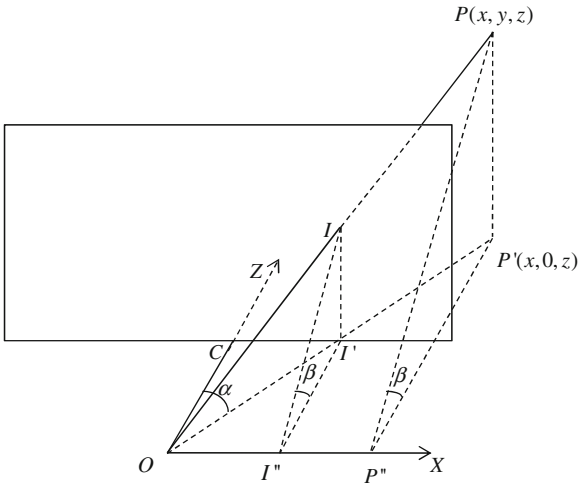


Fig. 70.5 Imaginary planar projection screen used to analyze the enlargement of projection picture

projection screen is $|OC|$. Set $|OC| = 1$, thus $x_c = 0, y_c = 0, z_c = 1$. Denote the projection picture on the imaginary planar projection screen as P_I. It is obvious that point C is the center of the imaginary projection picture.

This imaginary screen is constructed to analyze the enlargement of the projection picture. The times of enlargement can be obtained by comparing the size of P_c with that of P_I. The width of P_I is $w_i = 2|OC|tg(\gamma/2)$ and the height of P_I is $h_i = 3w_i/4$. Define horizontal enlargement factor e_h as the ratio of the width of P_c and that of P_I : $e_h = w_s/w_i$. Similarly, define vertical enlargement factor e_v as the ratio of the height of P_c and that of P_I : $e_v = h_s/h_i$.

Assume that the intersection point of the incoming ray OP and the imaginary screen is $I(x_I, y_I, z_I)$. The projections of P and I on XOZ plane are P' and I'. From

P' and I' draw vertical lines until intersecting with the x-axis. The intersection points are P'' and I''.

Additionally, $\angle COI' = \alpha$ and $\angle II''I' = \angle PP''P' = \beta$. Hence: $x_I = |OC| \cdot tg\alpha = x/z$, $y_I = -|I'I''|tg\beta = -y/z$.

The incoming ray OP goes through point I on P_I and the reflected ray goes through point S on P_c. The relations between the coordinates of I and S will be analyzed on both the horizontal direction and the vertical direction.

As it is discussed above, $C(x_c, y_c, z_c)$ and $O_c(x_0, y_0, z_0)$ are the centers of P_I and P_c respectively. Relative coordinate systems can be constructed in P_I and P_c with the origin of C and O_c respectively. Thus the relative x-coordinate of I is $x_I - x_c = x_I$. Consider point O'_c with coordinates of $O'_c(0, y_0, z_s)$. The radius is R and the center of the circle is O' and $\angle SO'O'_c = \theta$. The relative x-coordinate of S is arc length $\overset{\frown}{SO'_c}$, which can be computed by the following equation:

$$\overset{\frown}{SO'_c} = R\theta = R\arcsin(\frac{x_s - x_0}{R}) = R\arcsin(\frac{x_s}{R}) \tag{70.1}$$

In order to implement uniform enlargement of the projection picture, the x-coordinate of any point on P_c should be proportionably enlarged compared to the x-coordinate of the corresponding point on P_I. On the horizontal direction, the relation between $\overset{\frown}{SO'_c}$ and x_I can be written in the following form:

$$\overset{\frown}{SO'_c} = e_h x_I \tag{70.2}$$

where e_h is the horizontal enlargement factor which is introduced above. Furthermore, it can be obtained in Fig. 70.4 that $y_s = y_0 + R(1 - \cos\theta)$.

On the vertical direction, the relative y-coordinate of I is $y_I - y_c = y_I$. The relative z-coordinate of S is $z_s - z_0$. Similar to (70.2), the following equation can be obtained:

$$z_s - z_0 = e_v y_I \tag{70.3}$$

where e_v is the vertical enlargement factor. As shown in Fig. 70.4, OP is the incoming ray and PS is the reflected ray. Connect O and S. The intersection point of OS and the surface normal which goes through point P is $N(x_N, y_N, z_N)$. Denote the surface of the reflector as $z = f(x, y)$. The direction vector of the normal PN is $\overrightarrow{k} = (\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, -1)$. Thus the coordinates of N can be expressed as $(x + \frac{\partial f}{\partial x}t, y + \frac{\partial f}{\partial y}t, z - t)$, $t \in (0, +\infty)$.

$ON = (x + \frac{\partial f}{\partial x}t, y + \frac{\partial f}{\partial y}t, z - t)$ and $OS = (x_s, y_s, z_s)$. As O, N, S are collinear, we can get the following equations:

$$(x + \frac{\partial f}{\partial x}t)z_s = (z - t)x_s \tag{70.4}$$

$$(y + \frac{\partial f}{\partial y}t)z_s = (z - t)y_s \tag{70.5}$$

According to (70.4), $t = (x_s z - x z_s)/(\frac{\partial f}{\partial x} z_s + x_s)$. Hence:

$$x_N = x + \frac{\partial f}{\partial x} t = x + \frac{\partial f}{\partial x} \frac{x_s z - x z_s}{\frac{\partial f}{\partial x} z_s + x_s} = \frac{x_s (x + \frac{\partial f}{\partial x} z)}{x_s + \frac{\partial f}{\partial x} z_s} \tag{70.6}$$

According to angle bisector theorem, the following equation can be obtained:

$$\frac{|OP|}{|PS|} = \frac{|ON|}{|NS|} = \frac{x_N}{x_s - x_N} \tag{70.7}$$

According to (70.6) and (70.7), $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ can be expressed as:

$$\frac{\partial f}{\partial y} = \frac{|OP| y_s - (|OP| + |PS|) y}{(|OP| + |PS|) z - |OP| z_s} \quad \frac{\partial f}{\partial x} = \frac{|OP| x_s - (|OP| + |PS|) x}{(|OP| + |PS|) z - |OP| z_s} \tag{70.8}$$

Assume that the intersection point of z-axis and the curved surface reflector is$(0, 0, z_{ini})$. The initial condition of partial differential equations (70.8) is:

$$f(0, 0) = z_{ini} \tag{70.9}$$

Difference equations can be constructed based on the partial differential equations:

$$\frac{\partial f}{\partial x} \approx \frac{f(x + h, y) - f(x, y)}{h} \quad \frac{\partial f}{\partial y} \approx \frac{f(x, y + h) - f(x, y)}{h} \tag{70.10}$$

Solving equations (70.8)–(70.10), the surface of the reflector can be obtained.

70.4 Result

In this system, the values of the variable are shown in Table 70.1. The surface of the reflector is shown in Fig. 70.6.

Table 70.1 The values of the variables

Variable	Value
R	1.50 m
fv	140°
w_s	3.66 m
h_s	1.80 m
h_p	1.00 m
d_{ls}	2.00 m
γ	30°
(x_0, y_0, z_0)	$(0, 200, 100)$(cm)

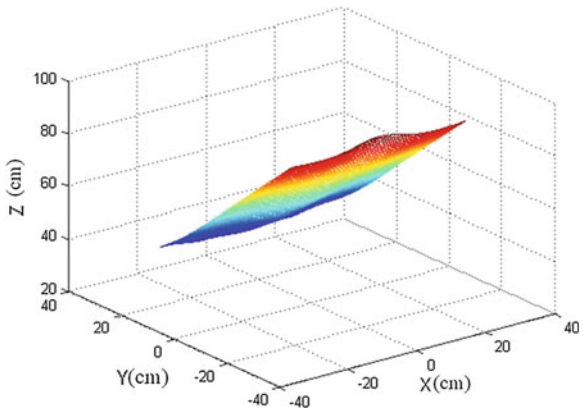


Fig. 70.6 The surface of the reflector.

Fig. 70.7 Immersive display in **a** Virtual roaming, **b** Virtual training and **c** 3D game

Using the reflector designed above, the immersive display effect which is obtained by 3ds Max simulation is shown in Fig. 70.7. In Fig. 70.7, the objects in the surrounding environment are in real size and the participant is surrounded by the virtual scene, which can make the participant feel like staying in the real world.

70.5 Conclusion

The immersive display system presented in this paper employs curved surface reflector to reflect projecting light from a single projector, and to implement the displaying of wide-angle virtual scene. The projection experiment proves that this system can achieve satisfying immersive display quality.

Using single projector to project image with a large continuous field of view, it is necessary to choose a projector with higher display resolution. Improving the immersive display quality and putting this system into application will be considered as the future work.

References

1. Slater, M., Steed, A., Chrysanthou, Y: Computer Graphics and Virtual Environments: From Realism to Real-Time. Addison Wesley/Pearson, Boston (2001)
2. Rheingold, H.: Virtual Reality. Simon & Schuster, New York (1991)
3. Huang, D.J., Bo, S.K., Chen, B.H.: Research on immersion display technology study. Comput. Syst. Appl. **3**, 43–46 (2007)
4. Li, K., Chen, H., Chen, Y.Q., Clark, D.W., Cook, P., Damianakis, S., Essl, G., et al.: Building and using a scalable display wall system. IEEE Comput. Graph. Appl. **20**, 29–37 (2000)
5. Simon, A.: The i-Cone^{TM}: a panoramic display system for virtual environments. In: Proceedings of the 10th Pacific Conference on Computer Graphics and Applications, pp. 3–7 (2002)
6. Henden, C., Champion, E., Muhlberger, R., Jacobson, J.: A surround display warp-mesh utility to enhance player engagement. Comput. Sci. **5309**, 46–56 (2009)
7. Paul, B.: Spherical mirror: a new approach to hemispherical dome projection. Planetarian **34**, 6–9 (2005)
8. Hashimoto, N., Jeong, S., Takeyama, Y., Sato, M.: Immersive multi-projector display on hybrid screens with human-scale haptic and locomotion interfaces, Proceedings of the 2004 International Conference on Cyberworlds, pp. 361–368 (2004)
9. Chen, L.Y., Chang, H., Dai, S.L.: A survey of building multi-projector tiled display systems. In: Proceedings of the 4th Virtual Reality and Visualization Conference of China, pp. 250–254 (2004)
10. Harville, M., Culbertson, B., Sobel, I., Gelb, D., Fitzhugh, A., Tanguay, D.: Practical methods for geometric and photometric correction of tiled projectors on curved surfaces. In: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop, p. 5 (2006)

Part IV
Data Processing

Chapter 71
Self-Adaptive Cloth Simulation Method Based on Human Ring Data

Wenhua Hou and Bing He

Abstract As an integral part of Computer Graphics, cloth simulation technology has been extensively applied in many fields, such as computer games, entertainments, film special effects and computer animation productions. The cloth simulation technology can realize efficient and vivid simulation as it draws upon the latest achievements in computer graphics, applied mathematics and engineering mechanics. In this paper, researchers adopt finite element method to simulate cloth simulation model, and make an in-depth discussion and study on self-adaptive time step and human body collision processing. Researchers have improved cloth simulation implicit integration method and proposed a collision detection algorithm based on human ring data, which achieved a faster and more realistic result.

Keywords Cloth simulation · Self-adaptive · Collision processing

71.1 Introduction

In the field of clothing simulation, the realistic effects and efficient algorithm are always contradictory. Recently, researches on cloth simulation focus on how to improve the existing physical model [1, 2], calculation methods and cloth-body collision processing, in order to show the details of the clothing, and accelerate the efficiency of cloth simulation.

Implicit integration or explicit integration method can be adopted to solve cloth kinetic equation [3, 4]. The explicit integration method is closer to the actual kinetic rule, and it allows adopting larger iteration step in cloth simulation to reduce iteration times and generate accurate simulation results, so it has

W. Hou (✉) · B. He
State Key Laboratory of Virtual Reality Technology and Systems,
BeiHang University, BeiJing, China
e-mail: houwh@vrlab.buaa.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_71,

advantages in algorithm stability and iteration convergence. In this paper, we put forward a self-adaptive time step detection method and achieved a better visual effect in terms of detail performance.

Cloth-body collision can be seen as external collisions between complex flexible body and complex rigid body. In cloth simulation, hierarchical bounding box in octree which is the same as used in the cloth is usually adopted to organize somatic data, so this algorithm has a certain generality and also time-consuming [5]. In this article, we proposed a collision detection algorithm based on human ring data, which simplifies somatic data organization and improves collision detection efficiency.

71.2 Cloth Simulation Method Based on Self-Adaptive Time Step

In this paper, we put forward a self-adaptive time step detection method, that is, to correct the next simulation time through collision prediction, and then calculate rational time step.

71.2.1 Self-Adaptive Time Step Based on Collision Prediction

Each time we make simulation computation, position and velocity can be updated, and we make collision detection prediction based on the current time step. Multiply current velocity of each unit node by time step to get predicted node position. Instead of updating the node position immediately, we carry out collision prediction on all unit nodes at the current predicted node position.

Collision involves two finite elements. As shown in Fig. 71.1, penetration occurs, and the finite element corresponding to the current node X collides with finite element ABC. We can get collision position through collision detection algorithm.

Constant step algorithm can directly update current position of a node to the collision position, thus getting the new velocity and direction through impulse model so as to ensure that no collision case will occur after elapsed time. However, such treatment method overlooks simulation details in the time, while other collisions may occur during the time when step is large.

Through current position collision prediction, we can get a predicted collision time for each node. Adopting this time as time step of the next simulation ensures that no possible collisions may be ignored during the time, thus getting more simulation details.

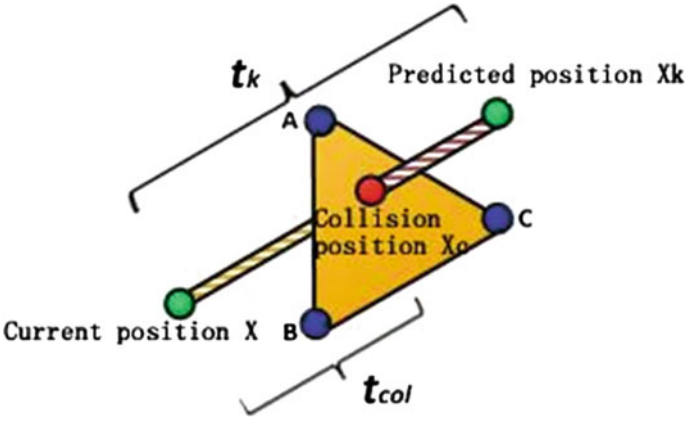


Fig. 71.1 Collision prediction schematic diagram

71.2.2 Improved Self-Adaptive Time Step

Simply adopting the minimum collision time as time step of the next simulation step will stagnate simulation. Suppose the collision time is very short, that means that although no collision occurs at present, simulation step must be updated within a rather short period of time, while two adjacent steps are different. Under limiting circumstances, if collision occurs at the present moment, simulation step of the next frame will be set as zero.

Therefore, the simulation step should change along with collision prediction time, but it should not be changed too fast so as to guarantee both details and smooth simulation process.

To this end, a simulation step threshold interval is set in this article, and this interval is updated during iteration computations so as to ensure that the simulation time at the next step can only change within this interval. The main steps are recording collision prediction time of all nodes, adding up all items within the interval to calculate average time, using this time as simulation time of the next step, making collision processing on all nodes which are less than this, and updating velocity and position.

In a simulation model with frequent collision, the simulation time will be finally stood at a constant value. The root cause of such phenomenon is that the simulation time can only be shorter instead of longer if we use the above self-adaptive method.

To solve this problem, a step increment is added in this article. During each simulation, record the quantity of nodes with collision, and compare them with that detected during the last collision. If the current greater than the front, it suggests that there is a trend of collision increase. In this circumstance, we need to reduce simulation step to add simulation details. Otherwise, it suggests that there is

a trend of collision decrease. In this circumstance, we need to enlarge simulation step.

By adopting the above strategy, the simulation step can be adjusted as per current collision situation, thus making simulation more vivid while guaranteeing simulation stability.

71.3 Collision Detection Based on Human Ring Data

In this paper, we build human ring data structure through pre-treatment somatic data characteristics, which effectively simplifies somatic data organization and improves collision detection efficiency.

71.3.1 Human Ring Data Pretreatment

Somatic data can be expressed by different organization forms including parametric curve, triangular mesh and ring mesh. The organization form of somatic data will affect the process of cloth patch and human body collision processing, and improper data structure may reduce simulation velocity or even cause collision penetration and other distortions.

We adopt ring mesh data proposed by Wang etc. to organize human body patches [6].

Compared with the traditional method of organizing somatic data in a tree structure (such as spatial quadtree and octree structure) [7], the ring data hierarchies fit well with physiological features of human body, as it divides human body into different parts, and each part is composed of different circles. These ring data interact with human body in a more convenient and flexible way, which suits well real dressing process. Real cloth is made up of warp threads and weft threads; similarly, human skin, muscle and other parts are also arranged with skeleton at the center, human ring data in combination with linear mesh model of cloth reflect this corresponding relation, so human ring data have obvious advantages in both simulation precision and efficiency.

For pre-treatment process of human ring data, see Fig. 71.2. We use a cross section to cut the human model and obtain a slice. The distance between each slice is h, and p is the center of gravity of section ring. Each slice is formed by connecting the point set.

Somatic data is complex, but its geometric structure features can be used for data simplification. Head, hands and feet of a human body generally do not involve in cloth-body collision, so these parts can be ignored. Besides that, somatic data are divided into nine major parts, they are, trunk, left arm, left forearm, right arm, right forearm, left thigh, left shank, right thigh and right shank. We conduct ring data pretreatment on these nine parts to generate a series of regular array of

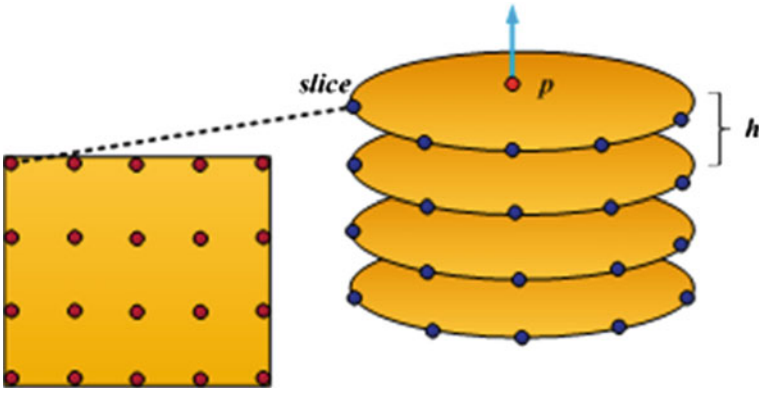


Fig. 71.2 Human ring data organization

striping data, then organize these data in a rational way through linear data structure and set up bounding box information.

All data are divided into different sections and cut into different rings, and each ring has certain basic properties such as position of the center of gravity, normal line direction, circle index, bounding box information. As human body may be seen as static rigid body during cloth modeling, these parameters are constant quantities, thus avoiding calculating repeated values and improving simulation efficiency.

71.3.2 Cloth-Body Collision Detection Based on Human Ring Data

At stage of rough detection on external cloth-body collision, we need to get rid of impossible collision primitive pairs based on the bounding box information.

Hierarchical bounding box in octree are adopted to organize cloth data, and ring data is used to organize somatic data. By referring to inter-octree collision detection method, we adopt the following steps to detect cloth-body collision:

Step one, determine human body collision part. Carry out bounding box detection on root node of hierarchical bounding box in octree and different parts of human body, so that we can acquire information on parts which might have cloth-body collision, and then record the part number.

Step two, determine interval of human ring with collision. For human body part with possible collision, calculate the sum of the two projective points of cloth bounding box on central axis of human ring data, then divide the distance between circles by the distance between the projective point and the initial central point to get circle number with possible collision,

$$number1 = \frac{|x_{s1} - x_c|}{h} \tag{71.1}$$

$$number2 = \frac{|x_{s2} - x_c|}{h} \tag{71.2}$$

In this way, we get the ring interval with possible collision between cloth and sections of human body, and other ring data of human body can be get rid of during collision detection.

Step three, iterative intersection. For non-leaf node of current cloth with octree structure, repeat the intersection process as stipulated in step two, calculate the result of bounding box detection at sub-node and sections within the ring collision interval. The leaf node of the cloth with octree structure is a triangular patch of cloth, detection on collision between primitives of the triangular patch of the cloth and that of the human ring structure needs to be carried out to get primitive pair with actual collision and make the corresponding collision response.

Conducting cloth-body collision detection based on ring data is a simplified algorithm. We can position bounding box information of somatic data only by virtue of section number and circle number during computation, and can quickly get rid of areas without possible collision according to the bounding box information, thus improving collision detection efficiency. Triangular patch of each human ring data is limited in quantity, so it will not take long for us to carry out full traverse.

Due to complex three-dimensional structure of human body and discrepancy of data, if we adopt hierarchical bounding box in octree which is deep, it will affect retrieval efficiency during cloth-body collision detection. Compared to this, human ring data is generated based on special structure of human body, it maintains basic form of data while improves cloth-body collision efficiency, so it is a feasible and simplified method targeting at human body.

71.4 Experiment Result and Analysis

Hardware environment of the experiment are: CPU, Intel Core Q6600, 2.40 GHz; internal memory: 4.0 GB.

Software environment of the experiment are: Windows XP Professional operating system; programming environment: Visual Studio 2008.

71.4.1 Experiment I: Cloth Simulation Effect Based on Self-Adaptive Time Step

In experiment I, we simulate motor process of a cloth, the cloth moving under gravity collides with the sphere below it and generates corresponding deformation drape. The cloth is initially placed horizontally. The cloth mesh is made of

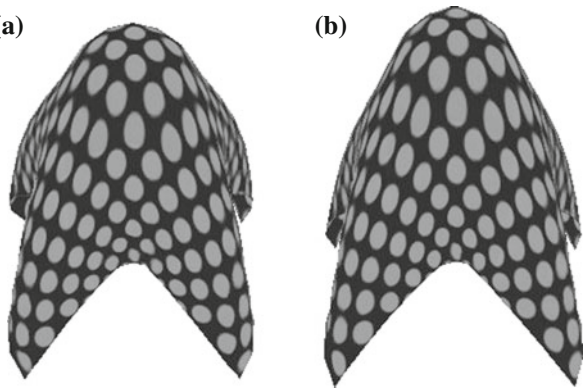


Fig. 71.3 Simulation effect of self-adaptive integration step **a** strategy 1, **b** Strategy 2

Table 71.1 Self-adaptive simulation step under different strategies

Frame number	Simulation step of strategy 1 (s)	Simulation step of strategy 2 (s)
1	0.05	0.05
50	0.046	0.043
100	0.04	0.04
150	0.031	0.034
200	0.023	0.035
250	0.016	0.043
300	0.012	0.041
350	0.01	0.033
400	0.01	0.035
450	0.01	0.037

33×33 nodes, finite element quantity is 2,048, piecewise linear constraints method is adopted for elastic modulus, Poisson's ratio is 0.25, constant simulation steps is taken as 0.001, the maximum and minimum threshold of step are 0.05 and 0.001 respectively, and step increment is 0.001.

Self adaptive time steps can be adopted with two strategies. In simulation adopting strategy 1, the simulation step may decrease due to inspected collision, and it will not change when reaching the minimum threshold. According to collision trend predicted based on collision quantity, strategy 2 increase or decrease simulation step as per preset step variation, thus realizing the goal of self-adaptive adjustment of simulation time.

Figure 71.3 shows the experiment results adopting these two strategies. From Fig. 71.3 we can see that, the two strategies have similar overall effect, but strategy 2 has better visual effect in terms of detail performance.

Table 71.1 records average simulation step of different frames.

Besides, due that simulation time changes constantly, it is not proportional to actual drawing time, thus simulation animation effect is discontinuous. If we want to get smooth simulation animation effect, we need to store position of all

simulation times to generate continuous animation finally, and have intermediate data as per difference strategy. Further discussion in this regard will be conducted in the future.

71.4.2 Experiment II. Cloth-Body Collision Treatment Method

The T-shirt model adopted in experiment II has 6,750 vertexes, 13,312 triangular facets, manikin is adopted, and human body is organized by using hierarchical bounding box in octree and ring structure.

Figure 71.4 is a screenshot of one fine cloth simulation effect. During motion solving, we use self-adaptive time step based on collision prediction and Poisson ratio is 0.25. The average simulation time is 1,560 ms by using hierarchical bounding box in octree structure, while using human ring data structure, the average time is 1,433 ms. Experiment result suggests that the comprehensive method described in this article can better handle with cloth simulation problems and can achieve certain simulation effect.

Nonetheless, human ring data have certain limitations, hierarchical bounding box in octree method has fast velocity at some simulation times when the manikin is relatively simple and regular, but when the manikin becomes more complex, it is not as fast as ring data can do in terms of retrieval speed due to deep octree human body structure. The experiment illustrates that simplified human ring data can improve collision detection efficiency.

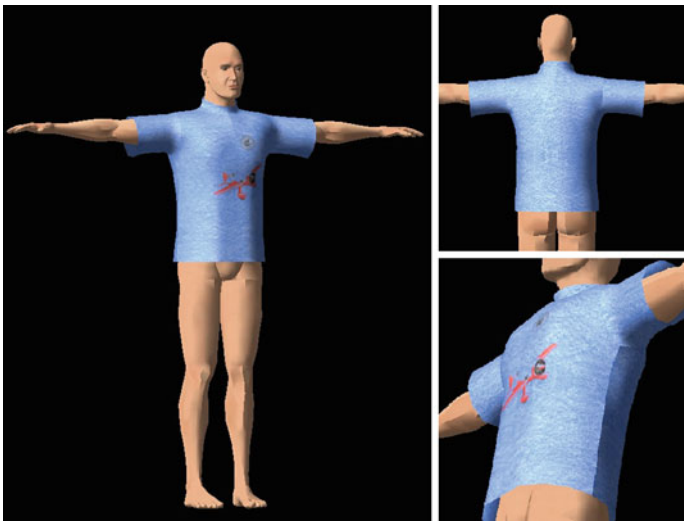


Fig. 71.4 Cloth simulation effect

71.5 Conclusions and Prospect

Based on kinetics rules and the characteristics of finite element cloth simulation model, researchers compute velocity solution by using implicit integration method, and put forward a self-adaptive time step adjustment strategy based on collision prediction. The strategy can add simulation details while ensuring stability of simulation algorithm.

Meanwhile, an external collision detection strategy based on human ring data is also proposed in this article. Through this strategy, somatic data are divided into different parts, and each part corresponds to different circles. One can quickly position a part of human body through interval number and circle number, and the corresponding bounding box information can be acquired. Thus the strategy effectively simplifies somatic structure and improves collision detection efficiency.

Although research results mentioned above have been achieved in this article, still there is room for further discussion and research on cloth simulation technology, which mainly include:

1. If ordinary bounding box data is adopted in circular somatic data treatment, collision detection efficiency is low, thus a tighter bounding box (for example, ellipsoidal bounding box) can be adopted as it suits well with structural rules of human body and can be quickly positioned through axis length during rough collision detection, and as a result, improves detection efficiency.
2. Cloth simulation with multiple-resolution has become a hot topic of research over the past 2 years. Through rough simulation, researchers get cloth mesh data, through which they can compute finer mesh form. In the future, researchers may add cloth simulation with multiple-resolution into the existing algorithm frame, and try to optimize the final simulation effect by adopting data-driven algorithm.

Acknowledgments This work was supported by grant No. 61272346 from NSFC (National Natural Science Foundation of China).

References

1. Volino, P., Megnenat-Thalmann, N., Faure, F.: A simple approach to nonlinear tensile stiffness for accurate cloth simulation. ACM Trans. Graph. **28**(4), 300–309 (2011)
2. Kavan, L., Gerszewski, D., et al.: Physics-inspired upsampling for cloth simulation in games. ACM Trans. Graph. **30**(4), 311–321 (2011)
3. Baraff, D., Witkin, A.: Large steps in cloth simulation. In: Proceedings of SIGGRAPH, pp. 43–54 (1998)
4. Terzopoulos, D., Platt, J., et al.: Elastically deformable models. In: Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, pp. 205–214 (1987)
5. Teschner, M., Kimmerle, S., Heidelberger, B.: Collision detection for deformable objects. Computer Graphics Forum pp. 61–81 (2005)

6. Wang, J., Lu, Guodong, et al.: Interactive 3D garment design with constrained contour curves and style curves. Comput. Aided Des. **41**(9), 614–623 (2009)
7. Tang, M., Curtis, S., et al.: Interactive continuous collision detection between deformable models using connectivity-based culling. IEEE Trans. Visual Comput. Graph. **15**(4), 544–552 (2009)

Chapter 72
Combination Approach of SVM Ensembles and Resampling Method for Imbalanced Datasets

Xin Chen, Yuqing Zhang and Kexian Wu

Abstract Datasets in real world are often predominately composed of normal examples with only a small percentage of interesting or abnormal examples. A new approach is applied in this paper to address the imbalance problem by combining SVM ensembles and resampling method. Through empirical analysis, researchers cluster majority classes by k-means algorithm into subclass which decreases the imbalance ratio. Additionally, they use resampling method which concludes oversampling and undersampling techniques to deal with the problem of long training time and low training efficiency in SVM ensembles. Experimental results show that the SVM ensembles with resampling method outperform individual SVM classifiers. The proposed combination approach can effectively solve the imbalance problem.

Keywords SVM ensembles · Imbalanced datasets · K-means · SOMTE

72.1 Introduction

In a classification problem, the datasets is said to present a class imbalance [1] if at least one of the classes is represented by significantly less number of instances than the others. Examples of applications with such datasets include, but are not limited to, text categorization [2], identifying fraudulent credit card transactions [3], detecting certain objects from satellite images [4] and telephone fraudulent detection [5]. In such applications, the class boundary learned by standard learning algorithms can be severely skewed toward the minority class, which means that the classifier misclassifies the minority class instances.

X. Chen · Y. Zhang (✉) · K. Wu
China University of Geosciences, Beijing, China
e-mail: yqzhang@cugb.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_72,

This paper is concerned with improving the performance of the Support Vector Machines [6] (SVM) on imbalanced datasets. We propose a complementary method and study the ensemble techniques as well as use of sampling to deal with the class imbalance problem. Firstly, we take an ensemble of SVM based on k-means algorithm [7] to boost the performance. Our observation indicates that SVM performance is influenced by the inhomogeneity of spatial distribution of minority instances. A collection of SVM is trained individually by k-means clustering on minority class, and the final prediction is obtained by combining the results from those individual SVM. By this mean, more robust results can be obtained by alleviating the information loss due to sampling, as well as by reducing the randomness induced by a single classifier. Furthermore, we propose to integrate the two types of sampling strategies by oversampling the minority class to a middle extent and undersampling the majority class to the similar size. The experimental results demonstrate that the approach, as a supplement of ensemble SVM, reduces the training time obviously. In the rest of the paper, we constrain our discussion to a standard two-class classification problem and refer to the minority and majority classes as "positive" and "negative" respectively.

72.2 Related Works

Recent researches on class imbalance problem have focused on two major directions. One is to resample the original training datasets, either by oversampling the positive class and/or undersampling the negative class until the classes are approximately equally represented [8]. As one of the successful resampling approaches, the SMOTE [9] (Synthetic Minority Oversampling Technique) oversamples the positive class by generating interpolated data. Another direction of techniques on class imbalance problem focuses on improving the existing algorithm. Wu et al. [10] report that they propose to enlarge the resolution around the decision boundary by revising kernel functions. Moreover, Veropoulos et al. [11] proposed a method which uses pre-specified penalty constants on Lagrange multipliers for different classes to indemnify for the skewness of the decision boundary. Using an ensemble of classifiers to boost classification performance has also been reported to be effective in the context of imbalanced data. The work in this direction includes that Yan et al. [12] propose an ensemble-based approach that applies SVM to address the issue of predicting rare classes in scene classification, and that Chen et al. [13] use random forest to unite the results of decision trees induced from bootstrapping the training data.

However, all these methods have their defects which deal with the class imbalance problem in a single direction. SMOTE brings more computational cost to the system for preprocessing, on the other the ensemble of classifiers increases number of training data making the classifiers training very costly. To address this

concern, this paper proposes an approach to control the scale of the ensemble of classifiers by k-means algorithm and resampling method, which boost prediction accuracy in a reasonable period of time.

72.3 SVM Ensembles

To address the defect that the inhomogeneity of spatial distribution of positive instances in class imbalance problem, We first take an ensemble of SVM to boost the performance by k-means algorithm clustering negative class instances into K subclasses which have a similar attributes in space. Moreover, we combine all the positive class instances with each negative subclass to be an individual subset, and then train SVM independently with each subset and combine the result in certain strategies.

72.3.1 Dataset Clustering

K-means clustering, as a method of cluster analysis, aims to partition n observations into K clusters in which each observation belongs to the cluster with the nearest mean. Given a set of observations$(x_1, x_2, \ldots, x_n)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k sets$(k \leq n)S = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{X_j \in S_i} \left\| X_j - \mu_i \right\|^2 \tag{72.1}$$

where μ_i is the mean of points in S_i.

After k-means clustering, negative class instances is divided into K subclasses as shown in Fig. 72.1. Each of negative subclass is more uniform in spatial distribution as well as decreases the imbalance ratio. However, K value, which we focus on, is difficult to choose since it depends much on negative class. Extremely small K value leads to bad clustering effectiveness while it causes the problem of long training time and low training efficiency as a result of generating more subclasses that K value is too large.

72.3.2 Combination of Multiple SVM

As illustrated in Fig. 72.2, we cluster negative class instances into K subclass, where K is depending on the number of positive examples. Then, each negative subclass is combined with the positive instances to form a train set to train an

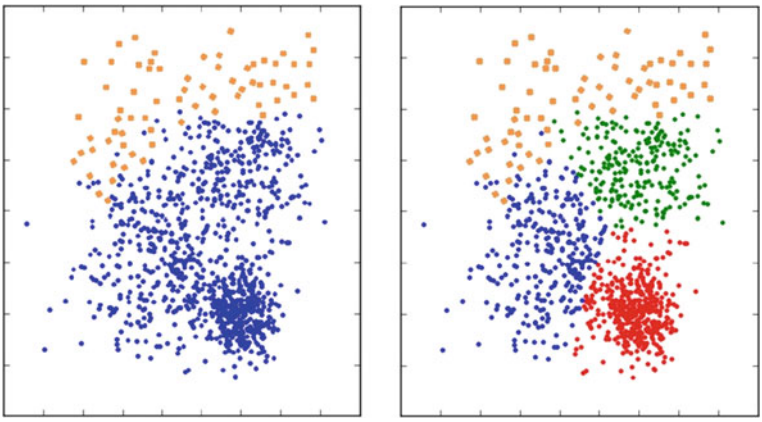


Fig. 72.1 An example of k-means clustering on negative class instance

SVM. After each classifier is trained independently, we come to aggregate their results in an appropriate combination approach. In this condition, we use direct combination strategies since continuous-valued outputs such as posteriori probabilities are available.

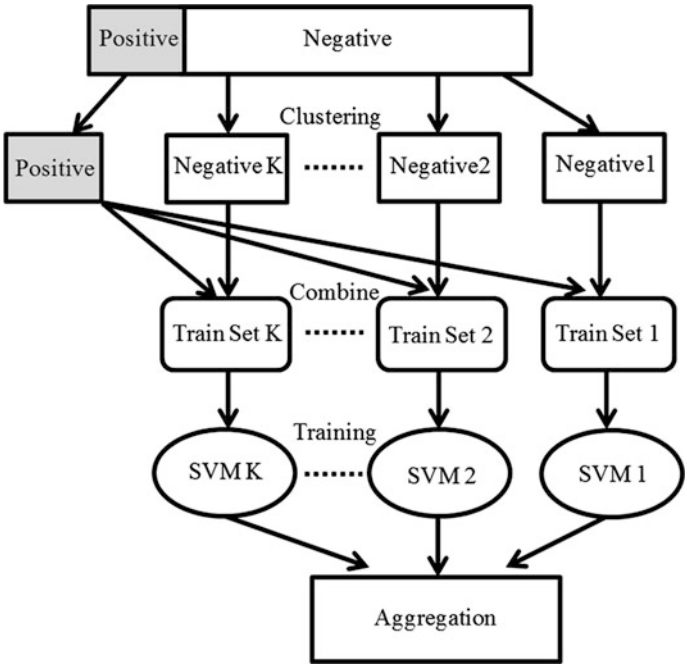


Fig. 72.2 Architecture of the SVM ensembles

72.4 Resampling Methods

SVM ensembles are effective to address the class imbalance problem. Nevertheless, the situation often happens that imbalance ratio is extremely high, 1:1000 for instance. In order to re-balance datasets which means the number of negative instances need as little as possible, we have to set a large K value to generate many SVM classifiers. The problem of long training time and low training efficiency will get worse if we use ensemble techniques only. Thus, resampling Methods which are independent of the classifier used are more versatile. In this section, we combine both undersampling and oversampling to balance the data.

The undersampling approach, which has been reported to outperform oversampling approach in previous literatures, selects a subset of the examples which represents the initial problem better, and avoids the bias to the negative class by removing redundant examples. It also has the advantage of creating a reduced set of examples to the induction process which less costly. However, undersampling throws away potentially useful information in the majority class, thus it could make the decision boundary trembling dramatically.

Instead of avoiding bias examples of the negative class, in oversampling approach SMOTE for instance, the positive class is oversampled by taking each positive class sample and introducing synthetic examples along the line segments joining any of the K positive class nearest neighbors. Depending upon the amount of oversampling required, neighbors from the k-nearest neighbors are randomly chosen. Figure 72.3 shows the performance of the adopted SMOTE. The left figure is the distribution of the original data while the right figure is the distribution after adding double synthetic minority class samples through SMOTE. From Fig. 72.3, the synthetic instances can basically keep the distribution of original samples, and cause the classifier to create larger and less specific decision regions.

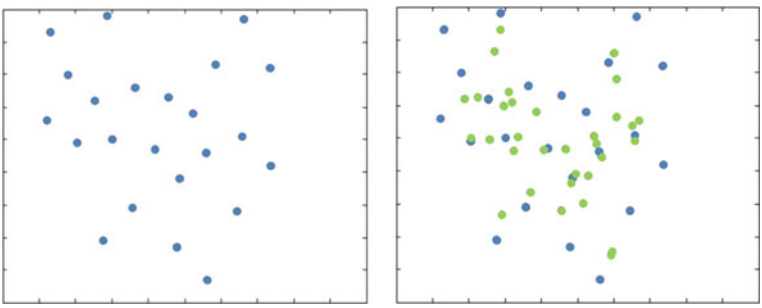


Fig. 72.3 The performance of SOMTE algorithm

72.5 Experimental Evaluations

In this section, we first describe the experimental data used in our test bed, and then introduce the evaluation standard. Finally, we report the experimental results that discuss the influence of K value on the performance and compare our combination approach with other methods.

72.5.1 Experimental Data

Five datasets are used to test the approach we proposed. All of these datasets are from the UCI (University of California Irvine) Machine Learning Repository. The five UCI datasets are Abalone, Hypothyroid, Prima, Vehicle and Glass. Information about these datasets is shown in Table 72.1. In order to calculate conveniently, both Vehicle and Glass which are more class problem transform into the standard two-class classification problem, that is, one class is considered to be positive and all the other classes are considered to be negative.

72.5.2 Evaluation Standard

The evaluation standard used in our experiments is based on the Confusion Matrix as illustrated in Table 72.2 for the two-class classification problem.

In our experiments, the performance measures are defined as follows:

$$g = \sqrt{acc^+ \times acc^-} \tag{72.2}$$

where $acc^+ = \frac{TP}{TP+FN}$ is the accuracy in positive instances and $acc^- = \frac{TN}{FP+TN}$ is the accuracy on the negative instances. And the geometric mean of the g reaches high value only if both acc^+ and acc^- are high and in equilibrium. The accuracy on positive instances can be increased at the cost of accuracy on negative instances.

Table 72.1 Dataset information

Dataset	Feature	Positive	Negative	Imbalance ratio
Abalone	9	61	5745	94.18
Hypothyroid	29	104	3513	33.78
Prima	8	268	500	1.87
Vehicle	18	176	2532	14.39
Glass	9	29	223	7.69

Table 72.2 Confusion matrix

	Predicted positive	Predicted negative
Actual positive	*TP* (true positive)	*FN* (false negative)
Actual negative	*FP* (false positive)	*TN* (true negative)

72.5.3 Experimental Results

As mentioned in Sect. 72.3.4, the selection of K may impact on the prediction accuracy. To make a better understanding, we make the experiment of the influence of K value on prediction accuracy. We are told form Fig. 72.4 that both five datasets achieve their highest accuracy with different K value since they have different imbalance ratio. By further observation, we can realize that the higher imbalance ratio the dataset has, the bigger K value need to be chosen.

Table 72.3 shows the performance for each method in G-mean, where SVM represents the original SVM method, Resampling Method denotes oversampling and undersampling the dataset and then training with original SVM, SVM Ensembles represents using ensemble of SVM based on k-means algorithm. SVM Ensembles with Resampling denotes the new approach we proposed. From the results we can see that SVM Ensembles with Resampling achieves the best results on all datasets except the Prima dataset for which SVM is the best since the imbalance ratio of Prima dataset is only 1.87. Thus, SVM Ensembles with Resampling may cause unsatisfactory results in balance datasets.

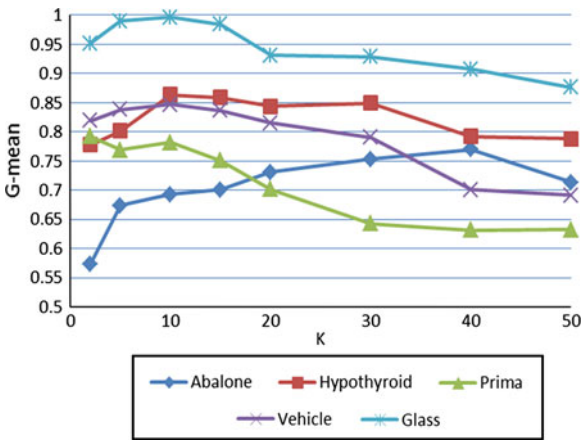


Fig. 72.4 G-mean with reference to different K values

Table 72.3 Performance in g-mean

Dataset	SVM	Resampling method	SVM ensembles	SVM ensembles with resampling
Abalone	0.513	0.559	0.744	**0.772**
Hypothyroid	0.729	0.783	0.845	**0.864**
Prima	**0.812**	0.801	0.784	0.792
Vehicle	0.708	0.790	0.845	**0.847**
Glass	0.876	0.937	0.965	**0.996**

72.6 Conclusion

In this work, researchers present a new approach which combines SVM ensembles and resampling method to address the class imbalance problem. They first use an ensemble of SVM based on k-means algorithm to boost the performance. In order to make the K value in k-means algorithm appropriate, which means more robust results can be obtained and reducing the randomness induced by classifier, researchers further propose to integrate the two types of sampling strategies by oversampling the positive class to a middle extent and undersampling the majority class to the similar size. The experimental results indicate that the proposed approach can receive better performance than the original approaches.

Acknowledgments Supported by Beijing Key Laboratory of Network Systems and Network Culture(Beijing University of Posts and Telecommunications) and the Innovative Experiment Plan for College Students of China University of Geosciences, Beijing.

References

1. Japkowicz, N.: The class imbalance problem: significance and strategies. In: Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI) (2000)
2. Dumais, S., Platt, J., Heckerman, D.: Inductive learning algorithms and representations for text categorization. In: Proceedings of International Conference on Information and Knowledge Management (CIKM) (1998)
3. Chan, P.K., Stolfo, S.J.: Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (1998)
4. Kubat, M., Holte, R.C., Matwin, S.: Machine learning for the detection of oil spills in satellite radar images. Mach. Learn. **30**(2–3), 195–215 (1998)
5. Fawcett, T., Provost, F.J.: Adaptive fraud detection. Data Mining and Knowledge Discovery, pp. 291–316 (1997)
6. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (1995)
7. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs (1988)
8. Japkowicz, N.: A novelty detection approach to classification. In: Proceedings of European Conference on Principles and Practice of Knowledge Discovery in Databases (2006)

9. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority oversampling technique. J. Artif. Intell. Res **16**, 321–357 (2002)
10. Wu, G., Chang, E.Y.: Aligning boundary in kernel space for learning imbalanced dataset. In: IEEE International Conference on Data Mining (ICDM), pp. 265–272 (2004)
11. Veropoulos, K., Campbell, C., Cristianini, N.: Controlling the sensitivity of support vector machines. In: International Joint Conference on Artificial Intelligence (IJCAI99) (1999)
12. Yan, R., Liu, Y., Jin, R., et al.: On predicting rare classes with SVM ensembles in scene classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 21–24. Hong Kong (2003)
13. Chen, C., Liaw, A., Breiman, L.: Using random forest to learn imbalanced data. In: Technical Report 666, Statistics Department, University of California at Berkeley (2004)

Chapter 73
Join Optimization for Large-Scale Data Analysis in MapReduce

Li Zhang, Shicheng Xu and Chengbao Peng

Abstract As the coming of the big data age, there is a new hot spot on how to handle and process huge amounts of data. The MapReduce parallel computing framework is increasingly being used in large-scale data analysis. Although there have been many studies about the join operation in the traditional relational database, join algorithms in MapReduce are inefficient. In this paper, we describe a number of well-known join algorithms in MapReduce, and present an experimental comparison of these join algorithms based on Hadoop cluster. An optimization algorithm for map side chain is proposed.

Keywords MapReduce · Join algorithm · Map side joins

73.1 Introduction

In 2004, a parallel computing framework called MapReduce was first developed by Google engineers-Jeffrey Dean and Sanjay Ghemawat [1]. It was designed for processing huge amount of web access logs in Google. The MapReduce program enables parallelization on a cluster of commodity machines. It provides a simple interface to achieve automatic parallelization and large-scale distributed computing, while hiding the complex details of fault tolerance, load balancing and data distribution in parallelization.

Based on the idea of MapReduce, the Apache Software Foundation has developed Hadoop [2], an open-source version of MapReduce framework. Hadoop

L. Zhang (✉) · S. Xu · C. Peng
Neusoft Corporation, Shenyang, China
e-mail: neu-zhangli@neusoft.com

C. Peng
Northeastern University, Shenyang, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_73,

has attracted a number of large internet companies like Yahoo, Facebook, Linkedln, etc. In the case of Facebook, it is used to store the huge amount of web logs (usually PB) for data mining and machine learning. Compared with the traditional relation database, Hadoop only provides developers with the underlying programming interface, it is a time-consuming and skilled job to write map and reduce functions with low-level programming language. So Facebook developed a data warehouse system, namely Hive [3, 4], on top of Hadoop. The hive can provide similar SQL query, so it is very convenient to analyze huge amounts of data stored in Hadoop for the people who are familiar with the SQL. Now Hive is also the project of the Apache Software Foundation.

Because of its easy to use, installation and implementation, MapReduce is being widely used in various types of tasks, one of which is the massive data analysis. The join query is the most critical operation in the large-scale data analysis. Compared to other kinds of operation, it is generally the most common and time-consuming, and has a great influence on the overall performance. It is quite simple to join two datasets in MapReduce, but for the majority of join algorithms, which first read both tables from disk and transfer the intermediate result data to reducers via the network, the performance of join is influenced by the network speed. When the two data sets are very large, the network transmission time becomes the performance bottleneck, which reduces the computational resource availability.

For a large-scale data analysis application, the join operation is an essential function. Now many applications on top of Hadoop MapReduce framework, such as Hive and Pig [5], have implemented join operations, which can be completed by the simple SQL statement or data manipulation script [6]. However, neither of these tools has solved the joint problem efficiently. In this paper, we introduce the existing common implementations of join algorithms and propose an optimized solution for joining datasets in MapReduce. The optimized algorithm is based on the idea of reducing the time of network communication.

73.2 MapReduce Overview

The simplicity is one of the most attractive characteristics in the MapReduce programming model. A MapReduce program is composed of two computations, the map and the reduce computations, which are executed on every node. The map function reads a series of records from an input file, processes these records, and produces intermediate results. The final outputs are a set of intermediate records, which are composed of key/value pairs. The MapReduce framework will group the intermediate key/value pairs by the key so that the key/value pairs with the same key will be presented to the same reduce function. The reduce function is responsible for reading the local output of the map function, combining and processing them to produce the final result which is stored in the Hadoop Distributed File System (HDFS) that provides high-throughput access to application

data. In detail, a file used as input to a MapReduce job is split into a number of blocks distributed across the nodes of the cluster. Each map function reads the input one by one and converts it into a key/value pair. The intermediate output pairs are grouped and sorted by their keys. During the reduce phase, the reducers retrieve the intermediate results from the mappers, combine these values and return a new key/value pair. The data flow as following:

$$(k, v) \rightarrow map \rightarrow (k1, v1), (k1, v2), (k2, v3) \rightarrow partiton \,\&\, sort \rightarrow (k1, list(v1, v2)), (k2, v3) \rightarrow reduce \rightarrow (k3, v4)$$

73.3 Join Algorithms

In this section, we present and analyze the well-known join implementations in the MapReduce framework. There are two main join algorithms: the map side chain and the reduce side join. Their name denotes the phase during which the actual join operation is performed. Additionally, we introduce an optimized version of the simple map side join which engages the Distributed Cache on the map phase [7].

73.3.1 Reduce Side Join

The reduce side join is also known as the reparation joins and similar to the repartition sort-merge join in the traditional relation database. In this algorithm, the map phase only reorganizes the datasets by the join key, the actual join takes place during the reduction phase. It is most generally join approach in the MapReduce, as it comes with no restrictions regarding the input data sets.

In this algorithm, the mapper reads one tuple at a time from both datasets, tags each tuple with its sources, and reorganizes them in terms of the join key. The output key/value pairs as following: <the join key+tag, tuple>, Tagging is necessary since it is used to differentiate the tuple from which dataset. The map outputs with the same key are assigned to the same reducer. During the reduction phase, the reducer identifies each tuple's parent dataset by the tag value, buffers the first dataset's tuples in order to join them with the tuples of the second dataset, and then writes the final result into HDFS.

From the above, it is clear that the reduce side chain is not so efficient because of the shuffle phase. Actual join doesn't take place until the reduce phase begins. In other words, it is only during the reduce phase that disjoint tuples are discarded so that we will shuffle both datasets over the network. And in the most case, we will get rid of most of these data during the reduce function. This point reduces

remarkably the efficiency of the reduce side join. Therefore, it is expected to perform the entire join operation during the map phase.

73.3.2 Map Side Join

The reduce side join in the last section seems to be the most straightforward approach, but it can also be quite inefficient, as the shuffle phase is a very time consuming. Another join algorithm is the map side join, which implements the join at the map phase and eliminates the reduce phase. This algorithm has some constraints on the input datasets. All datasets must be partitioned and sorted in the same way. Each dataset is divided into the same number of partitions and sorted by the join key. All tuples with the same key will be placed in the same partition.

This algorithm is not only applied to the two-way joins, but also multi-way joins. As long as the jobs have the same number of reduce functions, same keys, and the output file can't be cut (such as smaller than HDFS block size or gzip compression), the map side join can be used to join these jobs' outputs. The CompositeInputFormat class in org.apache.mapred. Join package can be used to perform the map side join, whose input dataset and join type (inner join or outer join) can be configured by an expression.

In some applications, one dataset is small enough and can be directly put into memory. For example, it is common that a small user database may need joining a large user log in order to analyze user behavior. In this situation, the map side join is a better choice. First, the smaller dataset will be loaded into memory, and the mapper reads each tuple from the larger data set, at the same time probing the smaller dataset in memory in order to joining. But if the small dataset is not small enough, there is out-of-memory exception. This algorithm can be optimized by making use of the hash algorithm, with which the small table can be load as hash table to reduce the space and accelerate the speed of probing.

All in all, the map side join is quite efficient because of the elimination of the reduce phase. Each map task reads the data directly from the HDFS and no reduce is needed. The map side join is the best choice when there are a smaller dataset fit into memory and a lager dataset. However it is clear that the map side join lacks the generality of the reduce side join, since it needs the input data to be partitioned and sorted in the same way.

73.3.3 Distributed Cache

As the map side join reduces the time of transferring the data to the reducer and sorting the data, it is more efficient than the reduce side join. There are many researchers who study and improve the map side join about mass data processing.

MapReduce Distributed Cache is used in combination with the map side join for it to work with more cases.

The merit of Distributed Cache is eliminating the network communication. Before launching the job, the smaller dataset is assigned to Distributed Cache. The data is copied to each node that has map tasks so that it can be accessed through local file system instead of HDFS. During the map phase, the mapper only needs reading data from the local file system. Compared to reading data from HDFS (similar another node), this can save the network transmission time and improve query performance. In this case, all the map tasks on one node will share one data copy. With Distributed Cache, the application scope of the map side join will be extended.

73.4 Improving Join

Since the simple map join can only deal with a huge dataset and a small dataset. In the worst case, two datasets are huge and neither can be placed in the memory. In order to solve this problem, MapReduce Distributed Cache is proposed to copy the smaller dataset to each node. But the Distributed Cache is a double-edged sword, when the dataset is gradually increasing, it will become the performance bottleneck. The mappers will wait for the datasets distributed by Distributed Cache. In the worst case, when there are many nodes are waiting, the performance using the Distributed Cache will be worse than the reduce side join.

The idea of a two-stage map join is intuitive: when the duplicate dataset becomes large, the bottleneck of Distributed Cache is that many nodes are waiting for the data. It may make sense to read as much data into memory and perform the map join on this part of the data; the remaining data is compressed and copied by Distributed Cache. As the duplicate data is processed in two stages, the algorithm is called two-stage map join. The map in the first stage will process the smaller dataset, filter the tuples, and produce two types of output. The first type of output will be read by the map in the second stage, and then fit into memory in order to perform the map join as soon as possible. The second type of output will be organized in the form of text file and compressed. When the first type of output is processed completely, we assign the second type of output to Distributed Cache, so that it is copied to each node that running tasks.

73.5 Experimental Evaluations

All experiments run on the Hadoop cluster which consists of 100 nodes, each node is equipped with two quad-core CPUs, 1G RAM and 1T storage. TPC-H data are our source data. The Map task input split size is the default 128 MB. The number

of reduce tasks is 20 for the reduce side join. The size of one dataset is fixed at 20 GB, while the scale of other dataset goes from 20 M to 5 G.

73.5.1 Comparison Between Map Side Join and Reduce Side Join

Figure 73.1 shows the performance of Map Side Join (MSJ) and Reduce Side Join (RSJ). As shown in the Fig. 73.1, the map side join performs better than the reduce join. This result is due to the following reasons: 1) map side join gives up the reduce phase, which saves the time of data transferring via network and sorting at shuffle phase. When a job launches too many map tasks, the joining performance will degrade due to the massive network data transmission. As we can see, the limitation of map side join is the out-of-memory exception when the "replicated dataset" is becoming too large.

73.5.2 Performance of Two-stage Map Join

When the "replicated dataset" is less than 100 M, the two-stage map join (TMJ) algorithm needs only one stage, and is same as the map side join. In this case, the performance of two-stage map join is consistent with that of the map side join. When the smaller dataset increases gradually, TMJ splits the table into two parts, launches two joins to perform join operation. As a result of reducing the waiting time for the data copy, the map task can work as early as possible. From the performance perspective, the performance of TMJ is better than that of Distributed Cache join (DCJ) and map side join. As shown in the Fig. 73.1, TMJ also extends the simple map side join algorithm for it to work with more cases.

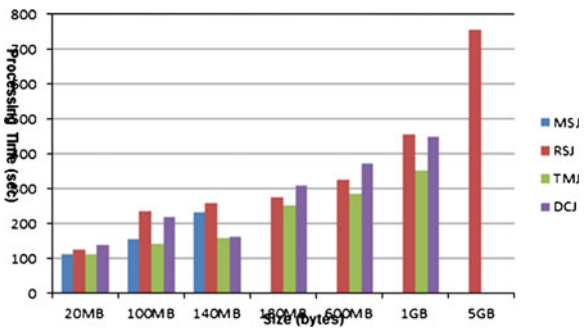


Fig. 73.1 Query test

73.6 Conclusion

For the large-scale data analysis in MapReduce, the join query is one of the most important queries. Based on the analysis of the existing join algorithms in MapReduce, we propose a new idea about join query in MapReduce framework, and present a comprehensive experiment on the Hadoop cluster with 100 nodes in order to evaluate the join algorithm in the context of MapReduce. The respective advantages and disadvantages of all join algorithms are shown through the experiment. The insights obtained from the study can help an optimizer to choose the appropriate join algorithms according to the characteristics of datasets.

References

1. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. Commun. ACM **51**(1),107–113 (2008)
2. Hadoop.: http://hadoop.apache.org
3. Thusoo, A., Sarma, J.S., Jain, N., Shao, Z., Chakka, P., Anthony, S., Liu, H., Wycko, P., Murthy, R.: Hive: a warehousing solution over a map-reduce framework. Proc. VLDB Endow **2**(2), 1626–1629 (2009)
4. Thusoo, A., Murthy, R., Sarma, J.S., Shao, Z., Jain, N., Chakka, P., Anthony, S., Liu, H., Zhang, N.: Hive—a petabyte scale data warehousing using hadoop, ICDE (2010)
5. Gates, A.F., Natkovich, O., Chopra, S., Kamath, P., Narayanamurthy, S.M., Olston, C., Reed, B., Srinivasan, S., Srivastava, U.: Building a high-level dataflow system on top of map-reduce: the pig experience. Proc. VLDB Endow **2**(2), 1414–1425 (2009)
6. http://wiki.apache.org/hadoop/hive/languagemanual/joins
7. Distributedcache.: http://hadoop.apache.org/common/docs/current/mapredtutorial.html#DistributedCache

Chapter 74
Key Technologies of Data Preparation for Simulation Systems

Xiangzhong Xu and Jiandong Yang

Abstract The speed and quality of data preparation become one of the bottlenecks of the execution of large-scale simulation systems, which has not gained enough attention so far. The paper discusses the intrinsic characteristics of three types of simulation data, that is, relational data, hierarchical data and spatial data. It expounds the following three key technologies of data preparation for simulation systems. For efficient management of hierarchical data, XML is proposed and the XML-based component is developed for reuse. For seamless management of heterogeneous data, the three-tied architecture is put forward. To improve data quality, quality of simulation data is divided into applicability, effectiveness and validity, and the theoretical framework for quality assurance of data preparation for simulation systems is given. Research results are of great significance to improve the efficiency of data preparation for simulation systems.

Keywords Simulation data · Data preparation · Data management · Data quality assurance

74.1 Introduction

Nowadays, following theoretical research and experimental research, simulation has gradually become the third important means to cognize the world and to reconstruct the world [1]. Platform-independent-applications, data-independent-models are commonly used to improve the agility and applicability of simulation

X. Xu (✉)
Academy of Armored Forces Engineering, Beijing, China
e-mail: xuxz02@21cn.com

J. Yang
China Satellite Maritime Tracking and Control Center, Jiangyin, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_74,

systems. As the complexity of simulation objects and the magnitude of simulation systems increasingly grow, it is more difficult to develop simulation systems and to prepare data for simulation systems. In fact, the speed and quality of data preparation become one of the bottlenecks for simulation execution which has not gained enough attention so far.

Simulation data can be divided into three groups, that is, relational data, hierarchical data and spatial data. Quality of simulation data is determined by the following factors, such as applicability, effectiveness and validity. In data preparation for simulation systems, there are several thorny issues, such as efficient management of hierarchical data, seamless management of heterogeneous data, and quality assurance of data preparation.

The eXtensible Markup Language (XML) is proposed for efficient management of hierarchical data and the XML-based component is developed for reuse. The three-tied architecture is put forward for seamless management of heterogeneous data. To improve data quality, quality of simulation data is divided into applicability, effectiveness and validity, and the theoretical framework for quality assurance of data preparation for simulation systems is given.

The rest of this paper is organized as follows: Section 74.2 gives a brief analysis of simulation data according to their intrinsic characteristics. Section 74.3 discusses efficient management of hierarchical data based on XML. In Sect. 74.4, seamless management of simulation data based on the Commercially-Off-The-Shelf components are proposed. Section 74.5 gives the framework for quality assurance of data preparation for simulation systems. Section 74.6 concludes the paper.

74.2 Brief Analysis of Simulation Data

Models and data are two basic components of simulation systems, and the initialization and execution of simulation models depends heftily on data. There are stochastic data and deterministic data. Stochastic data are usually produced by pseudorandom number generators and used for Monte Carlo simulation [2]. According to their intrinsic characteristics, deterministic data fall into the following three categories.

74.2.1 Relational Data

Relational data mean those regular two-dimensional structured data, such as the attribute of persons, organizations, and other entities, and the performance data of weapons. It is suitable to apply relational model to manage relational data, and the management technologies of relational data are very mature, viz. relational database management systems (RDBMS). RDBMS can provide users with efficiency,

data integrity, consistency and reliability due to their highly tuned index, transaction processing, concurrent access control, recovery, trigger and other powerful mechanisms. Therefore, RDBMS are playing the leading role for the management of large amounts of regularly structured data. In fact, modern large simulation systems have been constructed upon the basis of RDBMS. However, in such situations, data must be tabular and conform to specific schema, which promotes integrity but discourages the management of hierarchical data.

74.2.2 Hierarchical Data

Hierarchical data mean those data characteristic of hierarchy (such as parent/child, ancestor/descendant), sequence of sibling nodes, or switches. There are many hierarchical data in the real world, such as organization, unit, form, weapon system, taxonomy and pedigree, just name a few. They are most naturally modeled as a hierarchy. In fact, the management of hierarchical data is a common issue facing data preparation for simulation systems, which is therefore worthy of good consideration.

74.2.3 Spatial Data

Spatial data mean those data characteristic of geographic distribution, such as battlefield environments, and deploy, maneuver of entities. Spatial data are common in simulation systems, especially in military simulation systems. Spatial data differ vastly from relational data, so it is inefficient for RDBMS to manage spatial data, unless they are extended for spatial data.

74.3 Efficient Management of Hierarchical Data

As mentioned above, hierarchical data are very common in simulation systems. RDBMS are suitable to the management of relational data, but not hierarchical data. By contraries, the eXtensible Markup Language (XML) can play a great role in managing hierarchical data.

74.3.1 Weak Points of Relational Data Model

RDBMS contain only a fraction of the world's data for several good reasons. Their theoretical underpin is relational theory, and their data model is relational model.

In user's opinion, the relational model is a regular table consisting of rows and columns in terms of logical structure. With the rapid development and evolution of semi-structured data and non-structured data, RDBMS show the following shortcomings [3]:

- The real world has to be hardly mapped to a collection of tables; therefore, many semantics of complex objects, e.g. aggregation and specification, have to be discarded.
- In order to designate the sequence of sibling nodes, extra fields must be attached to relational tables to describe semantics data possess in the real world (For the record set, the sequence is meaningless, which is mandated by the relational theory. Thus, queries can be greatly optimized through highly tuned index). Moreover, it is difficult to maintain the semantics dynamically.
- Similarly, for the description of hierarchy (such as parent/child, ancestor/descendant) of data in the real world, extra fields must also be added to relational tables. In fact, hierarchy can be treated as a special form of semantics.
- The relational model demands that relations should be normalized. In order to enhance the normalization level of tables to reduce redundancy and avoid modification exception (including add, update and delete), it's necessary to split relations, i.e. normalization design, in design-time and perform join operation when needed in run-time. Thus, it may perplex system design and afflict system run-time performance for the very reason that joins is a most cost operation in RDBMS. A balance should therefore be carefully made between design-time performance and run-time performance.

74.3.2 Advantages of XML Data Model

Applying XML to manage hierarchical data has the following several distinctive advantages, such as building and maintaining the hierarchy readily, preserving the sequence easily and incorporating XML into tree control conveniently provided by develop platforms.

74.3.2.1 Characteristics of XML Data Model

XML is an international standard stipulated by W3C, which is designed as lingua franca to facilitate information interchange between distinct programs and between the user and the program. It has several outstanding characteristics, for example, expansibility, simplicity, and self-description.

From the viewpoint of data modeling, XML provides two different means, that is, DTD and XML Schema [4]. They can stipulate normal forms for a group of XML documents with the same logical structure and help developers to

materialize their ideas. It is through these two modeling means that XML promises to create more smart documents (e.g., proving error-looking up function to certain extent) and application programs can abstract needed information from texts more easily and efficiently and present themselves as people wish.

74.3.2.2 Characteristics of XML Data Model

XML data model is tree-formed in itself and, once a well-formed and validated XML document has been parsed successfully, a tree will be built in memory. To manage the semantics and hierarchy is therefore an easy job for XML data model. What is more, the dynamic maintenance of them is also simple [5].

Additionally, an XML parser is a component, providing standard interfaces. Developers thus can avoid develop and distribute interface program repeatedly, and reduce the cost of test and maintenance.

74.3.2.3 Combination of XML and Tree Control

The management of hierarchical data is a routine in data preparation of simulation systems. It's necessary to exploit the advantages of software reuse to improve efficiency and quality of software process and avoid introduction of new errors. XML's strong base of freeware and commercial tools afford flexibility at greatly reduced development costs. The characteristics and functions of the component CXmlOleTree are as following (Fig. 74.1): building the tree easily, meanwhile preserving the hierarchy and sequence of data, maintaining the tree graphically, supporting ole, encoding the node flexibly which is the key to the join of hierarchical data and relational tables, applying user custom redraw technology, which enables powerful representation.

The main member attributes of CXmlOleTree include IXMLDOMDocumentPtr domDocPtr, IXMLDOMNodePtr domNodePtr,IXMLDOMElementPtr domElementPtr, CTreeDropTarget m_CTreeDropTarget,_ConnectionPtr m_pADOConn, RecordsetPtr m_pADOSet, CBitmap m_bitmap,CImageList * m_pImageList.

74.4 Seamless Management of Heterogeneous Data

It is necessary to perform seamless management of hierarchical data, relational data and spatial data during data preparation of simulation systems.

The seamless management of heterogeneous data is constructed upon three-tied architecture (Fig. 74.2). From bottom to up, they are the storage layer, the middle layer and the representation layer respectively. The key to the solution is that XML is applied to the management of small-scale hierarchical data, that RDBMS are applied to the management of large-scale relational data and spatial data, and that

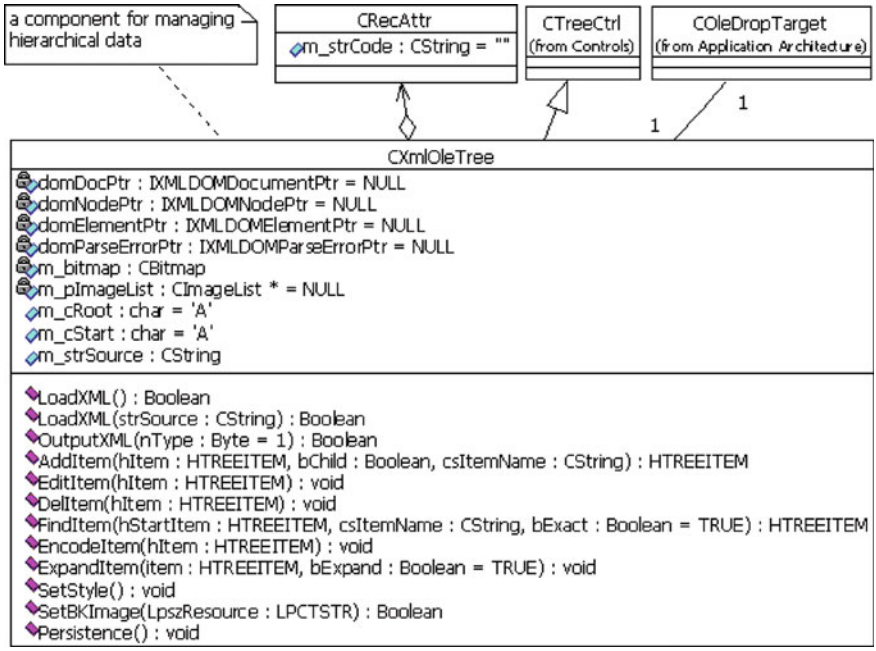


Fig. 74.1 UML model of the XML-based component CXmlOleTree

these data are joined by the node code of tree items. In representation layer, hierarchical data, relational data and spatial data are mainly represented as trees, lists and maps respectively. This goal is achieved through the components including CXmlOleTree, ClistCtrl and MapObjects.

The solution is also component-based. The third-party components include MSXML3 (the parser for XML documents), ADO (an application-level programming interface to data and information), Map Objects (a develop platform for geographic information systems), and MFC (an application framework for

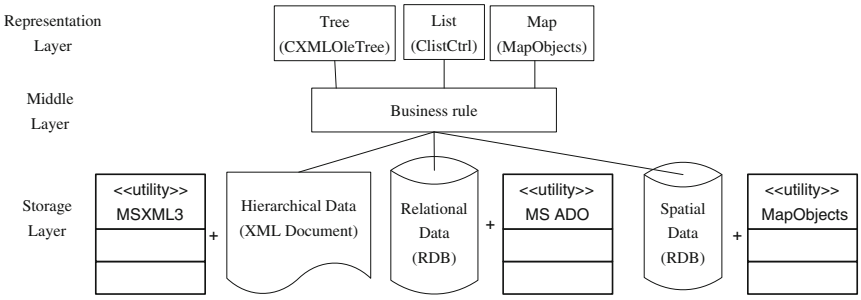


Fig. 74.2 Seamless management of heterogeneous data

programming) Library. The middle layer is business rule for data preparation, which is related to specific application domain and open to developers.

74.5 Theoretical Framework for Quality Assurance of Data Preparation for Simulation Systems

The quality of simulation data has become the bottleneck of simulation systems which can be divided into three categories, namely applicability, effectiveness and validity [6]. Applicability means data provided by simulation systems are in accordance with data needed by simulation users for their decision-making. Effectiveness means the update frequencies of simulation data are in accordance with simulation users' needs. From this viewpoint, simulation data can be divided into static data and dynamic data. They have different requirement. The former focuses on stability, whereas the latter focuses on timeliness. Validity means there are not any incorrectness, inconsistency, incompleteness, or repetition in simulation data. In fact, poor quality data have yielded severe impacts on the effectiveness of information systems, including simulation systems. They should be detected and amended before the simulation execution.

It is urgent to find out the critical influencing factors of data quality, put out methods for improving the quality of simulation data, and set up the theoretical framework for quality assurance of data preparation for simulation systems under the guide of the data quality assurance theory from three aspects, viz. theory, method and technology (Fig. 74.3).

The theoretical framework for quality assurance of data preparation for simulation systems consists of three parts, that is, data usableness theory and method based on metadata, data updating theory and method based on real-time monitoring, and poor quality data detection theory and methods based on knowledge tree. The relevant technologies include data modeling based on metadata, data monitoring based on regression prediction, real-time data monitoring based on rules, data detection based on knowledge tree, reasoning of business rule, and so on.

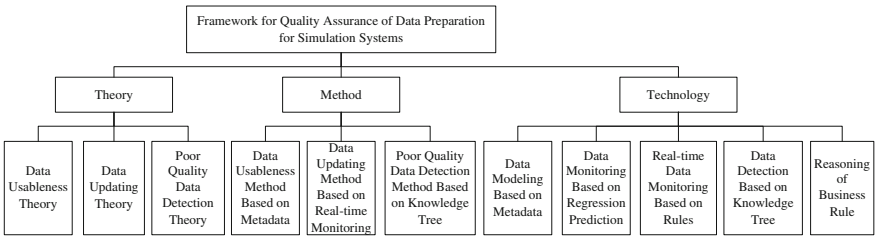


Fig. 74.3 Seamless management of heterogeneous data

74.6 Conclusion

Based on these research results, the data preparation tool for the Corps equipment support simulation and evaluation system is developed. Firstly, batch input through worksheets and individual modify through user interface are combined to speed up data preparation. Proper ways are used to prepare specific type of data. For example, the framework of unit and form are suitable to be prepared through batch input. On the other hand, logistic relationship, maneuver route are suitable to visually input through user interface. Secondly, data visualization and check rule are combined for poor quality data detection, so as to avoid the introduction the error of initial conditions into the simulation executions.

In further work, the automatic data generation based on certain rules are considered to be combined with manual modify to further improve the response and quality of data preparation for simulation systems.

References

1. President's Information Technology Advisory Committee.: Computational Science Ensuring America's Competitiveness. USA National Coordination Office for Information Technology Research & Development, pp. 10–13 (2005)
2. Xu, X.Z, Pan, L.J.: Study on high performance random number generators for combat simulation systems. J. Acad. Armored Forces Eng. **23**, 67–70 (2009)
3. Wang, S., Chen, H.: A Course in Database System Principle. Tsinghua University Publishing Company, Beijing, China (2001)
4. St. Laurent, S.: XML A Primer, 2nd edn. Electronic Industry Publishing Company, Beijing, China (2000)
5. Xu, X.Z, Wang, J.Y, Shao, L.S.: Towards better management of hierarchical data. In: Proceedings of the International Conference on System Simulation and Scientific Computing. Beijing, China (2005)
6. Zhang, Z.B.: Research on data quality control theory and methods of the equipment command information system. A Dissertation Submitted for the Doctoral Degree. Academy of Armored Forces Engineering. Beijing, China, pp. 26–27 (2011)

Chapter 75
Suspend-to-PCM: A New Power-Aware Strategy for Operating System's Rapid Suspend and Resume

Chenyang Zi, Chao Zhang, Qian Lin, Zhengwei Qi and Shang Gao

Abstract Modern data centers provide good performance and many kinds of services accompanying considerable power consumption. Reducing power consumption becomes essential for decreasing the operating costs. Unlike conventional ways, authors propose the novel Suspend-to-PCM hibernation strategy to provide rapid suspending and resuming of the operating system (OS). Characterized by its low access latency and low energy consumption, Phase change memory (PCM) is a kind of non-volatile flash media. Strategy in the paper is to suspend OS from memory to PCM rather than disk, and vice versa for the resuming. Since PCM owns much better property of access speed and power utilization than traditional storage media, the Suspend-to-PCM strategy is able to achieve improved performance of system suspending and low power consumption.

Keywords Rapid suspending and resuming · PCM · Power saving · Data center · Performance

C. Zi (✉) · C. Zhang · Q. Lin · Z. Qi · S. Gao
Shanghai Key Laboratory of Scalable Computing and Systems,
Shanghai Jiao Tong University, Shanghai, China
e-mail: zichenyang001@163.com

Q. Lin
e-mail: linqian.qian@hotmail.com

Z. Qi
e-mail: qizhwei@sjtu.edu.cn

S. Gao
e-mail: chillygs@sina.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_75,

75.1 Introduction

With the rapid growth of applications spurred by the Internet, power consumption issue of enterprise servers is increasingly critical in the design and operation in contemporary data centers. It is reported [1] estimates that the servers and data centers in the United States consumed about 61 TWh. This energy would cost $6B annually at a common price of $100 per MWh and generating this electricity would release about 36 M tons of new Carbon Dioxide annually [2].

The most common way to reduce power consumption is to hibernate some computers when they are not involved in computing and wake them up on demand. Generally, power consumption of a normally running computer is about tens of watts. When OS goes into hibernation, the power consumption will be less than running normally and resume overhead will be less than shutting down. Academia and industry are making efforts to provide faster and more convenient wake-up service, such as Advanced Configuration and Power Interface (ACPI) which is the interface of all the power management for applications of OS.

Hibernation strategies are usually divided into three categories: Standby, Suspend-to-ram and Suspend-to-disk. Standby costs less time for saving less data. Suspend-to-ram strategy can save more energy than Standby, but its wake-up speed is a little slower. Furthermore, once OS loses power provisioning the data in memory will be totally lost. Suspend-to-disk need to save the full image of memory, it is necessary for the disk to provide additional space of memory image size. Besides, lower speed of disk access leads to larger time overhead of hibernation and wake-up. Under the existing hardware and software conditions, we cannot save much energy while rapidly suspending and resuming OS by these three strategies.

Recently, researchers proposed using a new type of storage material called PCM [3] as part of the main memory to replace the Dynamic Random Access Memory (DRAM). As PCM doesn't need dynamical refreshing to maintain data, the OS will not lose its context when it is in hibernation state and does not supply power to PCM. Besides, reading and writing operations against PCM are faster than disk's, so the time overhead of hibernation and recovery is less than the Suspend-to-disk strategy. However, PCM is not widely used in commercial computer system nowadays. In order to prove our design of the innovative Suspend-to-PCM hibernation strategy, we leverage the simulation environment of QEMU to emulate the feature of PCM.

The major work of this paper is designing and implementing the prototype of the Suspend-to-PCM strategy, as well as applying it to achieve rapid suspending and resuming of enterprise server. Our contribution includes:

1. The impact factors affecting virtual machine (VM) performance are analyzed by our approach. Memory size and kernel modification are both considered in this paper. The evaluation releases that memory size has certain impact on the VM performance.

2. The comparison is made between Suspend-to-PCM and traditional suspend strategies in this paper. The final results show performance of our approach is better than traditional method. The experiment demonstrates less time is cost and more power consumption is reduced by our approach.

The rest of this paper is organized as follows. Section 75.2 introduces the related works. Section 75.3 describes the design and implementation of the Suspend-to-PCM strategy in detail. Section 75.4 provides the evaluation of the prototype along with the associated discussion and Sect. 75.5 concludes our work.

75.2 Related Work

Power consumption is becoming increasingly significant for mobile devices, laptops, large scale machines and especially large data centers. With regard to VMs, Nathuji et al. [4] proposed a set of cluster-level management components and abstractions to manage power between different VMs.

Power consumption was also widely concerned in networks area. Zeng et al. [5] proposed SOFA which helps to save energy by minimizing the wake of the Power-Saving Mode clients. Jung et al. [6] presented a holistic controller framework named Mistral to optimize power consumption, performance benefits and the transient costs in cloud infrastructures. In addition, schemes were also developed to both reduce the connection delay time and save energy in wireless sensor networks [7, 8].

Energy savings are very significant for data centers today. A power budgeting system for virtualized infrastructures [9], as well as Auto controller of multiple virtualized resources [10] was employed in data centers to save a large amount of electrical energy efficiently. Other works such as power-aware application placement controller [11] in the context of an environment with heterogeneous virtualized server clusters and [12] were also practical.

In recent years, with the rapid development of computer science and mobile devices, the demand for the storage's performance and energy-saving grows greatly. A lot of research has been done on PCM. Qureshi et al. [13] proposed a new memory architecture. The buffer could improve the write efficiency of PCM and reduce write frequency to extend PCM's usage as well. By simulation experiments, using this kind of architecture could save energy of memory. Wu et al. [14] proposed the combined cache architecture of inter cache Level HCA and intra cache Level HCA.

75.3 Design and Implementation

In this paper, we choose PCM as main memory instead of the traditional DRAM to improve speed of sleep and wake-up of OS and to reduce power consumption of the computer system in sleep mode. There are following two goals in our design.

1. Reduce more time overhead of sleep and wake-up and get higher performance: owing to faster the PCM read and write than disk, time overhead of sleep and wake-up is less than the Suspend-to-disk strategy so that user can get better experience and higher performance.
2. Reduce more power consumption only sacrifice minute performance: as the PCM doesn't need refresh dynamically to maintain the data like DRAM, it needs no power in sleep mode. Our approach aim to get more power saving than Suspend-to-ram strategy by sacrificing minute part performance.

PCM is the key component to the Suspend-to-PCM strategy. Upon considering the design of the Suspend-to-PCM strategy, essential is it to figure out the alternative way to actualize the features of PCM. Theoretically, either customized hardware or software simulation is feasible to be used for emulating the PCM features. No matter which method, in the design and implementation of this paper, it is necessary to guarantee the non-volatile property of PCM and minimize the performance overhead and side-effect of system suspending and resuming.

The whole process of our implement is split into two phases: first is the Suspend-to-PCM phase, second is Resume phase.

The circuit of Suspend to PCM phase is shown in Fig. 75.1. The OS first freezes process in the system and then saves the state of CPU registers after a user makes a sleep command. The data are stored in DRAM as main memory. Next ACPI sleep command will be issued to the underlying hardware platform by the OS and control permission will be turned over to the hardware platform. Due to using QEMU emulator, operation of hardware platform is to be completed. Therefore QEMU emulator is easy to insert the command which saves contents of DRAM to PCM at the moment. QEMU emulator then simulates ACPI sleep state and turns off power supply.

Resume from PCM phase is shown in Fig. 75.2. When the user issues a wake-up operation, a wake event will be generated by devices with a wake-enabled. For the hardware system, this is a wake-up event signal which will trigger the initialization of the chip on the BIOS. For QEMU, this event means creating a new VM. In the process creating the machine, the data stored in the PCM can be

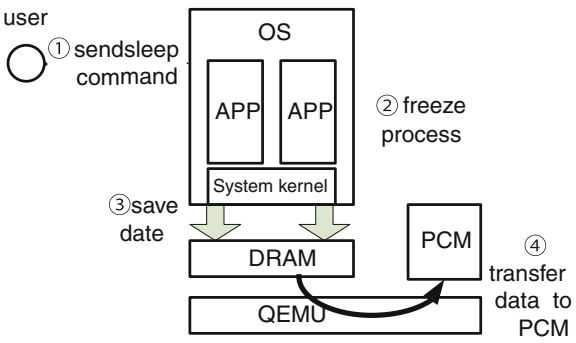


Fig. 75.1 Logical control flow of suspend to PCM

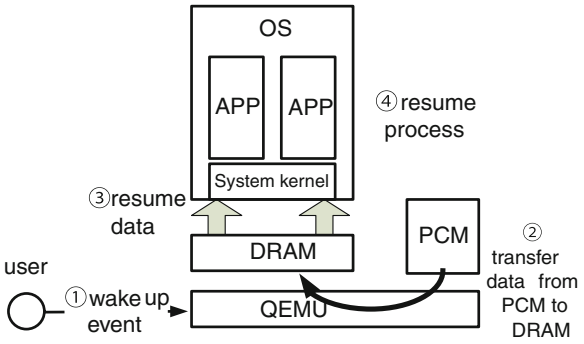


Fig. 75.2 Logical control flow of resume from PCM

directly regarded to memory data of the current VM. Once VM starts, the data are read from the memory to recovery data state and thaw process.

75.4 Experiment Result and Analysiss

We use Intel Core 2 Duo CPU with E8400 3.GHz and version of BIOS is V02.61. The version of QEMU is 0.12.3. And the OS is Fedora 2.6.32.11 with patch TuxOnIce, which is a famous suspend program on Linux. ACPI Suspend Type is S3 (STR).

75.4.1 Performance Evaluation

Figure 75.3 shows the cost of suspend and resume in both real and virtual environments. 'Bare-ram' and 'bare-disk' is the time overhead of suspend-to-ram and suspend-to-disk in real environments. 'Virt-ram' and 'Virt-disk' is the corresponding time in VMs. And 'Virt-PCM' is the time of our implementation of Suspend-to-PCM strategy in VMs.

It is illustrated from the Fig. 75.3 that the performance of Suspend-to-PCM is better than that of suspend-to-disk. We can also see that in the Suspend phase, it takes more time for Suspend-to-disk and Suspend-to-PCM to suspend than for Suspend-to-ram. This is because the former two require moving data from memory to disk or PCM. The process of the latter is much simpler but saves less power. The situation is similar in the Resume phase. Then Suspend-to-PCM can get less time overhead than Suspend-to-disk and less power than Suspend-to-ram.

The most important steps in Suspend-to-PCM strategy are writing data in memory to PCM in the Suspend phase, and reading data from PCM to memory in

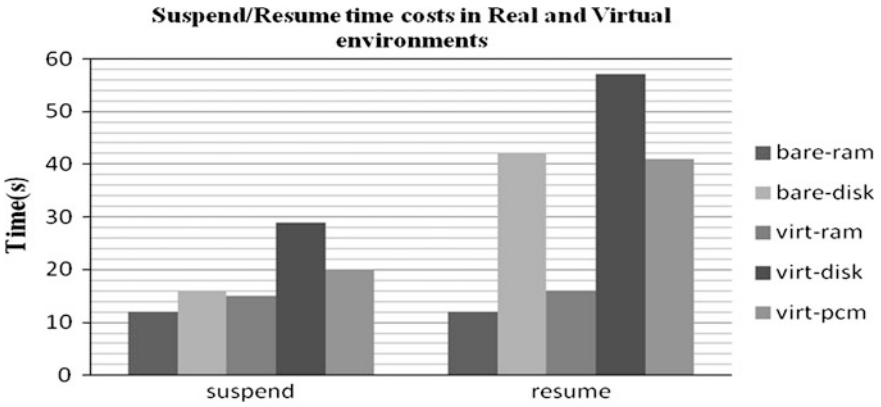


Fig. 75.3 Logical control flow of resume from PCM

the Resume phase. We design the experiment to evaluate the effect of changing the memory sizes on the performance of suspend.

The result is shown in Fig. 75.4. As the memory size increases, the time varies from 16 to 26.5 s in the Suspend phase and varies from 36.7 to 66 s in the Resume phase. The suspend time is consisted of two parts. One is the time of freezing the processes and storing the registers. The other one is the cost of copying data from memory to PCM. In the experiment, there is a few program running in the guest OS, thus causing few time for the first part. With the size of memory increasing, the cost of the second part also increases. Therefore, the total time increases. It has the same situation as in the Resume phase.

The modification to the kernel does not affect the performance of the overall operating system much. As shown in Table 75.1, CPU performance decreases less than 6 %.

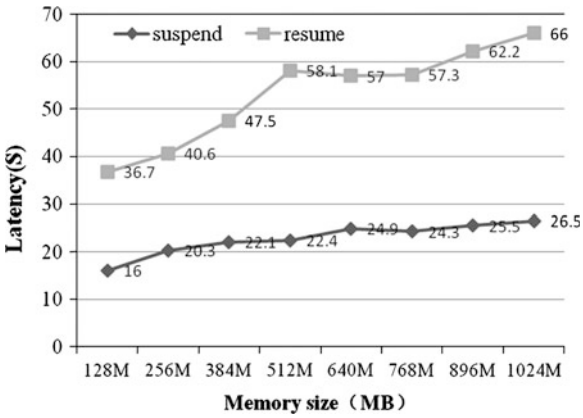


Fig. 75.4 Overhead of suspend and resume phase

Table 75.1 Performance of CPU

Processor, processes—times in microseconds, smaller is better							
	Open clos	Slct TCP	Sig inst	Sig hndl	Fork proc	Exec proc	Sh proc
Kernel	2.63	2.66	0.34	1.27	95.4	341	1293
Kernel modified	2.67	2.72	0.36	1.31	99.4	353	1343
Diff	−0.04	-0.06	−0.02	−0.04	−4	−12	−50
Overhead	1.52 %	2.26 %	5.88 %	3.15 %	4.19 %	3.52 %	3.87 %

Table 75.2 Performance of memory and file system operation

$T_{full}\%$	10 %	10 %	10 %	10 %	10 %	10 %	10 %
$T_{idle}\%$	90 %	75 %	60 %	45 %	30 %	15 %	2 %
$T_{sleep}\%$	0 %	15 %	30 %	45 %	60 %	75 %	88 %
Power save (W)	0 W	9 W	18 W	27 W	36 W	45 W	52.8 W
Power save (%)	0 %	13.69 %	26.97 %	41.06 %	54.75 %	68.44 %	80.30 %

75.4.2 Power Consumption Evaluations

According to the data provided by [14], a general desktop consumes about 80-100 W of electronic power and another 35–80 W of the monitor power when it is active. If the CPU is idle, the consumption of the power is 60–80 W.

Assume that the power consumption is 6.25 W of an idle computer, and is 95 W when the CPU is fully used, and is 2.5 W when the computer is sleeping. The ratios of idle, fully loaded and sleeping of the computer are represented by $T_{idle}\%$, $T_{full}\%$ and $T_{sleep}\%$ respectively. The three ratios must comply with the following condition:

$$T_{idle}\% + T_{full}\% + T_{sleep}\% = 100\% \tag{75.1}$$

Therefore, if the computer uses our implemented Suspend-to-PCM strategy, the saved energy of the computer can be calculated by the following formula:

$$T_{sleep\%} \times (P_{idle} - P_{sleep}) = x\% \times (6.25 - 2.5\,\mathrm{W}) \tag{75.2}$$

In which, x% is the time ratio of sleeping. And we just assume $T_{full}\%$ equals 10 % for simplicity. With x changing, the result of saved power is show in Table 75.2.

75.5 Conclusion

With PCM's advantages of non-volatile and low power consumption, the paper introduces a new design that accelerates the processing of system suspend and resume. Since PCM is still in the experimental development stage, prototype in the

paper is implemented in the emulator of QEMU. Compared with the conventional strategies, the experimental result shows that the Suspend-to-PCM strategy has a better performance than Suspend-to-disk and sacrifices a small part of performance in exchange up to 80.3 % energy saving than Suspend-to-ram. Besides, factors influencing Suspend-to-PCM in virtual environments are analyzed in the experiment. With the size of memory increasing, the cost of the Suspend-to-PCM increases. Meanwhile, effect of kernel modification on performance is less than 6 %. To sum up, it is believed that the Suspend-to-PCM strategy is capable lowering the power consumption in data centers especially when the cluster volume is large.

References

1. Brown, R., et al.: Report to congress on server and data center energy efficiency. Public law 2008, 109–431 (2008)
2. Chase, J., Anderson, D., Thakar, P., Vahdat, A., Doyle, R.: Managing energy and server resources in hosting centers. ACM SIGOPS Oper. Syst. Rev. **35**(5), 103–116 (2001)
3. Qureshi, M., Srinivasan, V., Rivers, J.: Scalable high performance main memory system using phase-change memory technology. ACM SIGARCH Comput. Architect. News **37**(3), 24–33 (2009)
4. Nathuji, R., Schwan, K.: Vpm tokens: virtual machine aware power budgeting in datacenters. In: Proceedings of the 17th International Symposium on High Performance Distributed Computing, ser. HPDC'08, pp. 119–128 (2008)
5. Zeng, Z., Gao, Y., Kumar, P.: Sofa: a sleep-optimal fair attention scheduler for the power-saving mode of WLANs. In: Distributed Computing Systems (ICDCS), pp. 87–98 (2011)
6. Jung, G., Hiltunen, M.A., Joshi, K.R., Schlichting, R.D., Pu, C.: Mistral: dynamically managing power, performance and adaptation cost in cloud infrastructures. In: International Conference on Distributed Computing Systems, pp. 62–73 (2010)
7. Ghidini, G., Dasn, S.: Energy-efficient markov chainbased randomized duty cycling scheme for wireless sensor networks. In: Distributed Computing Systems (ICDCS), pp. 67–76 (2011)
8. Chen, X., Jin, S., Qiao, D.: M-psm: mobility-aware power save mode for IEEE 802.11 WLANs. In: Distributed Computing Systems (ICDCS), pp. 77–86 (2011)
9. Lim, H, Kansal, A., Liu, J.: Power budgeting for virtualized data centers. In: Proceedings of the 2011 USENIX conference on USENIX annual technical conference, ser, pp. 5–5 (2011)
10. Padala, P., Hou, K.-Y., Shin, K.G., Zhu, X., Uysal, M., Wang, Z., Singhal, S. Merchant, A.: Automated control of multiple virtualized resources. In: Proceedings of the 4th ACM European conference on Computer systems, ser, pp. 13–26 (2009)
11. Verma, A., Ahuja, P., Neogi, A.: Pmapper: power and migration cost aware application placement in virtualized systems. In: Proceedings of the 9th ACM/IFIP/USENIX International Conference on Middleware, ser. Middleware'08, pp. 243–264 (2008)
12. McCullough, J.C., Agarwal, Y., Chandrashekar, J., Kuppuswamy, S., Snoeren, A.C., Gupta, R.K.: Evaluating the effectiveness of model-based power characterization. In: Proceedings of the 2011 USENIX conference on USENIX annual technical conference, ser, pp. 12–12 (2011)
13. Wu, X., Sun, G., Dong, X., Das, R., Xie, Y., Das, C., Li, J.: Cost-driven 3d integration with interconnect layers. In: Proceedings of the 47th Design Automation Conference, ser. DAC'10, pp. 150–155 (2010)
14. Das, T., Padala, P., Padmanabhan, V.N., Ramjee, R., Shin, K.G.: Itegreen: saving energy in networked desktops using virtualization. In: Proceedings of the 2010 USENIX conference on USENIX annual technical conference, pp. 3–3 (2010)

Chapter 76
A Web Content Recommendation Method Based on Data Provenance Tracing and Forecasting

Zuopeng Liang and Yongli Wang

Abstract How to choose an appropriate releasing strategy for site content, and which one caters to user's habits, have become the main challenges. This article provides a provenance-aware model to design the content of the website. Based on the user's browsing history data, it constructs timed automaton that can trace the provenance of the data to find what the user may be interested in, and it establishes a Markov chain model to determine the content of the link relationship. Experiments show this model not only meets the dynamic needs of users when they browses the site, but also gives certain options to the administrator of site content. It provides recommending result efficiently and should have a bright application prospect.

Keywords Content recommendation · Data provenance · Timed automation · Markov chain

76.1 Introduction

With the advances in Web technology, the amount of information inherent in the Internet became more. How to provide targeted, appropriate information to user presents many challenges for the field of information retrieval (IR).

Z. Liang (✉)
Department of Economics, Nanjing University, Nanjing, China
e-mail: 896073265@qq.com

Y. Wang
School of Computer Science and Engineering,
Nanjing University of Science and Technology,
Nanjing, China
e-mail: yongliwang@mail.njust.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_76,

The site administrators not only need to obtain reliable data in a complex, interconnected network, but also need to explore all kinds of network users' needs. Web-based development has changed the traditional development method such as data flow, storage, and statistics. Firstly, it is extremely easy to access data and copy data on the network environment, which causes the reliability of data is difficult to be guaranteed; secondly, the pages in the browser are evolving and expanding, and the relationship between the pages is relatively unstable.

In recent years, Web applications are booming, and the study on the Web personalized recommendation are quietly rising [1]. However, the defect of these algorithms is that it cannot meet the needs of most users and only provide a recommendation for the specified user. A data provenance consists of the entire processing history of the data, which includes its source and all subsequent processing steps [2]. If we regard the click history or browsing log as the data provenance of certain user, we can use provenance workflow to trace the habit of the user. There are two approaches to calculate the data provenance: query inversion mode ("lazy" approach), and labeling mode ("eager" approach) [3]. This article uses "eager" approach to calculate the data provenance.

There are some existing methods, such as timing diagram, provenance diagram, XML DTD (XML Schema), to realize a labeling mode based on workflow provenance [4, 5]. In the forecast, policy makers always expect subjective judgments as much as possible close to the objective judgments [6]. Markov chain algorithm provides us with the scientific method to resolve these problems [7]. However, the existing prediction models only process the basic data structures and do not fully mine inter- relationship of access log.

We annotate these data to establish the state of the automaton, and the Content Manager can achieve the content semantics that users concern. On this basis, we use the extended Markov chain model to predict the order of the browsing the web content by user.

76.2 Definition of Timed Automata Model

The proposed recommendation model supports time constraints about accessing Web network, the Web Workflow is a real-time workflow.

Definition 1 (Web real-time workflow): Web real-time workflow consists of activities, participants and dependencies between activities. An activity refers to a separate step in the business processes; it can be viewed web content that users browsed. A participant mainly refers to the user. A dependency determines the execution order of activities and data flow between activities, which is the conversion between the web content.

A timed automaton is widely used in the modeling and verification of real-time systems. Constructing a timed automaton for Web accessing can record timing constraints relationship between which the user browses the web pages.

Definition 2 (Time automata): timed automata A is a seven-tuple $(S, S_0, \Sigma, X, I, E, F)$, where S is a set of finite states, which indicates all state of web content. $S_0 \subseteq S$ is the initial state set, which denotes the first web page content. Σ is a finite event set, which includes clicking on the link, closing the page. X is a finite clock set. I is a mapping, which assigns a timed constraint in clock constraints set $\Phi(X)$ for each state s in S.

When the clock of the page state does not satisfy the timing constraint, the automata must be able to perform the migration to leave the web page content. $E \subseteq S \times S \times \Sigma \times X \times \Phi$ is a collection of content conversion. A conversion $< s, s', \alpha, \lambda, \delta >$ denotes that when an event α occurs, the web content converse from s into s', where $\lambda \subseteq X$ is the resetting zero value of the clock collection while the conversion occurs; δ is a time constraint in X, which specifies the time constraint to met the conversion condition. We denote the time constraints of users browsing in specified contents as $s \xrightarrow{\alpha,\lambda,\delta} s'$, where α, λ and δ can by default. $F \subseteq S$ is the set of states of termination of the web content.

We construct a four-tuple (t, ip, s, s') for a specific website. This data structure represents that a user, whose IP address is *ip*, transferred from the content *s* to the content *s*' during time *t*. For example, Fig. 76.1 shows a timed automaton based on 3 web pages content.

While the state transition occurs, timed automata record and transfer the identity of the relevant web content. The timed automaton can constantly update data set to reflect the latest status of the user. Thus, it provides considerable flexibility and scalability.

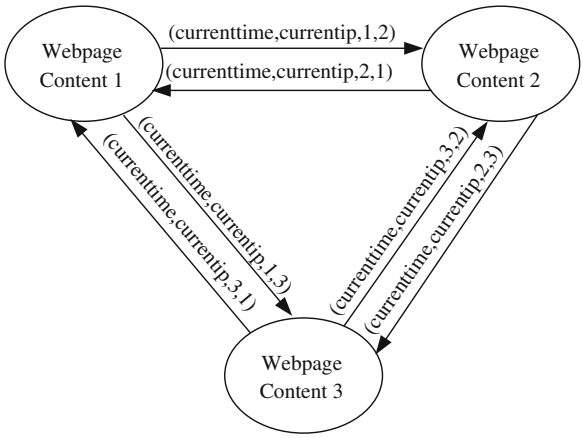


Fig. 76.1 Three web pages of timed automata

76.3 Forecasting the Content of Links Based on Markov Chain

When users are browsing the web, it is difficult to identify the inherent regularity to discover clicking on which link and selecting the link in what order in the complex environment. We propose a method to predict the linking content that the user select based on time automata.

76.3.1 The Markov Chain for Web Content

Definition 3 (Web Markov chain): Suppose that $\{X(n), n = 0, 1, 2, \ldots\}$ is a random sequence and Q is a discrete state space, if for any m non-negative integers $n_1, n_2, \ldots, n_m (0 \leq n_1 < n_2 < \ldots < n_m)$, any natural number k, and arbitrary $i_1, i_2, \ldots, i_m, j \in Q$ meet:

$$\begin{aligned} &\mathrm{P}\{X(n_{m+k}) = j | X(n_1) = i_1, X(n_2) = i_2, X(n_m) = i_m\} \\ &= \mathrm{P}\{X(n_{m+k}) = j | X(n_m) = i_m\}. \end{aligned} \tag{76.1}$$

We call $\{X(n), n = 0, 1, 2, \ldots\}$as a Markov chain. If n_m represents the present moment, $n_1, n_2, \ldots, n_{m-1}$ represents the last moment, and n_{m+k} represents the future moment. Eq. (76.1) shows that the webpage content j in the future moment n_{m+k} only depends on the webpage content in the present moment n_m, In the other word, the webpage content j in the future moment n_{m+k} is independent of the webpage content in m-1past moments $n_1, n_2, \ldots, n_{m-1}$. This reflects the characteristic of Markov process.

Markov chain is a particular case of the Markov process. Markov chain model of the Web content describes that the state of webpage content change from the past to the present, and from the present into the future, which changes one by one. It like a chain and has no aftereffect. The Markov chain reflects the randomicity of user's browsing behavior.

We denote the data to be forecasted as an instance of seven-tuple from timed automata. The transition probability matrix can be updated dynamically, and the calculation process can be executed according to the recurrence relation. As long as the initial web content that the transition matrix obtained is accurate, the future of predicted link results has certain credibility.

A random sequence with the characteristics of the Markov chain can be divided into m states, for example, $i_1, i_2 \ldots i_m$ in Eq. (76.1)… and j. In this paper, the interlinked webpage content can be seen as the different status. The state space is $Q \subseteq S$, which represents the collection of webpage content. For example, $Q = \{1, 2, 3, 4, 5, 6\}$, each element corresponds to the contents 1 to the contents 6.

76.3.2 State Transition Matrix

The basic idea of Markov prediction is to obtain the state transition matrix of sequence using the original data sequence. The goal of Markov prediction is to estimate the future development trend according to the state transition matrix.

Definition 4 (Conditional probability): The condition probability $P\{X(n_{m+k}) = j|X(n_m) = i\} = P_{ij}(m,k)$, we call $P_{ij}(m,k)$ as k-step transition probability at moment n_m. After k-step transition, Web content i inevitably reach one webpage content in set Q, and only to reach one webpage content. Thus, k-step transition probability meets the following conditions:

$$P_{ij}(m,k) \geq 0, i,j \in Q; \tag{76.2}$$

$$\sum_{j \in E} P_{ij}(m,k) = 1, i,j \in Q. \tag{76.3}$$

Assume that the transition probability P_{ij} (m, k) of Webpage content does not depend on the Markov chain of m, we call P_{ij} (m, k) as homogeneous Markov chain. The status of the webpage is relevant to the starting content i, transfer step number k and the reaching content j. It is not relevant to m. At this point, we denote k step transition probabilities as P_{ij} (k), namely:

$$P_{ij}(k) = P_{ij}(m,k) \tag{76.4}$$

We use a transition probability matrix to represent the changed probability during transferring from one state to another state in Markov chain. For Webpage content space $Q = \{1, 2, 3, 4, 5, 6\}$, the corresponding one step state transition matrix is as following:

$$P(1) = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{16} \\ P_{21} & P_{22} & \cdots & P_{26} \\ \cdots & \cdots & \cdots & \cdots \\ P_{61} & P_{62} & \cdots & P_{66} \end{bmatrix} \tag{76.5}$$

The i-th row, j-th column element P_{ij} in P (1) represents the one-step transition probability while the Markov chain model transferred from the webpage content i to webpage content j.

76.3.3 Calculation of Transition Probability Matrix

In timed automata model which annotates browsing workflow, when the transition of state occurs, the timed automata record two webpage content that corresponds

to the transferring, and denote it as the four-tuple (t, ip, s, s'). We let N_{ij} represents the number of the transition during the t period while transfer page i to page j. During t statistics period, the number of four-tuple meet the $s = i$ and $s' = j$ is N_{ij}.

$$P_{ij} = \frac{N_{ij}}{\sum_{j=1}^{6} N_{ij}} (1 \leq i, j \leq 6) \tag{76.6}$$

The k-step transition probability is $P_{ij}(k)$, we can obtain the recurrence relations using C-K equation:

$$P\ (k)\ = P\ (1)\ P\ (k-1)\ = P\ (k-1)\ P\ (1) \tag{76.7}$$

Thus,

$$P\ (k)\ = P\ (1)\ P\ (1)\ldots P\ (1) = P\ (1)^k \tag{76.8}$$

It is vital to keep the one-step-state transition matrix correctness, which ensures the forecast close to the true value of the k-step transition of webpage content.

76.4 Experimental Analyses

Experiments establish on the basis of a small website. We use the Visual Studio 2005 to implement all the algorithms. The base station server is an IBM compatible computer (CPU Intel(R) Xeon(R) E5620 2.40 GHz and RAM 12 GB), and the OS is Windows 7.

We found that the state transition matrix approach stable when $k = 5$. For this experiment, the one-step transition will be able to reflect the user's browsing habits. According to one-step state transition matrix, we use the web page, which was visited with the maximum probability, to speculate the user's browsing order. For example, we should push the webpage content in the following order for the above matrix P: webpage content 1 → webpage content 4 → webpage content 5 → webpage content 6 → webpage content 3…

76.4.1 Verification of Reliability

The algorithm calculates the transferring probability of webpage content based on the data provenance of timed automata. In order to verify the reliability of the probability, we set the order 1 → 4 → 5 → 6 → 3…, in which the user's actual browsing the webpage within a period as a standard. And we explore the variance between the probability matrix generated from timed automata and the actual probability matrix.

We sample the running example and achieve the statistical transferring sequence as follows:

(1) Start from the webpage content 1 and go to webpage content 2, 3, 4, 5, 6, probability: 5.1, 36.7, 48.2, 3.5, 6.5 %;
(2) Start from the webpage content 2 and go to webpage content 1, 3, 4, 5, 6, probability: 18.6, 14.7, 9.7, 10.5, 46.5 %;
(3) Start from the webpage content 3 and go to webpage content 1, 2, 4, 5, 6, probability: 15.5, 8.7, 57.2, 12.5, 6.1 %;
(4) Start from the webpage content 4 and go to webpage content 1, 2, 3, 5, 6, probability: 11.1, 1.7, 17.2, 23.5, 46.5 %;
(5) Start from the webpage content 5 and go to webpage content 1, 2, 3, 4, 6, probability: 14.6, 12.2, 24.7, 12.0, 36.5 %;
(6) Start from the webpage content 6 and go to webpage content 1, 2, 3, 4, 5, probability: 25.1, 16.3, 32.1, 4.2, 22.3 %;

The probability variances in one-step state transferring matrix are 0.068, 2.468, 1.07, 0.402, 3.102, 2.088. Variance range is less than 3 basically, which is within the acceptable range. Figure 76.2 shows the trend of the variance of state transition probability.

We construct the proposed algorithm on the basis of the access log and click historical data, and we fully take into account the actual user's browsing habits. Thus, the data set that algorithm obtained is authentic. One-step state transition matrix is been calculated statistically by analyzing the history of user's clicking on a link. Therefore, we believe the state transition probability that generated from timed automata is credible.

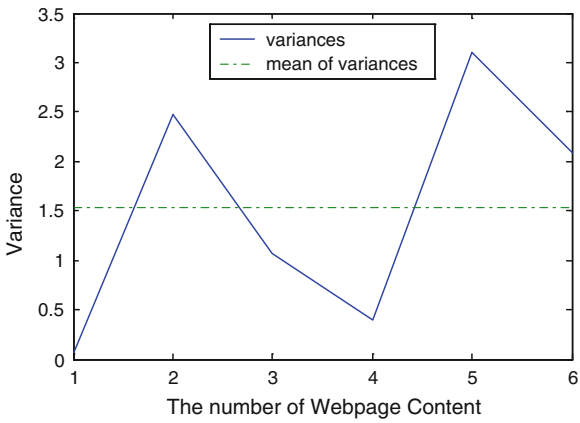


Fig. 76.2 The variance of the state transition probability distribution

76.4.2 Validation of the Recommended Quality

In order to verify the performance of the algorithm, we compare the proposed algorithm and the collaborative filtering algorithm that mentioned in the related work. Collaborative filtering algorithm is the most widely used personalized content recommendation algorithm.

In order to measure the quality of the recommended content intuitively, we use the average absolute deviation (MAE) [8] as referral service quality standards. Support that the forecasted set that consists of N user's scores is $\{m_1, m_2 \ldots, m_N\}$, the actual score set of user's ratings is $\{ n_1, n_2 \ldots, n_N \}$, the MAE defined as following:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N} |m_i - n_i|}{N} \tag{76.8}$$

Figure 76.3 shows the MAE comparisons of the collaborative filtering algorithm and the proposed in this paper. We can conclude that the accuracy of the automata-based Webpage content recommendation method is higher than one of the collaborative filtering algorithms.

Collaborative filtering algorithm uses the similarity between the users to filter information. However, the main drawback of this algorithm is that the similarity bears sparsity problems and scalability problems, and the similarity of the user calculated by this algorithm has a certain deviation. The proposed algorithm in this paper directly establishes in the habits of user's accessing content; thus the MAE is relatively low. In addition, we can update the recommended strategy and dynamically know the user's new interest; thus the proposed algorithm is high flexibility and wide applicability.

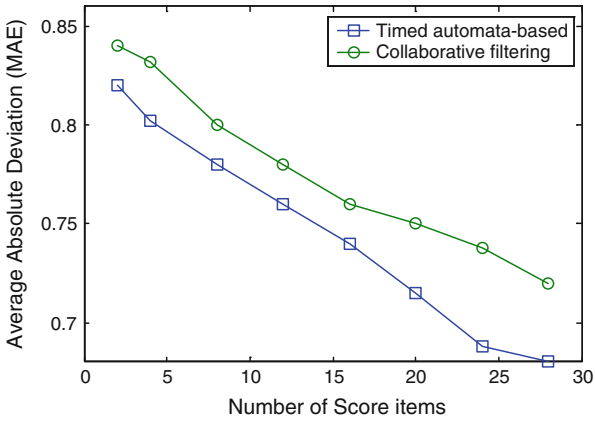


Fig. 76.3 MAE comparisons of two algorithms

76.5 Conclusion

In recent years, Network resources have increasingly become an indispensable part of people's lives, which brings a golden opportunity for businesses recommendation. Researchers creatively use timed automata, which is constructed by a labeling workflow method, to find Webpage content that is welcomed by users in this paper. Researchers employ Markov chain principle to establish the content link model. Based on this method, the website content administrator can design web pages that users are most interested in. This model is not only convenient for the user to view, but also improves the efficiency and quality of the user's view.

Acknowledgments This work is supported in part by China Postdoctoral Science Foundation (2012M511227), Jiangsu Province Postdoctoral Science Research Fund (1101073C), National Natural Science Foundation of China (61170035), Natural Science Foundation of Jiangsu (BK2011022, BK2011702).

References

1. Wang, J., Tang, X.: Personalized recommendation algorithm research based on content in social network. Applica. Res. Comput. **24**(8), 1248–1250 (2011)
2. Woodruff, A., Stonebraker, M.: Supporting fine-grained data lineage in a database visualization environment. In: Proceedings of the International Conference on Data Engineering (IEEE ICDE), pp. 91–102 (1997)
3. Liu, X., Wan, C.: Research on data provenance an overview. Sci. Mosaic **1**, 47–52 (2005)
4. Kiran-Kumar, M., Uri, B., David, A.H., Peter M., Diana M., Daniel M., Margo S., Robin S.: Layering in Provenance Systems. USENIX Annual technical conference (2009)
5. Zoé, L, Christophe, L., Spyro, M.: Storing Scientific Workflows in a Database. ACM (2009)
6. Geoffrey, R.G., David R. S.: Probability and Random Processes. Oxford University Press, USA; 3 edition (2001)
7. Deng, M.: Research on the top three places in men's modern pentathlon of olympic using gray markov chain prediction model. Mathemat. Pract. Theory **41**(2), 134–137 (2011)
8. Herrlocker, J., Konstan, J.: Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. **22**(1), 5–53 (2004)

Chapter 77
Research and Implementation of Massive Data Atlas Visual Strategy

Peng Wang and Shunping Zhou

Abstract In order to solve the display problem with massive data atlas, researchers propose five kinds of visualization strategy to improve the display efficiency in this paper. By using the strategy of establishing the atlas map frame index, hierarchical visualization, real-time dynamic projection, double buffering + multi-threading data dynamic annotation, local cache based on spatial data mining visualization, researchers basically solve the problems existing in the process of multi-source heterogeneous mass data atlas display. The strategy has achieved on the MapGIS K9-based platform and successfully applied to the basis of the outcome of mapping system of the national fundamental geographic information center. Researchers found that the display strategy is significantly more efficient than traditional visualization strategy. Thus, the effectiveness and accuracy of the algorithm are verified.

Keywords Massive data · Atlas · Visualization strategy

77.1 Introduction

With the widespread application of GIS, an increasing number of industries are involved in GIS. A variety of GIS data systems do not have a unified standard, but it's likely to intercross and need for integration of multi-source heterogeneous data [1]. In order to manage the multi-source heterogeneous mass data, we choose atlas model of MapGIS platform, and firstly the display efficiency of massive data atlas is an urgent problem to be solved.

P. Wang (✉) · S. Zhou (✉)
Faculty of Information Engineering,
China University of Geoscience(Wuhan), Wuhan, China
e-mail: wp19850104@163.com

S. Zhou
e-mail: zhouspin@yahoo.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_77,

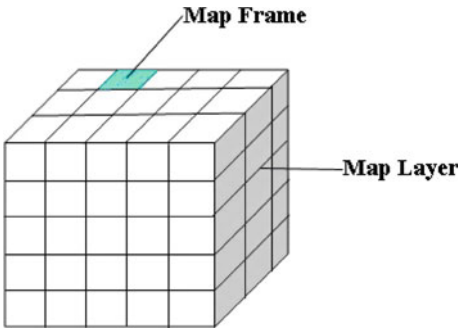


Fig. 77.1 Atlas cube management model

77.2 Atlas Model of MapGIS Platform

Atlas model management of MapGIS platform adopted a cube model (Fig. 77.1) based on map frames and map layers [2]. Map frame units manage spatial data in the atlas management model, which constitute horizontal grid. Individual frames are composed by several layers in vertical overlap [3], layers division correspond to the division of the input editing layer class, such as administrative boundaries layers, water system layers. Layer horizontal division makes gallery management more rational, more layered.

77.3 Atlas of Huge Amounts of Data Visualization Strategy

In order to reasonable and efficient display multi-source heterogeneous atlas of huge amounts of data; we propose the following visualization strategy.

77.3.1 Establish the Atlas Map Frame Index

The map frame management index is established by the characteristics of the spatial data's horizontal framing and vertical stratification [4], which is beneficial to query and select the atlas frame data. The atlas visualization process is divided into two processes (1) Using the particular range to query extract spatial data, (2) Drawing this particular data on the screen. Creating map index can improve the speed of querying extract data, is the first step to improve the efficiency of display process. Testing proved that when the space numbers of features are about 300 million or so, if you create a frame index and use the index filters the space data, the shortest response time was only 2 s, so we can create map frame index to improve the query efficiency of massive data atlas.

77.3.2 Hierarchical Visualization Strategy

When atlas of graphic information is displayed to the subscriber, the region of display information received by the user was only the region of the computer screen [5], due to the limited display area of the atlas; user is difficult to obtain their useful information unless we choose the display ratio control and content hierarchical display. To make atlas more naturally display the data layers and give user a better interactive experience, we can set the displayed radio to achieve the purpose of the outline. After setting a suitable displayed radio, the atlas can dynamically adjust displayed loading capacity. Due to adjust the display ratio, in small scale, the map information displayed relatively rough, in large scale, relatively detailed map information displayed. The hierarchical visualization strategy can be a good solution to the problem of multi-scale feature class displaying in the same view. Therefore, the use of hierarchical visualization strategy can quickly improve efficiency of massive data atlas display.

77.3.3 Real-time Dynamic Projection Strategy

In view of atlas manages massive heterogeneous data with different reference coordinate system; we must solve the problem of unified reference coordinate system before the integration of heterogeneous data show. We can set the reference coordinate system of the atlas when atlas integrated display [6], atlas will real-timely dynamic project the heterogeneous data to the setting coordinate system, and then unify display projected data, so end-users can preview of the multi-source heterogeneous data in the same view, such as raster and vector data overlay display in the same view is shown in Fig. 77.2, the use of dynamic projection can solve different reference coordinate system of heterogeneous data integration display.

77.3.4 Using of Double Buffering + Multi-threading to Achieve Efficient Mass Data Dynamic Annotation

Double buffering technique means to not modify visible graphics cache conditions, and request a equal size storage of screen area in memory, the image which will be displayed is simultaneously drawn in the virtual memory, then directly copy virtual memory to the visible pattern cache (because of memory copying, the entire copy process is very quickly). Drawing process need not to directly operate on the screen, but operate the background of virtual memory, so it is a good solution to

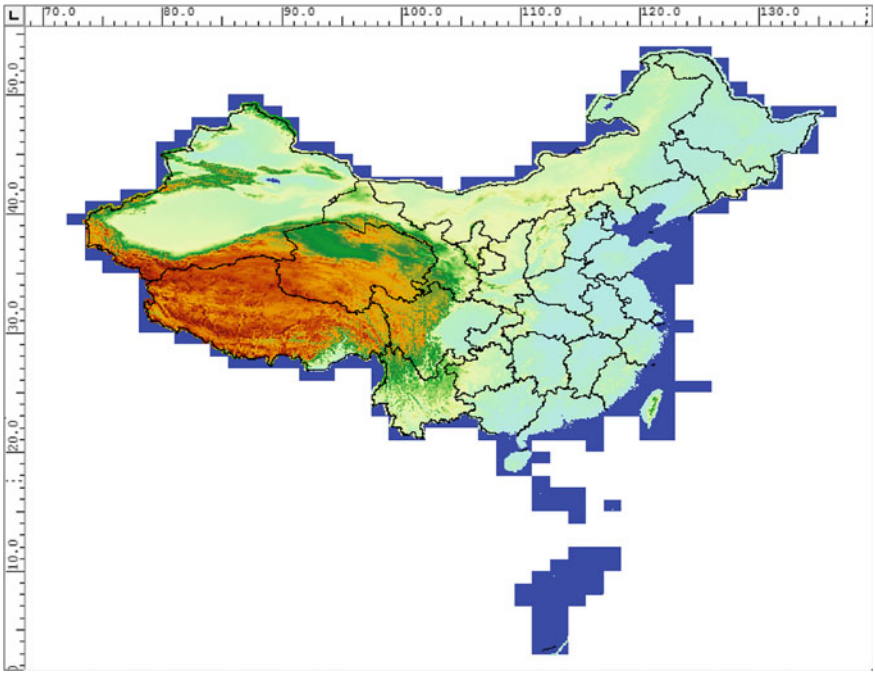


Fig. 77.2 Atlas of vector and raster data unified display

the jitter, flicker problem of screen draw. Therefore, using this method can significantly improve the drawing speed, enhanced graphics effects. Its principle is shown in Fig. 77.3.

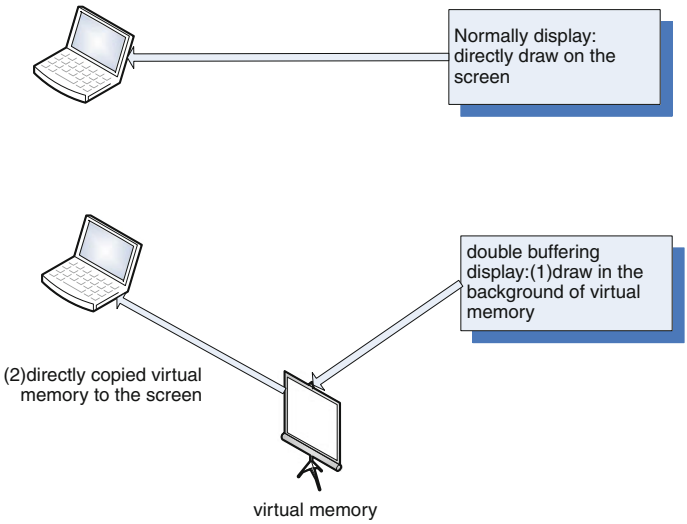


Fig. 77.3 Double buffering schematic diagram

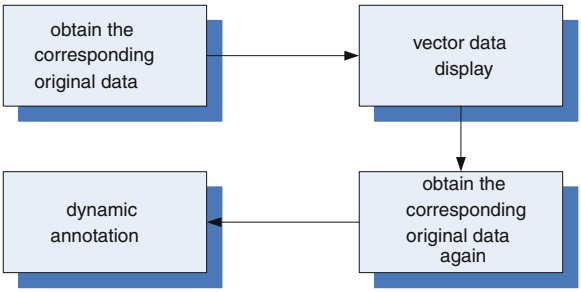


Fig. 77.4 Traditional dynamic annotation mode

Figure 77.4 depicts a conventional dynamic annotation process, which is the most intuitive and simple manner, after the completion of the vector data display, then obtain the corresponding original data from the database in accordance with the display range. After the dynamic projection of the original data, calculate the annotation point position and its related information, and then dynamic annotation according to the annotation information. The advantage of this way in the practical application of the process is intuitive and simple, its correlation function is easy to implement, but its obvious shortcomings is that the method process is extremely inefficient in the practical application.

Figure 77.5 describes improved double buffering dynamic annotation mode, In the process of the vector data atlas show, when the data's display range is acquired from the database, the projection data's spatial information associated with the attribute information stored in the corresponding memory buffer, using of general flow to display vector data with range, after completing the display of the vector data, then calculating the note according to the space within the data buffer (completed dynamic projection calculation) and attribute information. This way greatly improved the efficiency of the whole show, but affects the speed of the graphics display, due to adding a dynamic annotation process in the graphical display of the process.

Figure 77.6 depicts a multi-threaded + double buffering to achieve dynamic annotation, the double buffering dynamic annotation based on a multi-threading support, new threads to achieve double buffering of dynamic annotation, so as to avoid affecting the image display speed and solve the problem which exists in the double buffering dynamic annotation, multi-threaded + double buffering

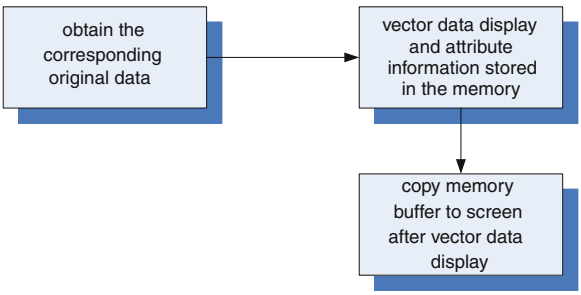


Fig. 77.5 Double buffering dynamic annotation mode

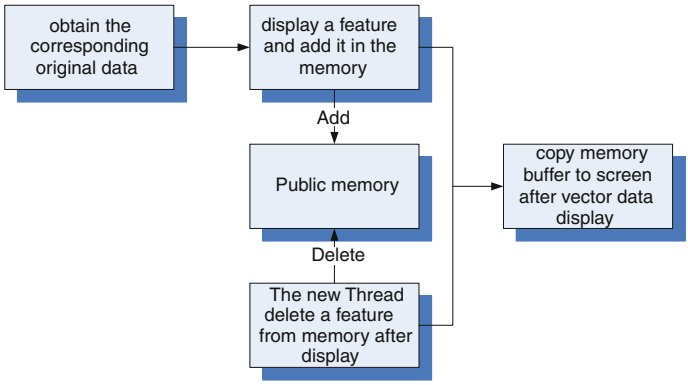


Fig. 77.6 Multithreaded dynamic annotations

combined dynamic annotation has a very high efficiency display, but the specific logic implementation is relatively complex, need to accurate handle the critical region, mutex semaphore.

77.3.5 The Local Cache Based on Spatial Data Mining Visualization Strategy

Spatial data mining refers to using space or attributes condition to extract user's interested spatial patterns, characteristics, spatial and non spatial data universal relation from the spatial database [7].

In order to efficiently display data atlas, we consider using a local cache. As the amount of data is too large, directly cached massive amounts of data from the database to the local which is time-consuming, wasting of disk space. Therefore we need to analyze the characteristics of the data, as well as the regular habits of users to preview data, then organize and manage the cached data, thus efficiently show users' most interested content. We carefully analysis of the spatial and attribute data according to the idea of spatial data mining, and ultimately achieve a unique visualization strategy based on local cache of spatial data mining.

Based on spatial data mining visualization of local cache strategy process is: firstly, we choose a layer of the original spatial data in the massive data atlas for data mining by attribute extraction classification, and then create a local cache atlas to manage the multiple local extraction results class. By the atlas hierarchical visualization strategy such as configuration transition display ratio or layer display sequence, we can make local cache atlas meet our users' preview habits, natural transition, and rapid response effect. The layer data in the atlas, no longer display the original spatial data, but display the configured local cache atlas. We can greatly improve the spatial data visualization efficiency by setting the local cache atlas.

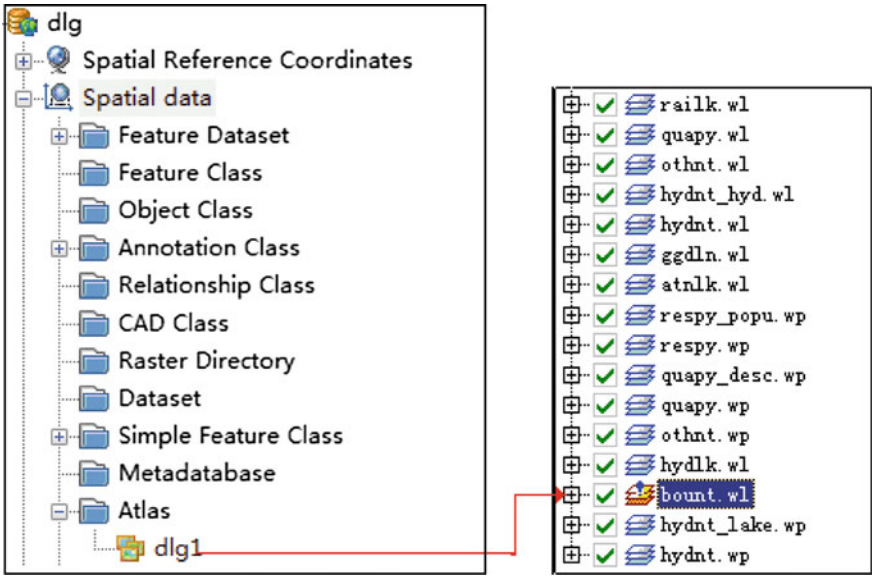


Fig. 77.7 Bount.wl layer example

In Fig. 77.7, we use the National Geographic Information Center's 1:25 million achievement database map set's boundary line layer (bount.wl) as an example.

77.4 Conclusion

Massive data atlas visualization strategy presented in this paper basically solves the problems that exist in the process of multi-source heterogeneous mass data display, and significantly improve the efficiency of displaying huge amounts of data. It also solves the multi-scale, multi-source heterogeneous data integration display problems, and massive data dynamic annotation inefficiencies. This paper innovatively has proposed visualization of spatial data mining strategy based on local cache, making the local cache with flexible configuration features. Caching strategy greatly optimizes the massive data's visual effects. Due to the complexity of the vast amounts of data, there are still some visual problems to solve in pending further study, such as how to ensure the consistency of the cache data with the original data, the establishment of dynamic tile cache.

Acknowledgments This research was supported by the National Science and Technology Support Program, China (No.2012BAB11B05) and by the Research Center for GIS Software and Application Engineering, Ministry of Education Program, China University of Geosciences, Wuhan (No.20111113; No.20111114). Peng Wang and Shunping Zhou thank the great efforts they have paid.

References

1. Zhou, S.P., Wang, P.: The Integration of Multi-source Heterogeneous Data Based on Middleware, ICISE (2009)
2. Fang, F., Wang, P.: The Algorithm and Realization of Vector-Data Attenuation-Junction of Arbitrary Scope, ESIAT (2009)
3. Zhou, S.P., Wei, L.P., et al.: A study of integration of multi-source heterogenous spatial data. Mapp. Sci. **25**, 25–27 (2008)
4. Song, G.F., Zhong, E.S., et al.: A study on seamless integration of multi-sources spatial-data (sims). Prog. Geogr. **19**, 110–115 (2000)
5. Huang, Z.Q., Feng, X.Z.: The research of spatial heterogenous data source integration of GIS. J. Image Grap. **9**, 904–907(2004)
6. Wu, X.C.: Design and Implementation of Geographic Information System. Publishing House of Electronics Industry, Wuhan (2002)
7. Wu, X.C.: Application Software development of Geographic Information System. Publishing House of China University of Geoscience, Wuhan (2001)

Chapter 78
The Autonomous System Topology Build Method Based on Multi-Source Data Fusion

Jingju Liu, Guozheng Yang and Huixian Chen

Abstract The Internet topology can be analyzed from four different granularities. This paper aims at the Internet AS level topology. Firstly, it introduces some current related researches about this project, and summarizes the hierarchy structure of AS topology. Then it analyses the obtaining method of three possible data sources which are BGP route info, IRR data and traceroute probing data. According to the different characteristic of each data source, researchers list the advantages and disadvantages when using these data to build AS topology. Due to the excellence of BGP route info in node perfectibility and recognizing correctness, a new AS level topology build algorithm is put forward based on BGP route info, and fused with IRR data and traceroute data. Each step of algorithm is discussed in detail. To prove the algorithm validity, three types of real data were collected to build an AS topology in experiment. The node count and link count are calculated by each step, and the node degree distribution curves of our result are compared with related research, which show that our algorithm is effective in building a more comprehensive AS topology.

Keywords Data fusion · AS level · Network topology · Build method

78.1 Introduction

Internet has developed to a large scale complex network today. In order to understand the Internet structure, we can recognize it from four levels. There are IP interface levels, router level, POP (Point Of Presence) level and AS (Autonomous System) level, as shown in Fig. 78.1.

J. Liu · G. Yang (✉)
Network Department 603 Lab, Electronic Engineering Institute, Heifei, China
e-mail: yangguoz0218@163.com

H. Chen
New Star technology Institute, Heifei, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_78,

The lowest level is IP interface topology which describes the logic connection between all IP interfaces from different routers. The second level is router topology which can be obtained from IP interface topology by combining several IP interfaces belong to same router into one network node. Combining routers in same geographic position into one network node, we can further get the third level POP topology. The last level is AS topology. AS is the short for Autonomous System which means a group of routers network constructed by one manage department. Using AS node denotes its interior network and AS link denotes the business relationship between different ASes, the AS topology can reflect the macrostructure of whole Internet.

At present, BGP route info, IRR data and traceroute probing data are three main sources to build AS topology. Mahadevan distills the common AS nodes from these three data sources, and builds three type of AS topology from each data source [1]. Oliveira analyzes the evolution of AS node and link by using these three data sources respectively [2]. In order to study the characteristic of different relationship between ASes, Cohen uses BGP route info and IRR data to build an AS topology [3]. Because different data sources always have bias in building AS topology respectively, we will consider how to fuse these three data info to build a more comprehensive AS topology.

78.2 The Structure of AS Topology

According to the AS role in Internet, the AS topology can contain three layers, they are network core layer, network transit layer and terminal customer layer. Network core layer is composed of several top tier ASes. These ASes have large scale and achieve the backbone transit between different continents. Due to the

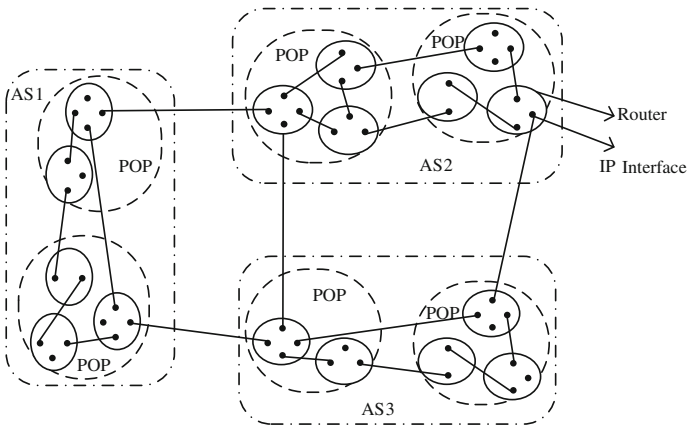


Fig. 78.1 Four levels internet topology sketch map

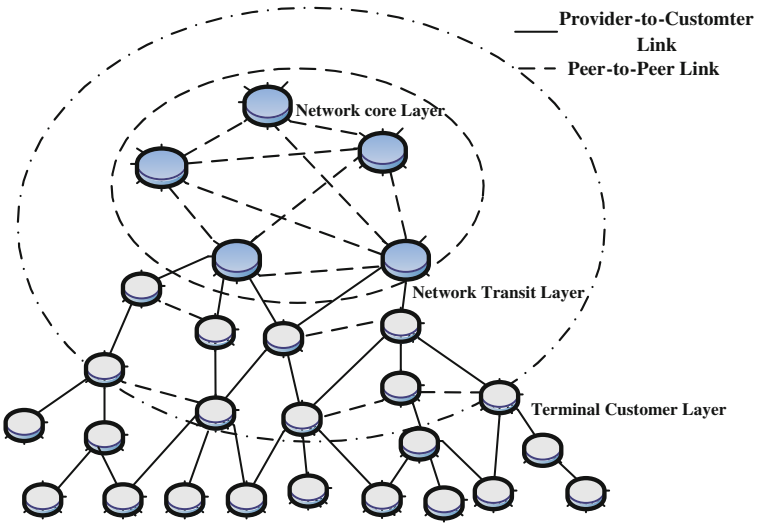


Fig. 78.2 As topology structure sketch map

connectivity is the most important factor between top tier ASes, they often build the peer-to-peer relationship to exchange each data flow. In network transit layer, ASes play a data transit role in some country or area interior. Generally speaking, the large AS provides the connection service for smaller AS, and builds provider-to-customer relationship with each other. So the network transit layer in AS topology displays loose connection state. In terminal customer layer, the count of AS node is far more than other two layers [4], but the structure is simple. These AS nodes often connect one or several provider ASes, and form a tree-like structure. The structure of total AS topology is shown as in Fig. 78.2.

The AS topology changes all the time. For example, some ASes prefer to connect to large ASes for shorten its data transit path, others want to strength its network robust and connect to several provider ASes to form multi-host connection. Furthermore, the appearance of backup links make quite a few AS links just work in network failure. All these factors cause AS topology become more complex.

78.3 AS Topology Building Method

78.3.1 Data Source Analysis

BGP route info contains a series of AS paths to target network. So we can analyze the AS link relationship from it. At present, the most famous project is Route-Views Project [5]. Oregon university has start RouteViews Project since 1999; they set BGP routers to connect with other ASes BGP routers, and receive route

updateinfo by creating BGP sessions. In RouteViews Project website, we can download route real-time info and history info directly.

Internet Routing Registry (IRR) is a distributed route info database by manual maintenance. There are 32 IRR organizations at present [6]. The most famous one in them is RIPE IRR. Due to the IRR data contain some special info which can hardly get from other methods, this type of info is also important to discover AS relationship.

By using traceroute mechanism, we can get the IP route path to some target address. Converting each IP address into AS number, the AS path is indirectly obtained. This info can also use to build AS topology.

78.3.2 The Characteristic Comparison of Three Data Source

Although each data source can use to build the AS topology, they have different characteristics. Firstly, BGP route info root from real BGP routers, this type of info has well validity and veracity, but lack of backup link info between ASes. Due to the limitation of route strategy, a large number of peer-to-peer links cannot be record [4].

Secondly, IRR data comes from manual database, the main problem to use IRR data is that the data validity cannot ensure. However, the IRR data also has some advantages which other data source haven't. IRR data contains newest route info, and the redundancy is low. A large number of peer-to-peer links and backup links are record in IRR data. Besides, the data properties from IRR are the richest in all data source, such as AS Name, Management, Geography position and Contact.

Thirdly, traceroute data need to probe a series of target IP addresses to achieve. This type of data has well controllable, and the IXP info can be found during IP path converting AS path. If one IP can convert into several ASes at same time, the confusion happens. It needs to set some rigorous rules to solve this confusion.

Table 78.1 has summarized these three data sources characteristics.

The way to get BGP route info and traceroute data are belong to active probing method, so they have high validity and local bias. Although the traceroute data obtaining method is hard, it can improve the IXP discovery rate in AS topology. IRR data is the knowledge info published in the Internet, the quality of IRR data is different in different areas, but analyzing these data is useful to discover a large number of peer-to-peer links and backup links. So it has an important role in building AS topology.

78.3.3 AS Topology Building Algorithm

Based on the excellence of BGP route info in node perfectibility and recognizing correctness, our algorithm chooses the BGP route info as original analysis data.

Table 78.1 The Characteristic Comparison of Three Data Source

Data source	BGP route info	IRR data	Traceroute probing data
Data obtaining	Easy	Easy	Hard
Validity	High	Low	High
Node perfectibility	High	Low	Determined by probe range
Link perfectibility	Low	Low	Low
Recognizing correctness	High	Median	Median
Link type bias	High	None	High
Area bias	Median	High	High
IXP discovery rate	Low	Low	Determined by probe range
Backup link discovery rate	Low	High	Low
Other info	None	Several	None

Firstly, use BGP route info to build a rough AS topology. Then add new AS nodes and links analyzed from IRR, and store additional info to describe each AS. At last, discover IXP info from traceroute probing data. We collect IXP list from some public websites [7], and adjust correlative nodes and links in current AS topology. The total algorithm sketch is shown as follows.

There are three main problems in analyzing BGP route info.

- Aggregation problem. Aggregation is one combination method in BGP routers to reduce the count of route. When several routes to one target network have same sub routes, they may aggregate to one route. For example, the aggregated route (2497 1668 10796 {11060, 12262}) contains two independent route. They are (2497 1668 10796 11060) and (2497 1668 10796 12262).
- Fake AS number problem. BGP routers often choose customer routes, peer routes and provider routes in turn according to its local strategy. When the type of route confirmed, the shortest path will be always chosen. The fake AS number can be added to change the route choosing in result. For example, the AS-Path (4513 701 6496) and (4513 8701 11853 6496) are two same type routes to certain target. According to route strategy, the first one will choose. By adding fake AS number 6496 we can make this route become (4513 701 6496 6496 6496), then the second route will choose. In our analysis algorithm, it needs to reduce the redundancy for each AS-Path.
- Private AS number problem. Such as private IP addresses exist, a section of AS numbers are private. The range is from 64512 to 65535. These private ASes are used to partition several areas in one AS inner. So if the private AS number appears in BGP route info, it must be a configure mistake. We just lose this route for simple.

After analyzing BGP route info, the IRR data need to be fused in current results.

- Choose aut-num object in IRR data which is updated in three years. If there are several records for same one aut-num object, just choose the nearest one.
- For each AS node, firstly check if it exist in anterior BGP analysis result. If not, this AS node cannot add in AS topology directly. Because some ASes just

register their route strategy in IRR, but never implement in real network. So it still needs to check correlative aut-nums to validate the symmetrical route strategy if exist. For example, if aut-207 has output strategy to aut-701, then need to check if aut-701 has input strategy to aut-207. When both strategies exist, the aut-num object is valid.

- The valid AS link analysis between two aut-nums follows above validation rule.
- Extract affiliated info. We extract affiliated AS info from aut-num object, such as AS Name, Management, Geography position, Contact and so on.

At last, the traceroute data need to be fused in current AS topology result. We adopt the longest mask matching method to convert each IP address to one network, and search the AS number which does this network belong to. When one IP address can convert to several ASes, we solve these ASes as an AS-Group.

For each AS-Group, firstly, our algorithm checks the IXP's AS number if it exists. Then we analyze the previous AS and the next AS of this AS-Group if connect directly in current AS topology. This AS-Group is replaced by the IXP AS number.

For each non-IXP AS-Group, we need to use current result to check the previous AS and the next AS of this AS-Group if connect each other. Based on Sect. 78.3.2 analysis, let $\mathrm{ASList}_{\mathrm{prev}}$ denotes for the AS list extracted from BGP result by previous AS, $\mathrm{ASList}_{\mathrm{next}}$ denotes for the AS list extracted from BGP result by next AS, then $\mathrm{ASList}_{\mathrm{common}} = \mathrm{ASList}_{\mathrm{prev}} \cap \mathrm{ASList}_{\mathrm{next}}$ denotes for the AS list simultaneity belong to both. Let $\mathrm{ASSet} = \mathrm{ASList}_{\mathrm{common}} \cap \text{AS-Group}$. If $\mathrm{ASSet} \neq \Phi$, it means any AS in ASSet can replace the AS-Group, so we choose a random one to replace the AS-Group. If $\mathrm{ASSet} = \Phi$, we treat this AS-Group as wrong info.

Pass through above steps, a more comprehensive AS topology can be build.

78.4 Experiment

By our independence probing in March 2012, we get the traceroute data. So the BGP route info and IRR data are collected at the same time. The BGP route info comes from Routeviews Project, the IRR data is obtained from 32 IRR database. The AS nodes and links analyzed from these three data source are shown in Table 78.2.

In Table 78.2, nodes and links from BGP take the biggest proportion. The main reason is that our BGP data comes from BGP router which connects to other AS backbone BGP routers; it includes a large number of global routes. Besides, we use BGP data as original analysis data, when analyzing IRR and traceroute data, many nodes and links info have been discovery by BGP. Traceroute data are analyzed from 228860 aggregated global network address probing. Fusing these three type of data, the AS topology contains 30775 AS nodes and 67838 links in total.

Table 78.2 Data analysis result

Data source	Node count	Link count
BGP	29895	61430
IRR	5429	11736
Traceroute	6687	10418
BGP U IRR	30592	65503
BGP U IRR U traceroute	30775	67838

Because the AS topology has large scale and complex characteristics, current researches often compare the network characteristics with others result for validation. We mainly calculate the node degree distribution of our AS topology and compare it with related result. Figure 78.3 displays the node degree CCDF (Complementary Cumulative Distribution Function) curves when fusing three types of data. k denotes node degree, and P_k denotes the complementary cumulative distribution value. In Fig. 78.3, in order to distinguish different result curves, the small picture in left down displays the CCDF curves of BGP and BGPU IRR. We can see that the CCDF curve of pure BGP data has obvious power-law characteristic. When fusing IRR analysis result, the middle of curve will raise a little. The reason is that more peer-to-peer links are found in IRR data, and make the count of middle degree node rapidly increase. When fusing traceroute data, the CCDF curve has little influence. The reason is that BGP data and traceroute data have similar characteristics. Although the traceroute data can find IXP info and change AS topology fractionally, this change cannot influence the CCDF curve.

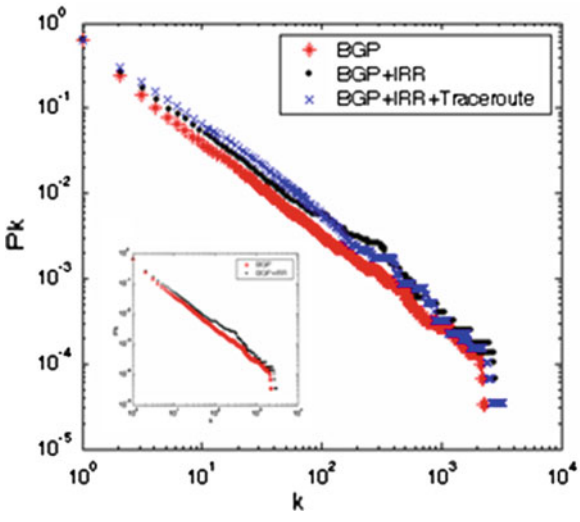


Fig. 78.3 The node degree CCDF of out result

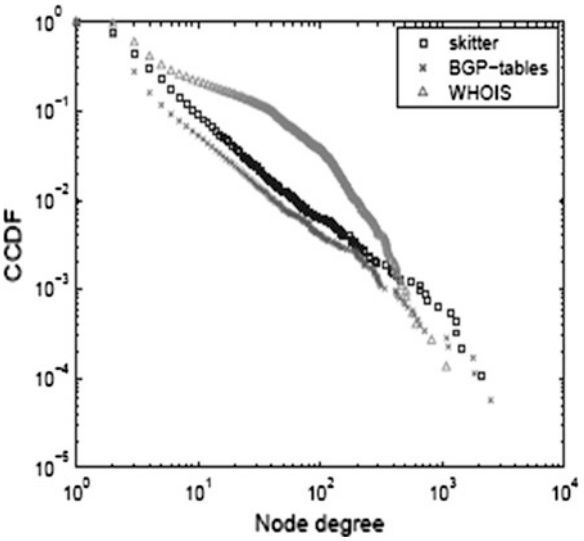


Fig. 78.4 The CCDF of each data source [1]

Figure 78.4 displays CCDF curves of each type of data source independently. Compared with Fig. 78.3, we can see the consistent result. It needs to point out that Fig. 78.4 only use the common info in all three data source. Our work is from the data fusion role, the data scale is larger, and keeps more extra information.

78.5 Conclusion

In summary, this paper analyzes the characteristic of BGP route info, IRR data and traceroute probing data. To solve the localization when using single data source to build AS topology, we put forward a new AS level topology build algorithm based on BGP route info, and fused with IRR data and traceroute data. In experiment, we use real data to execute algorithm and compare the result with related research which proved the validity and correctness of the algorithm.

References

1. Mahadevan, P., Krioukov, D., et al.: The internet AS level topology: three data sources and one definitive metric.ACM SIGCOMM. Comput. Commun. Rev. **36**(1), 16–26 (2006)
2. Oliveira, R., Zhang, B., et al.: Observing the evolution of internet as topology. ACM SIGCOMM **37**(4), 314–324 (2007)
3. Cohen, R., Raz, D.: The internet dark matter—on the missing links in the AS connectivity map. In: Proceedings of the IEEE INFOCOM, Barcelona, Spain (2006)

4. Oliveira, R., Pei, D., Willinger, W., et al.: In search of the elusive ground truth: the internet's AS-level connectivity structure. In: Proceedings of the ACM SIGMETRICS, Annapolis, USA (2008)
5. Routeveiws Project http://www.routeviews.org (2012)
6. List of Routing Registries.http://www.irr.net/docs/list.html (2012)
7. PeeringDB website. http://www.peeringdb.com/ (2012)

Chapter 79
Salinity Time Series Prediction and Forecasting Using Dynamic Neural Networks in the Qiantang River Estuary

Xingguo Yang, Hongjian Zhang and Hongliang Zhou

Abstract The early warning of saltwater intrusion is an important work for ensuring the drinking water supplies. To forecast and predict the daily maximum salinity for the water withdrawn for the waterworks located along the Qiantang River, the nonlinear autoregressive networks with exogenous inputs (NARX) model was applied. Since the multivariate time series of flow, the tide range, the salinities and the water levels observed at 8 gauging stations have great impact on the salt concentration in the river, this will bring in a large number of inputs when these variables directly fed into the NARX model and add unnecessary model complexity and poor performance. Therefore, the dynamic principal component analysis (DPCA) was used to reduce the data redundancy. Simulation predicted results show that the NARX model using DPCA can predict salinity in the river accurately, moreover, this method not only reduces the input dimension and over-fit the equation, but also enhances the model performance and the generalization ability considerably.

Keywords Time series · NARX model · DPCA · Prediction · Saltwater intrusion

79.1 Introduction

Water managers require a predictive model for the management of estuarine water resources as functions of geometry, freshwater flow and tide [1]. In many rivers, water is withdrawn for irrigation and drinking purposes, and if it is contaminated by salt from the sea, it is no longer useable [2]. The recommended threshold value of salinity for drinking water is less than 250 mg/L, namely, 0.45 ‰.

X. Yang · H. Zhang (✉) · H. Zhou
State Key Laboratory of Industrial Control Technology, Department of Control Science and Engineering, Zhejiang University, Hangzhou, China
e-mail: hjzhang@iipc.zju.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_79,

Although sophisticated three-dimensional numerical models are available to provide detailed solutions to critical environmental impact problems, a large amount of field data and much effort are often needed to calibrate and verify these numerical models for engineering applications. Therefore, in preliminary engineering studies, simple one-dimensional or time series models are quite useful for engineers to make rapid preliminary estimates of salinity change that may result from the modification of the tidal river [3]. For the strong tide, the saltwater and the fresh water mix well in Qiantang River, one dimension time series model can be used to predict the saltwater intrusion.

79.2 Site and Data Sets

The Qiantang River is one of the largest rivers in China, it springs from the borders of Anhui and Jiangxi provinces and passes through Hangzhou, the capital of Zhejiang province, before flowing into the East China Sea passes through Hangzhou Bay as shown in Fig. 79.1.

Hangzhou, is located downstream of the Qiantang River which is a typical tidal river. Currently, domestic, irrigation and industrial water supplies of Hangzhou are all mostly withdrawn from the Qiantang River, and will be subjected to the effect of saltwater intrusion during the dry season and spring tide every year.

The length of saltwater intrusion and the saline concentration in the Qiantang River estuarine area primarily depend on the upstream discharge and downstream tide. When the low water discharge increases, the saltwater intrusion will be suppressed. When the tidal range increases, the saltwater intrusion will be more serious [4]. Other factors which impact on the salinity include drainage pumping,

Fig. 79.1 Qiantang River estuarine basin and Hangzhou bay

Table 79.1 Available data

Location	Data type (daily)			
Fuchun River hydropower station	Discharge			
Tonglu				Water level
Fuyang				Water level
Wenjiayan				Water level
Zhakou			Salinity	Water level
Qibao		Tide range	Salinity	Water level
Cangqian		Tide range	Salinity	
Zhapu		Tide range	Salinity	

upstream and downstream salinities, the water level, and large-scale human activities [4, 5].

By forecasting the daily maximum salinity at Cangqian several days in advance, the saltwater intrusion early warning can be made for the water intakes located along the river. According to the Hangzhou pumping schedule, 5 days ahead forecasting period is chosen in this study. The factors affect the saltwater intrusion are shown in Table 79.1. Daily mean water levels, discharge, tide range and daily maximum salinities for the period 01-01-2006 to 31-12-2008 were collected from the 8 gauging stations along Qiantang River. Locations of these gauging stations can be seen in Fig. 79.1.

79.3 Artificial Neural Networks Model Development

Nonlinear autoregressive model with exogenous inputs (NARX) belongs to the class of recurrent dynamic networks which has been demonstrated that they are well suited for modeling nonlinear systems and specially time series. NARX has many advantages such as more effectively learning, faster converging and better generalize than other networks [6]. In the light of these characters, NARX neural networks are very suitable for forecasting and prediction the saltwater intrusion for the water intakes located along Qiantang River.

79.3.1 Data Division and Transformation

Regularization technique is used as a stopping criterion for NARX networks training, the available data was divided into training and test data sets. Correspondingly, the data (shown in Table 79.1) were used to create the training and test data sets. The 5 days ahead prediction results in a total of 1091 data records, from which 890 records (81 %) were used for training and 201 records (19 %) for testing. In this technique, the training set is used for computing the gradient and

updating the network weights and biases, while the test data are used to compare different models, and provide an independent measure of the network's performance after training.

In order to ensure that all variables receive equal attention during the training process and speed up the network convergence, the data of inputs and outputs are linearly transformed by using the original data range to rescale the series to a range that is commensurate with the output transfer function. The ranges used in conjunction with the hyperbolic tangent output transfer function were [−0.8 0.8] for the network inputs and outputs. The test data were normalized using the same methodology.

79.3.2 Determinatin of the Model Inputs

When the thirteen parameters which impact the salinity at Cangqian gauging station (shown in Table 79.1) directly fed into the neural network, it will make a complex network architecture and the training inefficiently, because of the strong correlation between these inputs and autocorrelation with the past. Therefore, the dynamic principal component analysis (DPCA) is applied to extract the principal components of the input variables to reduce the input dimension: the procedure is shown in Fig. 79.2.

DPCA is an expansion of PCA, and it through the 'time lag shift' method to include dynamic behavior in the PCA model [7]. Usually, DPCA is calculated by adding *l*time lag observations to obtain the input variables augmented matrix, X_A:

$$X_A(l) = [X(t) \quad X(t-1) \quad \cdots \quad X(t-l)] = \begin{bmatrix} x_t^T & x_{t-1}^T & \cdots & x_{t-l}^T \\ x_{t-1}^T & x_{t-2}^T & \cdots & x_{t-l-1}^T \\ \vdots & \vdots & & \vdots \\ x_{t-m}^T & x_{t-m-1}^T & \cdots & x_{t-l-m}^T \end{bmatrix} \tag{79.1}$$

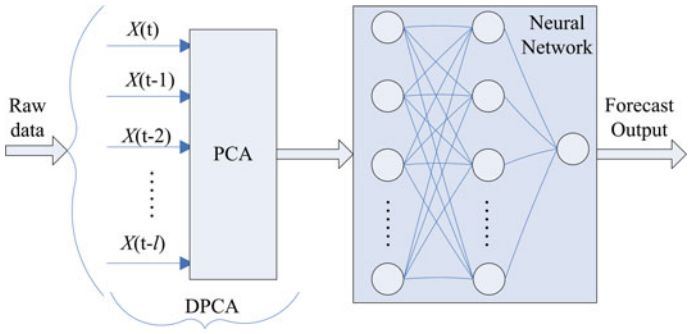


Fig. 79.2 The procedure of modeling the salinity time series

where $X(t)\quad X(t-1)\quad \cdots\quad X(t-l)$ are the time series affect the salinity intrusion, $x_t^T\quad x_{t-1}^T\quad \cdots\quad x_{t-m}^T$ is the observed data. The method to calculate l which indicates the order of the dynamic system as follows: first, calculate the static relation $l=0$, then calculate new relationship according the following equation:

$$r_{new}(l) = r(l) - \sum_{i=0}^{l-1}(l-i+1)r_{new}(i) \tag{79.2}$$

Until $r_{new}(l) \leq 0$, there is no static and dynamic relationship. Then, standardization the augmented matrix X_A to $\tilde{X}_A$, next, singular value decomposition the matrix $\tilde{X}_A$ according to the traditional PCA:

$$\tilde{X}_A = U\Sigma V^T \tag{79.3}$$

Only the first s eigenvectors are retained, while (k–s) smaller components are discarded, assuming that the latter describe mostly noise, the number corresponding for the principal component is s. Reduced the dimension of transformation matrix $V_{c\times c}$ for the $\tilde{V}_{c\times s}$, the principal component of the system is:

$$\tilde{Z} \approx \tilde{X}_A \tilde{V}_{m\times s} \tag{79.4}$$

where, k is the number of eigenvectors, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_s, \cdots \geq \lambda_k$ are singular values of the matrix X, and $U \in F^{r\times r}, \sum \in F^{r\times c}, V \in F^{c\times c}$.

Use Equation (5) to estimate the energy contribution η_j of a principal component. Retain the corresponding principal components if $\eta_j > \eta_0$, where η_0 is a constant, $0 < \eta_0 < 1$.

$$\eta_j = \lambda_j^2 / \sum_{i=1}^{k} \lambda_i^2, \quad j = 1, 2, \cdots, k \tag{79.5}$$

79.3.3 NARX-Model

Compared to Elman networks, NARX networks can approximate any nonlinear dynamical system. In general, the architecture and leaning algorithm are more complicated than static networks, but the capability described the nonlinear dynamical system is greatly enhanced [6]. Figure 79.3 depicts the NARX architecture.

The general prediction equations for computing the next value of the time series $y(t+1)$, the past observation $u(t), u(t-1), \cdots, u(t-m)$ and the past outputs $y(t), y(t-1), \cdots, y(t-n)$ as inputs, mathematically it can be described as follows:

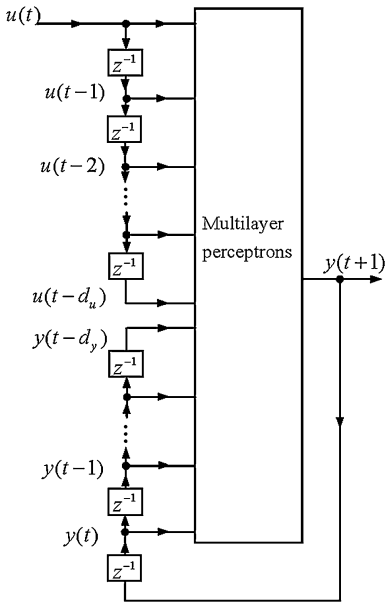


Fig. 79.3 NARX model with exogenous inputs [6]

$$y(t+1) = F(y(t), y(t-1), \cdots, y(t-d_y), u(t), u(t-1), \cdots, u(t-d_u)) \quad (79.6)$$

where the next value of the dependent output signal $y(t)$ is regressed from previous values of the output signal and previous values of an independent (exogenous) input signal. $F(\cdot)$ is a nonlinear function, d_u and d_y are the maximum lags in the input and output.

The network training function is a Bayesian regularization back propagation which updates the weight and bias values according to Levenberg–Marquardt optimization, and minimizes a linear combination of squared errors and weights, to produce a network that is well generalized. By adding a term that consists of the mean of the sum of squares of the network weights and biases (MSW) the performance is used to evaluate the prediction performance and compare it with the results of the models, described as follows:

$$MSEreg = \gamma MSE + (1-\gamma)MSW \quad (79.7)$$

where γ is the performance ratio, MSE is the mean sum of squares of the network error. MSE and $MSW$are expressed as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(e_i)^2 = \frac{1}{N}\sum_{i=1}^{N}(t_i - y_i)^2 \quad (79.8)$$

$$MSW = \frac{1}{n}\sum_{i=1}^{n}\omega_j^2 \quad (79.9)$$

Using this performance function causes the network to have smaller weights and biases, which forces the network response to be smoother and less likely to overfit the equation.

Another important work is to choose the initial weights and biases of the neural networks. This has a great impact on the network's convergence and local minimum. In this article, the initial weights and biases are selected using the Nguyen-Widrow initialization algorithm which chooses values that distribute the active region of each neuron in the layer approximately evenly across the layer's input space [8].

79.4 Results and Discussion

All the approaches mentioned above to model saltwater intrusion in the Qiantang River were implemented using the MATLAB 7.11.0 software. Unless stated otherwise, the default software parameters were used.

In this article, $l = 3$ was determined by the DPCA procedure Equation (2). And 12 principal components are finally chosen at the cumulative contribution rate $\eta = 0.96$, namely, the cumulative contribution rate of the first 12 principal components for the variance of the sample matrix is greater than 96 %.

The NARX network structure is configured by trial and error as follows: input delays $d_u = 2$ and output delays $d_y = 2$, the hidden layer neurons $N = 8$. The activation function was the sigmoid function and the performance ratio was set to $\gamma = 0.5$, which gives equal weight to the mean square errors and the mean square weights in this study. Under this configuration, using the data processed by DPCA, the NARX neural network trained ten times when the algorithm has converged, the mean test performance is about 0.5, the mean effective number of connections is about 180, and the mean training time is about 50 s respectively.

While using raw data without being processed by DPCA, the NARX network trained ten times when the algorithm has converged to reach the mean performance about 0.5, the mean effective number of connections is about 340, and the mean training time is about 80 s. Correspondingly, input delays $d_u = 3$, output delays $d_y = 3$, the hidden layer neurons $N = 10$, and the activation function and the performance ratio were same as the NARX model using DPCA. Obviously, DPCA not only reduce the data redundancy and extract the main features from the nonlinear dynamic system but also simplify the model structure. The comparison of NARX model and NARX model using DPCA are presented in Table 79.2.

Table 79.2 Comparison of NARX model and NARX model using DPCA

Method	Architecture	Performance	Number of effective parameters	Training time (s)
NARX	3-3-10	0.5	340	80
DPCA + NARX	2-2-9	0.5	180	50

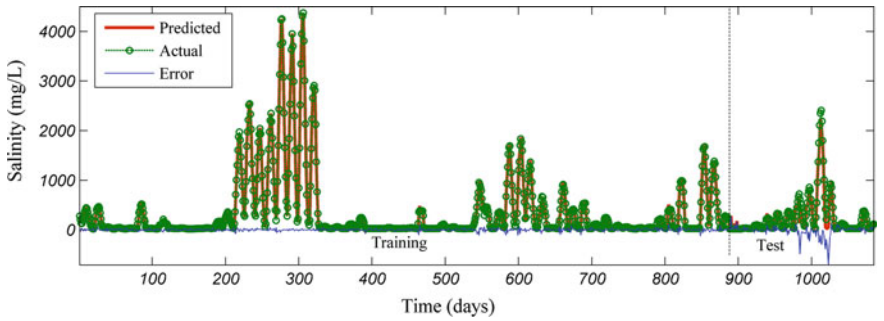


Fig. 79.4 Observed and the corresponding predicted salinities at Cangqian gauging station for 5-day ahead by NARX model using DPCA

The NARX network adopted a one-step-ahead prediction. The actual, the predicted and the errors between the predicted and the actual of the model using DPCA and NARX are plotted in Fig. 79.4.

As seen from the curves figure, the characters of salinity time series at Cangqian are nonlinear and non-stationary, moreover, the curve fluctuate greatly and wave crests are very sharp. Generally, for this characteristics of the time series, traditional modeling methods are difficult to capture the sharply crests well. Nevertheless, the correlation coefficient R between the outputs and targets of the NARX neural network plotted in Fig. 79.5. The perfect correlation between the targets and outputs means the NARX model outputs can fits the nonlinear observed well from Fig. 79.5.

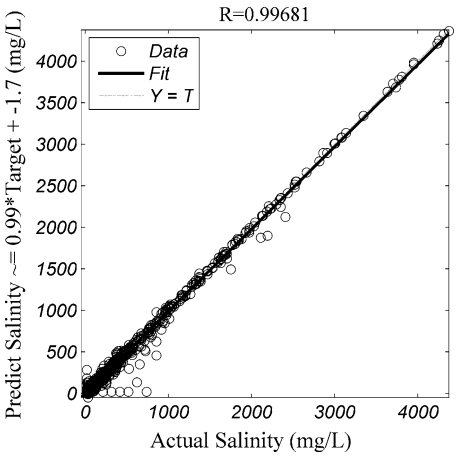


Fig. 79.5 The correlation coefficient R between the observed and predicted

79.5 Conclusion

In this article, for ensuring water drinking safety, the nonlinear salinity time series at Cangqian gauging station is modeled and predicted using NARX dynamic neural networks. One conclusion is that from the correlation coefficient R for the observed and predicted is close to 1, shows that NARX networks model has the capabilities of approximating the dynamic nonlinear system and is suitable for modeling the salinity time series at Cangqian gauging station located along the Qiantang River. Another conclusion is that DPCA is able to reduce the inputs dimension, and this can effectively cut down the neural network connections and the training time, and create a parsimonious model.

References

1. Savenije, H.H.G.: Predictive model for salt intrusion in estuaries. J. Hydrol. **148**(1–4), 203–218 (1993)
2. Aerts, J.C.J.H., et al.: Using GIS tools and rapid assessment techniques for determining salt intrusion: STREAM, a river basin management instrument. Phys. Chem. Earth **25**(3), 265–273 (2000)
3. Huang, W., Foo, S.: Neural network modeling of salinity variation in Apalachicola River. Water Res. **36**, 356–362 (2002)
4. Han, Z., et al.: Effect of large-scale reservoir and river regulation/reclamation on saltwater intrusion in Qiantang Estuary. Sci. China Ser. B. **44** (suppl): 221–229 (2001)
5. Lu, X.: Model experiment on saltwater intrusion of the Qiantang river estuary. J. Hydro-Sci. Eng. **3**(4), 403–410 (1991) (In Chinese)
6. Diaconescu, E.: The use of NARX neural network to predict chaotic time series. WSEAS Trans. Comput. Res. **3**, 182–191 (2008)
7. Ku, W., et al.: Disturbance detection and isolation by dynamic principal component analysis. Chemometr. Intell. Lab **30**, 179–196 (1995)
8. Nguyen, D., Widrow, B.: Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In: Proceedings of the International Joint Conference on Neural Networks, vol. 2, pp. 21–26 (1990)

Chapter 80
An Adaptive Packet Loss Recovery Method for Peer-to-Peer Video Streaming Over Wireless Mesh Network

Hamid Reza Ghaeini, Behzad Akbari and Behrang Barekatain

Abstract P2P video streaming over WMN includes different multimedia applications such as IPTV, video surveillance and video conferencing. It also introduces some challenges such as required level of QoS. Packet loss recovery methods can improve the experienced amount of QoS which leads to better video quality on peers. Although ARQ and FEC methods have been used in many video streaming applications, they are unable to provide enough level of QoS in P2P video streaming over WMN. Hybrid methods improve the performances of packet loss recovery schemes. But they do not carefully consider the characteristics of the source and the destination nodes, thus are not suitable for P2P video streaming over WMN. Therefore, in this study, an adaptive packet loss recovery method is proposed to select the loss recovery policy according to the source and the destination characteristics and loss probability of communication.

Keywords Packet loss recovery · Video streaming · QoS · P2P · WMN

H. R. Ghaeini
Multimedia Networking Lab, Research Institute of ICT (ITRC), Tarbiat Modares University, Jalal Ale Ahmad Highway, 14115-111 Tehran, Iran
e-mail: h.qaienee@modares.ac.ir

B. Akbari (✉)
Tarbiat Modares University, Jalal Ale Ahmad Highway, 14115-111 Tehran, Iran
e-mail: b.akbari@modares.ac.ir

B. Barekatain
Universiti Teknologi Malaysia, 81300 Johor Bahru, Malaysia
e-mail: Bbehrang3@Live.utm.my

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_80,

Acronyms

ARQ	Automatic Repeat reQuest
BMS	Buffer Map Status
FEC	Forward Error Correction
MANET	Mobile Ad-hoc NETwork
P2P	Peer-to-Peer
QoS	Quality-of-Service
QoE	Quality-of-Experience
WMN	Wireless Mesh Network

80.1 Introduction

WMN is an emerging communication network for seamlessly Internet access over the Internet. In WMN, each sent packet can be delivered at the destination node in a multi-hop manner according to the employed path selection routing protocol [8]. Self-healing, self-configuration and scalability are three important benefits of using WMN. On the other hand, low transmission coverage in most of the wireless mesh nodes such as laptops, tablets and mobile phones [5] and node mobility are two well-known drawbacks of them. Each node can either use nearby node or wireless mesh router for communicating to other nodes using multi-hop technique. This technique lets the network be more scalable and robustness, especially in peer churning. A WMN is a special type of MANET; however, some important differences between MANET and WMN is that wireless mesh networks consist of stable backbone, large coverage area and high power nodes i.e. wireless mesh routers [8]. In order to route data among existing nodes in WMN, there are three types of path selection algorithms including reactive, proactive and hybrid routing protocols [3, 10].

Recently, P2P systems have been used in many video streaming applications. A P2P system is a distributed system so that clients directly communicate with each other and there is no specific infrastructure [4]. Each peer has both the rules of a client and a server simultaneously. P2P networks can be setup over LAN, WAN or the Internet. Each peer needs a specific or compatible software for participating in P2P overlay [19]. One of the most interesting applications of P2P networking is multimedia communication. Nowadays, P2P video and audio conferencing can be adopted by P2P platforms using special applications such as Skype in order to provide better performances in conferencing [11, 16]. There are many P2P structures for P2P content sharing application [5]. P2P systems can be divided into three categories including structured, unstructured and hybrid systems [4]. Moreover, based on the employed topology, P2P systems can be implemented as mesh or tree structures [11].

Mesh and root are the two most important architectures for P2P video streaming [19]. The root architecture is suitable for live video streaming. However, the mesh-based architecture performs better in disruptive networks like wireless mesh

networks [11]. In mesh-based P2P architecture, a central server named Tracker keeps the statuses of all peers and their neighbours. If one peer wants to join the mesh, it first asks the tracker for neighbours' list and, then, sends joining messages to some of them randomly. Figure 80.1 describes a sample mesh topology in P2P video streaming over WMN which contain 2 central server including video source and tracker that have wired communication to mesh access point. Also there are 4 mesh access points which connect to each other with a physical wireless channel. Each wireless mesh node may have many overlay neighbours which represented by a dotted line.

In this paper different packet loss recovery methods for P2P video streaming over large scale disruptive networks such as wireless mesh networks are evaluated. In addition, an adaptive packet loss recovery method for P2P video streaming over wireless mesh networks will be proposed. Results show that the loss recovery ratio of this method is really considerable in comparison with other approaches. In other words, this method reduces end-to-end delay in video streaming. Therefore, live video streaming can be adopted in large scale networks.

80.2 Proposed Method

In disruptive networks like mobile networks, the loss probability is high [6]. In P2P video streaming, the effects of loss can be propagated in the whole overlay which leads to low video quality on receivers [1]. In order to cope with this problem, packet loss recovery is a conventional method in video streaming [13].

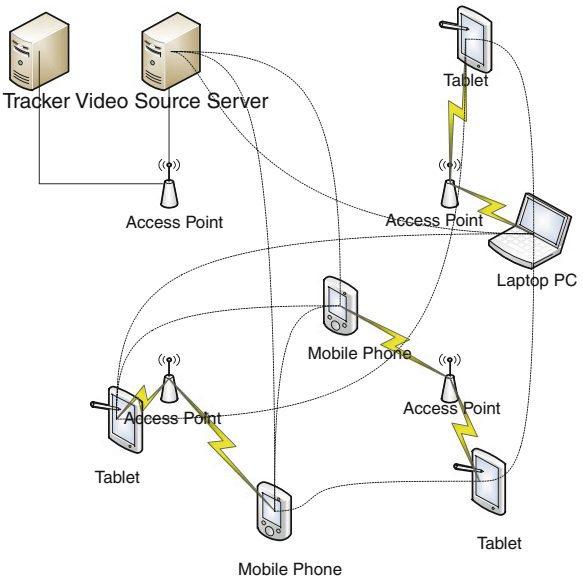


Fig. 80.1 Mesh topology in P2P video streaming over WMN

There are three types of packet loss recovery methods in video streaming including ARQ [15], FEC [17] and hybrid ARQ [2]. In adaptive packet loss recovery method, the redundancy of FEC codes can be computed and generated according to the packet loss ratio between source and destination nodes before packet transmission. As depicted in Eq. 80.1, the redundancy of FEC parity codes for maintaining a residual loss probability not more than p_{max} is [9]:

$$\begin{aligned} R_{FEC} &= \min\{R|\varepsilon \le p_{\max}\} \\ \varepsilon &= \sum_{k=R+1}^{D+R} \binom{D+R}{k} p^k (1-p)^{R+D-k} \frac{k}{D+R} \end{aligned} \tag{80.1}$$

where D is number of data packets, R is number of additional redundant packet, ε is the upper bound of residual loss probability and p is the probability of loss. Then, a suitable loss recovery policy will be adopted for packet protection against loss based on the performance of that policy in loss recovery between source and destination nodes. Figure 80.2 show the loss recovery algorithm in each frame sending process.

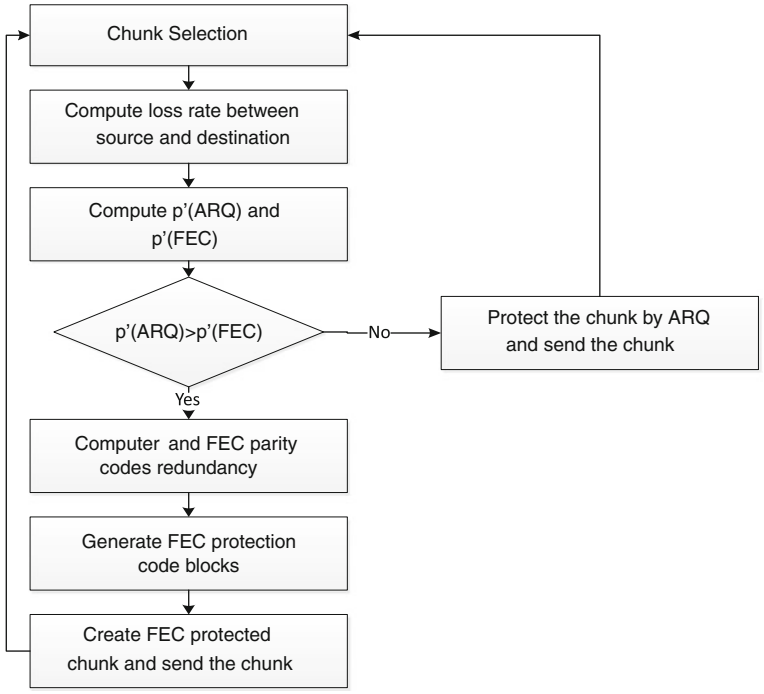


Fig. 80.2 Packet loss recovery selection algorithm

80.3 Problem Statements

In computer network, simulation is the common part of network research and design. In this technique, network behavior will be modeled by simulator software and the performance of the network will be evaluated. OMNeT++ [14], a popular tool, is a discrete event-based simulator for communication networks which includes several simulation frameworks. In this research, the OverSim and the INETMANET frameworks have been used for P2P video streaming over WMN simulation. The INETMANET framework is an extension of the INET package which is specifically designed for wireless networking. OverSim package is an overlay and P2P simulator and contains several solutions for structured and unstructured overlay networks. Moreover, different network performance metrics have been evaluated including end-to-end delay, video distortion, dependency loss and start playing time.

End-to-End delay is the required time for transferring a video packet from the source to the destination node in a multi-hop manner. In this manner, a video packet may forward through wireless overlay nodes or wireless mesh routers. End-to-end delay is one of the most important parameters in live video streaming [12]. Each communication protocol in P2P video streaming should mitigate an upper bound of this metric in order to provide high video quality on peers.

Distortion is the amount of video packet loss that a node experience due to network errors or interdependency among video frames. Dependency loss refers to the lost video frames due to existing dependency among video frames in a GoP [7, 18]. In other words, dependency loss is the percentage of lost video frames due to the loss of the base frames i.e. I or P frames. The amount of dependency loss is between 0 and 1. In video frame protections, this parameter shows the efficiency of the GoP based frame protection protocol against the loss due to inability of decoding the received video frames. As soon as a peer finishes its initial buffer stage, it can start the playback of video frames immediately. This time can be called start playback time. The start playback time is the average of time that takes for receiving and buffering enough video frames as well as decoding them for

Table 80.1 Conditions of P2P video streaming over wireless mesh network simulation

Variable	Value
Simulation time	600 s
Video Trace File, Fps, Codec	Silence of the Lambs, 25, MPEG4
Distribution model	Random
Entrance time interval	Uniform(1, 3)
Packet size, MTU	100 Kb, 7891 bytes
Propagation model, P_{MAX}	Path Loss Reception Model, 0.001
Peer Video Buffer, GoP (N_P, N_{PB})	100 s, (3, 2)
Peer neighbors in overlay	Random (3, 6)
MANET routing, Overlay topology	Reactive, Mesh topology-Pull
Node mobility type, speed	Pedestrian, 1.5 mps

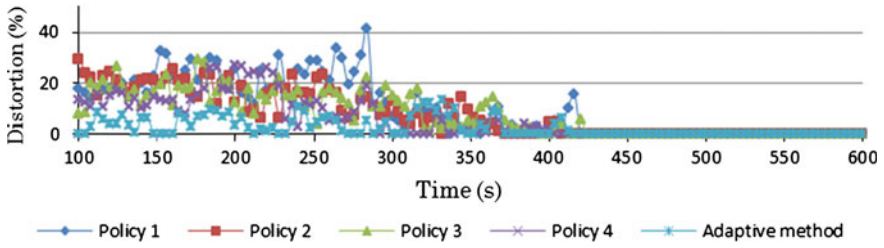


Fig. 80.3 Packet loss in different packet loss recovery methods against simulation time

Fig. 80.4 Packet loss in different packet loss recovery methods against network size

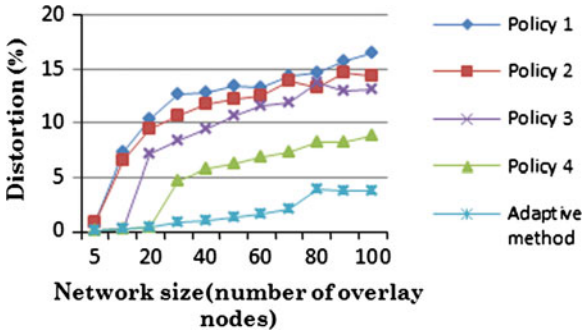
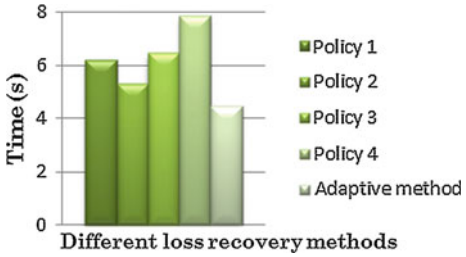


Fig. 80.5 End-to-end delay in different packet loss recovery methods



starting the video playback. This parameter indicates the efficiency of a packet loss recovery method in P2P video streaming. The simulation conditions for performance evaluation of adaptive packet loss recovery method in P2P video streaming over wireless mesh networks are depicted in Table 80.1.

80.4 Performance Evaluation

Here, five different packet loss recovery policies have been adopted in order to evaluate and compare different loss recovery methods in P2P video streaming over WMN. These methods are as follows:

1. ARQ: protecting the video chunks using simple ARQ method.

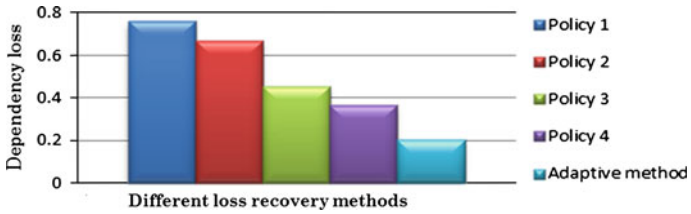


Fig. 80.6 Dependency loss in different packet loss recovery methods

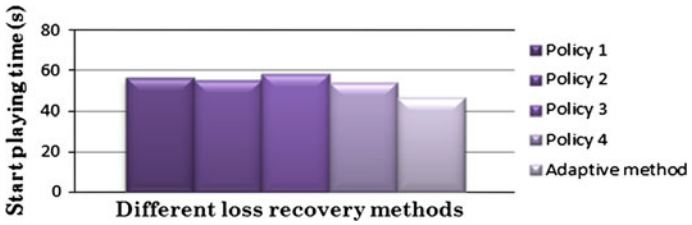


Fig. 80.7 Start playing time in different packet loss recovery methods

2. Unequal Importance Hybrid ARQ II: protecting the video chunks that contain I frames using Hybrid ARQ and other video chunks using the simple ARQ method.
3. Unequal Importance Hybrid ARQ II: protecting the video chunks that contain I and P frames using Hybrid ARQ and other video chunks using the simple ARQ method.
4. Hybrid ARQ II: protecting the video chunks using Hybrid ARQ method.
5. Proposed method: protecting the video chunks using Adaptive loss recovery method.

In proposed adaptive loss recovery method, each node selects different loss recovery approaches for each video chunk based on the probability estimation of loss (p') between source and destination. If FEC protection is selected, FEC parity codes will be generated according to P_{max} and size of the chunk. Figure 80.3 shows the amounts of distortions of different loss recovery methods in P2P video streaming over wireless mesh network across the simulation time.

Moreover, the proposed method is scalable by the size of network which works fine in high density loss situations. Figure 80.4 shows the amounts of distortions in different loss recovery methods across the overlay size.

As seen in Figs. 80.4 and 80.5, the proposed method performs very well in both mitigating the distortion and the end-to-end delay which are the two most important parameters in live video streaming. Moreover Fig. 80.6 compares the efficiency of protection methods in protection of important frames against loss. As seen in this picture the proposed method works very well in frame protection.

As can be seen in Fig. 80.7, the proposed method also works fine in P2P live video streaming over wireless mesh networks.

80.5 Conclusion

Based on the obtained results, the performance of the proposed adaptive packet loss recovery method is considerable. The main advantage of the method is its ability of decreasing end-to-end delay while increasing the QoE in peer to peer live video streaming over error prone networks like wireless mesh networks. This study showed that adaptive packet loss recovery methods can be adopted in error prone networks and overlays with high churning. Moreover, estimation of packet loss between source and destination nodes can improve the overall loss recovery performance and mitigate end-to-end delay and distortion.

References

1. Akbari, B., Rabiee, H.R., Ghanbari, M.: Packet loss recovery schemes for peer-to-peer video streaming. International Conference on Network Services, Athens, Greece (2007)
2. Akbari, B., Rabiee, H.R., Ghanbari, M.: Packet loss in peer-to-peer video streaming over the Internet. Multimedia Syst. **13**, 345–361 (2008)
3. Alotaibi, E., Mukherjee, B.: A survey on routing algorithms for wireless Ad-Hoc and mesh networks. Comput. Netw. **56**, 940–965 (2011)
4. Androutsellis Theotokis, S., Spinellis, D.: A survey of peer-to-peer content distribution technologies. J. ACM Comput. Surv. **36**, 335–371 (2004)
5. Buford, J.F., Yu, H., Lua, E.K.: P2P Networking and Applications. Morgan Kaufmann, Boston (2008)
6. Crow, B.P., Widjaja, I., Kim, J.G., Sakai, P.T.: IEEE 802.11 Wireless local area networks. IEEE Commun. Mag. **35**(9), 116–126 (1997)
7. Ghanbari, M.: Standard Codecs, Image compression to advanced video coding (3rd ed.) The Institution of Engineering and Technology, UK (2011)
8. Hiertz, G., Denteneer, D., Max, S., Taori, R., Cardona, J., Berlemann, L., Walke, B.: IEEE 802.11s: the WLAN mesh standard. IEEE Wirel. Commun. **17**, 104–111 (2010)
9. Khalil, I., Weippl, E.: Innovations in Mobile Multimedia Communications and Applications. New Technologies in Information Science Reference (IGI Global), pp. 175–200 (2011)
10. Lindeberg, M., Plagemann, T., Kristiansen, S., Goebel, V.: Challenges and techniques for video streaming over mobile ad hoc networks. Multimedia Syst. **17**, 51–82 (2011)
11. Liu, Y., Guo, Y., Liang, C.: A survey on peer-to-peer video streaming systems. Peer-to-Peer Netw. Appl. **1**, 18–28 (2008)
12. Lou, X., Hwang, K.: Quality of data delivery in peer-to-peer video streaming. ACM Trans. Multimedia Comput. Commun. Appl. **8**, 1–23 (2012)
13. Moltchanov, D.: Service quality in P2P streaming systems. Comput. Sci. Rev. **5**, 319–340 (2011)
14. OMNeT++. Available via DIALOG. http://www.OMNETPP.org/ (2012)
15. Peltotalo, J., Harju, J., Väätämöinen, L., Bouazizi, I., Curcio, I.D., Gassel, J.V.: Scalable packet loss recovery for mobile P2P streaming. Lect. Notes Comput. Sci. **60**, 107–120 (2010)
16. Rossi, D., Mellia, M., Meo, M.: Understanding skype signaling. Comput. Netw. **53**(2), 130–140 (2009)
17. Wehbe, H., Babonneau, G., Cousin, B.: Fast packet recovery for PULL-based P2P live streaming systems. The Second International Conference on Advances in P2P Systems, pp. 20–25 (2010)

18. Wu. H., Claypool, M., Kinicki, R.: Guidelines for selecting practical MPEG group of pictures. IASTED International Conference on Internet and Multimedia Systems and Applications (EuroIMSA), pp. 61–66, Innsbruck, Austria (2006)
19. Zhang, X., Hassanein, H.: A survey of peer-to-peer live video streaming schemes—an algorithmic perspective. Comput. Netw. **56**(15), 18–28 (2012)

Chapter 81
Measurement of Deformed Surface and Key Data Processing Based on Reverse Engineering

Yongjian Zhu, Jingxin Na and Shijie Wei

Abstract After the metal sheet forming, the strain change should be known through some ways in order to conduct the stress analysis of forming process. To achieve this goal, some special circle marks are imprinted on the sheet surface, and after forming, the deformed circle marks should be measured. The previous method uses a Vernier Caliper to manually measure the stretched rope, which is highly laborious and inaccurate. To overcome this disadvantage, a non-contacting method is adopted to achieve all the coordinates' data based on the structure-light principle. Then the achieved coordinates' data of spatial surface are fitted to be a CAD model through reverse engineering (RE). In RE, the CAD surface is unfolded to be a plane or two orthogonal lines are projected on the CAD surface, so the deformed circle marks—spatial ellipses could be measured in size completely at one time. The experiments prove that the measured accuracy is about 0.01 mm, higher than the previous manual one.

Keywords Deformed surface · 3D measurement · Reverse engineering

81.1 Introduction

Reverse engineering (RE) implies reproducing physical objects by directly extracting geometric information on the objects [1]. Traditionally, the product is manufactured according to the concept design and drawing. Oppositely, RE starts

Y. Zhu (✉) · J. Na
State Key Laboratory of Automotive Simulation and Control, Jilin University, Changchun, China
e-mail: zhuyongjian_hn@126.com

Y. Zhu · S. Wei
Zhejiang University of Science and Technology, Hangzhou, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_81,

from the digitization of prototype shape, and then reconstructs the CAD model. At last, manufactures the final products [2]. This method can turn the real object to be a CAD model and achieve the 3D size information for further data process. RE has become a key element in design and production, to fulfill today's needs of reducing time-to-market [3]. Many types of digitizing techniques have been proposed to realize RE including probe-based contacting method and optical non-contacting method. Among them, probe-contacting method mainly operates on coordinate measuring machines equipped with ball-tip probes producing continuous or touch-trigger signals [4, 5]. It has good measurement accuracy but is time-consuming. The non-contacting method is mainly based on laser-scanning or structure-light principle [6, 7]. This method features both high speed and high accuracy and meets the precision requirements of RE to some degree. In this paper, an advanced measuring instrument—Shining3D Scanner is used to measure the real stamping metal sheet in order to conduct the stain-stress analysis. This instrument is based on time-series structure-light principle and has a shape accuracy of about 0.01 mm. It aims at the 3D precision measurement of deformed circle marks after punch stamping.

As for the deformed circle marks, previously an inelastic rope is used to be stuck to the deformed marks at first and then stretched manually; afterwards, a Vernier Caliper is adopted to check the rope length. This is a highly laborious and impractical job. To solve this problem, Shining3D Scanner is adopted to digitize the deformed metal sheet, and then the digitized surface of metal sheet is fitted to be a CAD model (NURBS surfaces) in RE software or CAD software (UG NX). Finally the CAD surface could be unfolded to be a plane or two orthogonal lines be projected on the CAD surface, in this case, the deformed circle marks—spatial ellipses could be measured in size completely at one time. The experiments present the real measurement results. In this paper, Sect. 81.2 presents the digitization of real punched metal sheet, and Sect. 81.3 describes the measurement results, at last, Sect. 81.4 gives the conclusion.

81.2 Digitization of Real Object

81.2.1 Acquisition of Point Cloud and Mesh Generation

In the FEM simulation and experimental tests of metal sheet forming, the strain should be measured in order to analyze the forming defects and find the weak spot on sheets. However, strain couldn't be measured directly. For solving this problem, the shape deformation or displacement of metal sheet should be achieved firstly. Through displacement analysis, the strain could be obtained. In this process, the key step of measurement is to digitize the real object through a 3D scanning instrument. Hence a 3D scanner from Shining3D Tech. Co. is used to achieve the digitized data of stamping metal sheet. Figure 81.1 is the real object of

Fig. 81.1 The real object of deformed metal sheet

metal sheet with imprinted circle marks. It is 231 mm × 233 mm × 46 mm. The used 3D scanner—OpticScan series is based on time-series fringe projection with binocular vision, and the single-frame measurement area is 200 mm × 150 mm. Through automatic stitching, the measurement accuracy could reach about 0.01 mm. Figure 81.2 shows the picture of one of 3D OpticScan series.

To conduct the stress–strain analysis, the metal sheet is punched to be box-shaped. On the surface of box, the deformation of some circle marks should be known through RE method. Figure 81.3a shows the front face with sampled 4 circle marks (A, B, G and H). Figure 81.3b is the back face with 4 sampled circle marks (C, D, E and F). After punching, the circle marks on A-H are transformed into ellipse ones. The measuring goal is to obtain the length of major and minor axes from the deformed marks.

Fig. 81.2 The 3D scanner 231 mm × 233 mm × 46 mm

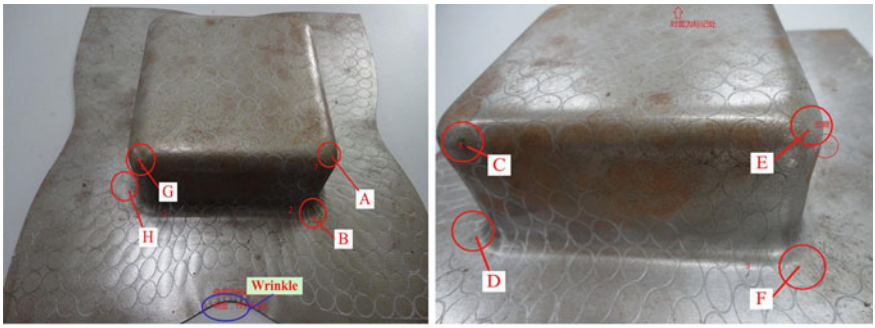


Fig. 81.3 The punch-stamped metal sheet with deformed *circle* marks, (**a**) the front face with deformed 4 *circle* marks (A, B, G and H), and (**b**) the back face with 4 deformed *circle* marks (C, D, E and F)

After digitization, PC of a part of metal sheet is shown in Fig. 81.4a. The 4 holes on the top of sheet are the additional artificial symbol marks used for stitching. There are 24 blocks of PC for this measurement. So the stitching process should be conducted in RE according to the stitching symbol. Figure 81.4b shows the stitching process of two blocks of PC according to artificial marks. Through stitching operation of 24 blocks of PC manually, the final digitized data are shown in Fig. 81.4c.

After achieving PC, it needs to delete noise or isolated points and simplify all PCs for mesh generation. After these operations, the meshed data are shown in Fig. 81.4d.

81.2.2 Mesh Mending and Contour Extraction

The holes on the meshed surface should be filled according to the neighbor mesh information. After filling the holes and checking the mesh information, the complete mesh surface can be fitted into the CAD surface. Figure 81.5a shows the whole mesh model composed of nearly 0.6 million triangles. Based on the triangle mesh,

theoretically the model could be fitted into NURBS (Non-Uniform Rational B-Splines) surface [8, 9]. But the deformed circles should be detected at first. So the mesh model should enter the contour detecting mode in Geomagic software. After the curvature sensitivity and delimiter sensitivity are set to be 20 and 10, respectively, the geometric model of metal sheet is shown in Fig. 81.5b, in which there is no surface patch but triangle meshes. Based on the contour-detecting model, the largely deformed circle marks should be extracted accurately in order to carry out the measurement. The deformed plane circles have been transformed into spatial distorted ellipses. The contour detecting goal is to achieve the whole shape and size of ellipses. Here the manual operation should be performed. Figure 81.5c

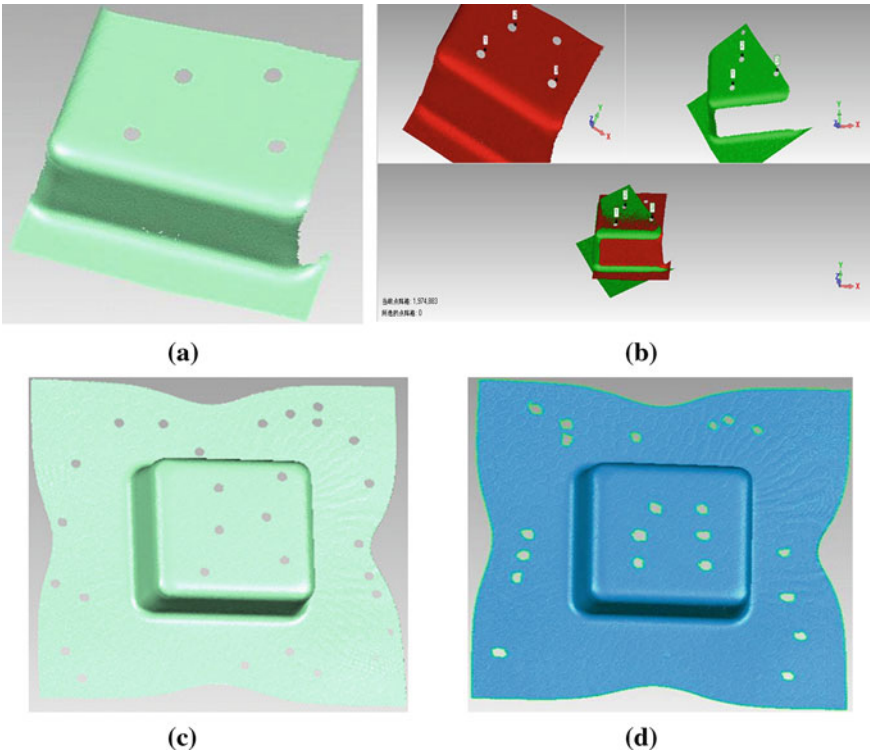


Fig. 81.4 **a** PC of a part of metal sheet; **b** the stitching process of PC; **c** the final PC of deformed metal sheet; **d** the meshed surface of metal sheet

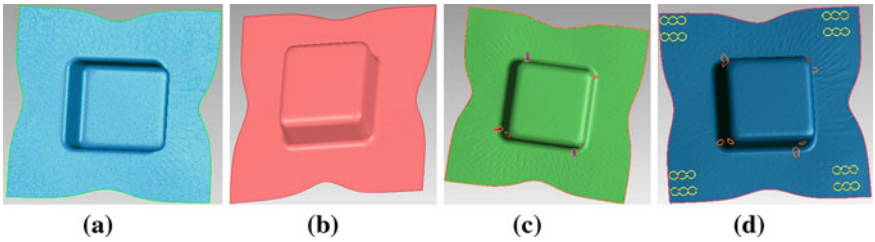


Fig. 81.5 **a** The filled whole mesh model; **b** the contour-detecting CAD model; **c** the detected contours of measured spatial ellipses; **d** the further edited contours of spatial ellipses

shows that all the measured ellipses are extracted manually. But these ellipse contours are not accurate. The following step is to edit the contours precisely and make them meet the design requirements. In addition, the segment length of contour is chosen to be 4 mm with even subdivision. The final contours of measured ellipses are shown in Fig. 81.5d.

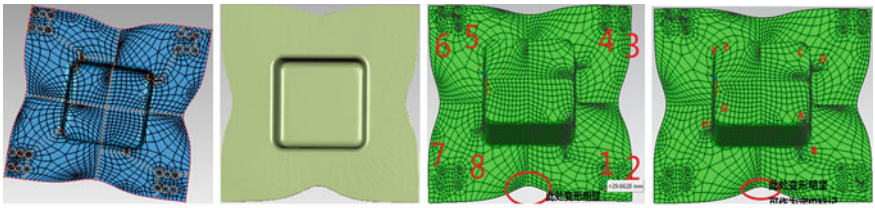


Fig. 81.6 **a** The patched surface; **b** the final CAD model; **c** the sampled circle marks used for error calculation; **d** the measured ellipses on the deformed CAD model of metal sheet

81.2.3 CAD Surface Model

Before forming the CAD model, the mesh model of metal sheet should be transformed into the geometric model consisting of surface patches. In Geomagic, 2583 patches are used to form the geometric model of deformed metal sheet. Patched surface is shown in Fig. 81.6a. At the same time, an option of surface relaxation could be selected to make the surface more even. In the patched surface, some large patches should be subdivided to meet the requirements of NURBS fitting operation. Following the above operation, the final CAD model could be obtained through NURBS fitting, which is shown in Fig. 81.6b.

81.3 Realization of Digital Measurement

After achieving the CAD model of deformed metal sheet, the data are imported into the CAD software of UG NX6. The measurement error is calculated by comparing the CAD data of metal sheet to the ones achieved by Vernier Caliper (accuracy of 0.01 mm). The sampled circle marks used for comparison are chosen from the unreformed ones on the plane of metal sheet, which are shown in Fig. 81.6c. They are marked from No.1 to 8. Each number stands for the area with 3 adjoining circle marks shown in Fig. 81.5d. By measuring the diameter sum of 3 adjoining circles, one can get errors shown in table. According to Table 81.1, the average error is 0.0036 mm, less than 0.01 mm. It means that the digitized measurement has an error of 0.01 mm compared to Vernier Caliper at least.

In the following step, the spatial ellipses are measured through UG algorithm. Because the ellipse marks is non-planar, the major and minor axes should be measured through surface unfolding method or orthogonal lines projecting method. Here the latter method is better. After four vertices on the ellipse are found and formed to be two orthogonal lines, the two lines are projected onto the corresponding CAD surface. The two curves projected onto the surface are the major and minor axes that need to be measured. On the other side, the previous manual method uses an inelastic rope to stick to the deformed marks along the

Table 81.1 The measurement errors compared to Vernier Caliper (unit: mm)

Circle team no. Method	1	2	3	4	5	6	7	8	Average value
Digitizing measurement	29.6628	29.7367	29.3043	29.5553	29.4271	29.3504	29.9033	29.4013	29.54265
Vernier Caliper	29.70	29.76	29.39	29.61	29.34	29.30	29.81	29.46	29.54625

Table 81.2 Data comparison (from A to H) between rope-stretching method and 3D-scanner digitizing method (unit: mm)

Ellipse no.	Axis	Method	A	B	C	D	E	F	G	H
Major		3D-scanner	12.2502	14.3657	14.4064	13.5676	13.4853	14.9118	14.7547	13.9684
		Rope-stretching	12.27	14.38	14.38	13.49	13.50	14.84	14.77	13.88
Minor		3D-scanner	8.1223	5.1037	9.6050	6.1086	7.4947	5.9636	6.6809	6.5012
		Rope-stretching	8.22	5.19	9.51	6.08	7.42	5.91	6.79	6.46

measured axis at first and then stretched it manually; afterwards, a Vernier Caliper is used to measure the rope length. Although this is a highly laborious and inaccurate job, it can meet the present requirements of low-accuracy measurement. Figure 81.6d shows the measured ellipses (from A to H) on the deformed CAD model of metal sheet. The measured data by both methods are shown in Table 81.2.

From Table 81.2, one can see that the rope-stretching method is deviated from the 3D-scanner method within 0.1 mm. Nevertheless, the 3D-scanner method has an average measurement error of 0.01 mm. Therefore, the rope-stretching method is proved to have a bigger error mainly cause of rope length deviation.

81.4 Conclusion

Spatially, it's hard to measure the deformed metal sheet by the traditional method. With the development of 3D measurement technology, the fringe projection method has become the mainstream solution to carry out the spatial size measurement during deformation. Combined with RE technology, the exact CAD model of metal sheet has been constructed in this paper. Through curve projection, the deformed ellipse marks have been measured precisely. The measurement accuracy has reached 0.01 mm. This method not only provides the precision measurement results for any spatially deformed surface, but also overcomes the laborious and tedious manual operation used in past. According to the experiments, the length data of major and minor axes have been changed dramatically on the largely deformed area of metal sheet, on which the ellipses sizes will lay a foundation for the further stress–strain analysis during deformation of metal sheet. In the next step, the strain could be achieved through digital measurement of deformed circle marks, and then stress could be also obtained. In this case, the FEM analysis of metal sheet forming could be comprehensively conducted based on the stress–strain analysis. At last, the optimized shape of original metal blank and some other conditions such as draw-bead parameters and pressing forces could be found.

Acknowledgments This work was supported by the Foundation of State Key Laboratory of Automotive Simulation and Control (No. 20111114); National Natural Science Foundation of China (No. 51005212 & No. 61275110); China Postdoctoral Science Foundation (No. 2011M500936 & No. 2012T50274); the Public-Service Technology Research Plan of Zhejiang Province (No. 2011C21003); the open fund of Key Laboratory of Space Laser Communication and Testing Technology from Chinese Academy of Sciences.

References

1. Seungwoo, K., Yibae, C., Jungtaek, O.: Reverse engineering: high speed digitization of free-form surfaces by phase-shifting grating projection moire topography. Int. J. Mach. Tools Manuf **39**(3), 389–401 (1999)
2. Quanqing, L., Hong, W., Yingjie, Z.: Key technology analysis in reverse engineering. Mech. Des. **6**(6), 4–7 (1999)
3. Giovanna, S., Franco, D.: Three-dimensional optical measurements and reverse engineering for automot-ive applications. Robot. Comput. Integr. Manuf. **20**, 359–367 (2004)
4. Claverley, J.D., Georgi, A., Leach, R.K.: Modelling the interaction forces between an ideal measurement surface and the stylus tip of a novel vibrating micro-scale CMM probe. Precis. Assem. Technol. Syst. **315**, 131–138 (2010)
5. Mahesh, C., Aarti, M., Rina, S.: Roughness measurement using optical profiler with self-reference laser and stylus instrument - a comparative study. IJPAP **49**(5), 335–339 (2011)
6. Göbel, W., Björn, M.K., Helmchen, F.: Imaging cellular network dynamics in three dimensions using fast 3D laser scanning. Nat. Method. **4**, 73–79 (2007)
7. Zhenzhong, W., Fuqiang, Z., Guangjun, Z.: 3D coordinates measurement based on structured light sensor. Sens. Actuators A Phys. **120**(2), 527–535 (2005)
8. Xuechang, Z., Xu, Z., Xuejun, G.: Reverse Modeling Technology and Innovative Design of Product. Beijing University Press, Beijing (2009)
9. Siyuan, C., Shaowang, X.: Reverse Engineering Technology and Application Based on Geomagic Studio. Tsinghua University press, Beijing (2010)

Chapter 82
Post-Hoc Evaluation Model for Development Workload of Gait Characteristic Database System

Dan Tang and Xiao-Hong Kuang

Abstract The software development workload is subject to such factors as field, staff, technique, environment and strategy. Thus it is difficult to measure the workload accurately, especially the measurement conducted by a third party after the completion of a project. With the gait characteristic database as a subject, this paper puts forwards a new post hoc evaluation model for workload by using workload algorithm model and expert review method based on full understanding of the system's basic information. Practices have proven that this model increases evaluation flexibility and effects and corrects the subjective bias of experts. It is quite applicable to the post hoc evaluation of software development workload.

Keywords Gait · Post-Hoc evaluation · Software measurement · COCOMO · System architecture

82.1 Introduction

The software development workload in software engineering field can be estimated in advance or evaluated afterwards. The advance estimation is usually conducted by the construction unit or developer at the early stage of a project for estimating the workload of software development so as to determine the project budget and underlie the staff, fund and material for the project. The post hoc evaluation is usually conducted by a third party evaluation organization, supervision organization or audit department at the late stage of a project for evaluating and verifying the actual workload of software development. The software projects become increasingly huge and keep on stretching in terms of time and space with

D. Tang (✉) · X.-H. Kuang
Department of Computer Science and Technology, Hunan police academy, Changsha, China
e-mail: tacom@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_82,

economic development. Especially in large-scale informatization construction, projects requiring heavy investments, long period and multi-industry and trans-department expertise are emerging one after another. Besides the construction unit and developer (undertaking unit) who are traditional subjects of the software project, many other parties like consulting institution, supervision organization, third party evaluation organization and audit department are also involved now. And there are some large-scale projects organized by several construction units and undertaken by different developers, thus greatly increasing the difficulty of project management. Therefore, the demand on post hoc evaluation model for workload of software development grows continuously, especially in the field of project acceptance, project audit, afterwards supervision and software test conducted by a third party organization. However, as software engineering master Roger S Pressman says, the estimation of the software cost and workload can never be accurate with too many variables like staff, technique, environment and strategy influencing the workload and final cost [1]. This paper takes the gait characteristic recognition database as an example and conducts a post hoc evaluation of its workload.

82.2 Post-Hoc Evaluation Model for Workload

There are three estimation techniques for workload of software development: a technique based on expert review, a technique based on algorithm model and a learning oriented technique. They have their own advantages and limitations, which means that no estimation technique may be applied to all development environments. Besides, the fast upgrade of software development method and technique challenges the estimation techniques. The estimation technique should be selected based on the characteristics and available information of the project and adjusted according to actual situation. Different techniques may be combined for estimation in the light of their characteristics to improve the accuracy [2]. The early stage workload should be fully considered during estimation and a standard library of workload should be established.

The techniques based on expert review and based on algorithm model are well developed and widely used. There are a lot of methods based on algorithm model, like COCOMO [3], SLIM, Checkpoint [4], SEER-SEM, Estimacs and PRICE-S, among which the COCOMO is a universally accepted and applied model that has been published. The techniques based on expert review are valuable in absence of quantized history data and are the most popular methods so far. However, such techniques greatly depend on estimators' experience and thus are highly subjective [5].

This project puts forwards a new post hoc evaluation model for workload based on algorithm model and expert review for the purpose of integration and correction.

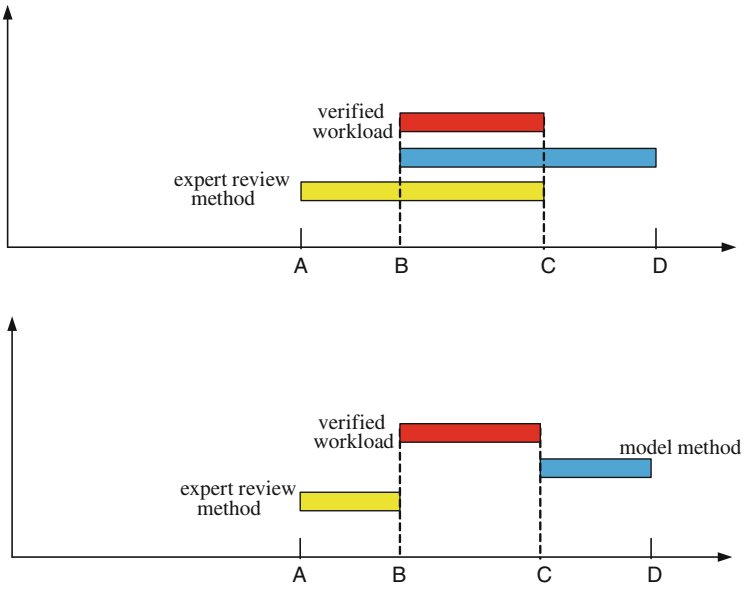


Fig. 82.1 Evaluation model for workload based on algorithm model and expert review

Due to the estimation deviation, the workloads estimated by the methods based on algorithm model and experts' reviews lie in certain ranges respectively. Two situations may appear like Fig. 82.1:

- When the workload intervals of the two method overlap each other, the overlapped part should be the result as shown in Fig. 82.1a: suppose that the workload range is [A, B] estimated by the expert review method, and [C, D] estimated by the algorithm model method, the [B, C] should be the verified workload after correction.
- When the workload intervals of the two method are independent from each other, the range between the two should be the result as shown in Fig. 82.1b: suppose that the workload range is [A, B] estimated by the expert review method, and [C, D] estimated by the algorithm model method, the [B, C] should be the verified workload after correction.

82.3 Basic Information of the System to be Evaluated

A great deal of factors influence the development workload such as service context, user's demand, system architecture, software function, development group, technical route, development environment, deployment environment and system data volume, etc. The post hoc evaluation should analyze these factors comprehensively, or the result may be much different from the actual workload.

Therefore, the post hoc evaluation must investigate and verify the basic information of a project.

This paper evaluates the workload of a gait characteristic database management system. The gait recognition is a distance non-contact recognition technique [6]. Different from such biological recognition techniques as iris and fingerprint recognition, the gait recognition is a much more reliable non-contact technique that will hardly be impacted by distance and environment. The database system for gait recognition focuses on gait data acquisition platform, gait data processing technique, gait characteristic database design and gait characteristic data management platform, etc. This project started in December 2007, the system architecture design started in April 2008, the data acquisition started in October 2008 and the project was put into operation successfully in February 2011. The system is capable of data acquisition, data processing, date query, basic data management and file management.

82.3.1 Basic Modules and Function Points of the System

The system functions depend on user's demand. The number of system modules and function points, selection of technical route and difficulty of the technique are main factors influencing the system workload estimation. As to the post hoc evaluation, the processes and details which should be investigated as required by the system cannot be fully understood, so the solution is to use the actually developed functions and the requirement specification as the evaluation points. This system mainly provides gait information query, gait recognition, gait analysis, gait recognition algorithm management, basic personal information management, gait database management and API interface management, basic gait data management, etc.

82.3.2 Running Environment of the System

(1) *Network and hardware environment*
The network and hardware environment determines the workload of system deployment. This system employs centralized deployment by 4 PC servers, including two database servers for hot standby, a file server and a WEB and application server. All servers are installed with windows 2003 and Oracle 10 g database. Apache is applied to web service.
(2) *System software and development environment*

(a) *Operation system*

- Server operation system: Windows 2003
- Workstation system: windows XP, windows vista and windows 7

(b) *Database management system*
The service database uses Oracle 10 g and workflow database uses SQL Server 2005.
(c) *Development environment*
Netbeans6.7, P/L Sql Developer 8.0.3 and rational rose 2003 as CASE tool.

(3) *Data volume*
Data volume is closely rated to the system workload as well. This system uses Oracle and SQL Server databases. The Oracle is applied to saving service data like personal information, user data, gait characteristics and file saving path. The SQL Server is applied to saving workflow process instantiation data. This system has acquired 1440 front, profile and 45 degree image sequences of 120 persons, making the data volume reach 18G.
(4) *System user number*
The system comprises four user roles: administrator, data maintenance administrator, normal inquiry user and external application. It supports simultaneous inquiries from 120 users in a testing environment.
(5) *Program code line data*
The program code line data directly reflects the workload but is equal to the workload. Furthermore, ongoing software industrialization brings more powerful case tools, IDE development tool and the third party components, making many codes able to be directly generated by tools and many others able to be simply customized. Additionally, design difficulty and system impact of codes for different layers like the service layer, interface layer and data layer are different. Therefore, the post hoc evaluation should consider all the factors. The data volume of this system includes all data volume of the integrated gait characteristic management system since 2007. Besides the service logic layer as the main body, interface, data access, storage process and functions are also included. The data statistics was conducted by a dedicated code line statistical tool SourceCount. It is verified that the system code line is 30152. Table 82.1 shows details of the code line data.
(6) *Development team and its R & D progress*
The development team is another important influencing factor of workload. The workloads of same project are totally different for development personnel at different levels. Our development team comprises: 1 system analyst, 2 senior programmers, 1 programmer and 1 tester. The core developers are quite

Table 82.1 Statistical chart of system code line data

Subproject	Source code statistics (lines)
Source code of service logic layer	12923
Source code of interface presentation layer	9078
Source code of data access layer	6841
Database storage process/Source code of function	1310
Total	30152

familiar with the business process and experienced in project development and management. This 5-person team spent 12 months on development and 12 months on early stage data acquisition and preprocessing.

82.4 Workload Evaluation

Besides the investigation of the basic system information and primary project data, the post hoc evaluation should combine the complete historical data and system analysis technique to provide an estimation result with acceptable risk and improve the estimation accuracy. Currently, our evaluator has collected the workload data of over 50 informatization projects in recent 5 years and established a sound evaluation system, a workload history database and a professional evaluation team. In following paragraphs, we will evaluate the workload of this project by the post hoc evaluation model which starts from the man-month standard and evaluation method.

82.4.1 Model Method

The workload of this project is estimated by a post hoc evaluation: middle COCOMO model and the formula:

$$E = C \times KLOC^{a} \times \prod_{i=1}^{15} fi \tag{82.1}$$

- E is the development labor (man-month)
- C is the model coefficient (development mode: organized, embedded and semi-independent)
- The model system and index data are shown in Table 82.2. This project is a semi-independent project, so the C is 3.0
- KLOC is estimated code line number

Table 82.2 Model coefficient and index

Development mode	Model coefficient C	A Model index
Organized	3.2	1.05
Semi-independent	3.0	1.12
Embedded	2.8	1.20

- We believed that the actual manual code volume is about 30 K, so the KLOC is 30 K.
- a = model index (corresponding to the development mode). This project is a semi-dependent project, so a = 1.12.
- fi (i = 1…15) are cost elements.

The system cost elements are: Product property, Computer property, Staff property, Project property.

$$\prod_{i=1}^{15} fi = 0.654 \tag{82.2}$$

Other suitable models are also available according to the actual situation, like SLIM, Checkpoint, SEER-SEM, Estimacs and PRICE-S.

The final formula of workload of this project E (man-month) is calculated as follows:

$$E = C \times KLOC^{a} \times \prod_{i=1}^{15} fi \tag{82.3}$$

$$E = 3 \times 30^{1.12} \times \prod_{i=1}^{15} fi \tag{82.4}$$

$$E = 3 \times 30^{1.12} \times 0.645 \tag{82.5}$$

The standard workload is E = 87 man-months. Due to the 20 % deviation, the workload figured out by the model method is 70–104 man-months.

82.4.2 Expert Review Method

The worldwide application status of estimation techniques shows that the estimation techniques based on expert review are the most popular methods around the world. Their great usability and flexibility may make contribution to their universality. Such techniques are very useful in absence of quantized history data.

We organized five experts in this field for evaluation by following steps:

(a) provide the experts with project specifications and estimated statements
(b) hold meeting for the development group and experts to discuss the factors related to the scale and workload
(c) make anonymous iteration forms by the experts
(d) summarize the estimation and return the result as an iteration form to the experts
(e) hold a meeting for the development group by a coordinator to discuss large estimation deviation

(f) ask the experts to review the estimation summary and submit another anonymous estimation as an iteration form
(g) repeat steps (4)–(6) for three times to get a minimal result and a maximal result.

The final result of this method: the minimal workload is 60 man-months and the maximal workload is 100 man-months.

82.4.3 Final Verification

We took the overlapped part of the model method and expert method as shown in Fig. 82.1. It is verified that workload of the gait characteristic database system should be 70–100 man-months.

82.5 Conclusion

The software development workload is influenced by user, region, business, development group, technical route, development mode, system hardware environment, etc., so it can be hardly measured accurately. The third party conducting the development workload post hoc evaluation has not gone through the software life cycle including project approval, design, development, test and deployment, so it cannot understand project details and development progress well, which makes the post hoc evaluation more difficult. This paper puts forwards a new post hoc evaluation model for workload which combines the algorithm model and expert review method. It takes advantages of the two methods, thus efficiently avoids subjective bias caused by the expert review and flexibly corrects the original algorithm model, making it fit current software development status in China. After conducting the post hoc evaluation for a number of projects and comparing the results with the workload standard libraries, this model is proven accurate, especially when the results of the two methods overlap each other. The workload of gait characteristic management system estimated by the post hoc evaluation is close to the actual workload submitted by the project group. When the results estimated by the algorithm model and expert review method are totally different, the estimation result obtained by the post hoc evaluation may lie in a large range and make the quantitative evaluation meaningless. In this case, another algorithm model should be used so as to make sure that the workload is in a reasonable range, which will be significant to further researches.

References

1. Pressman, R.S.: Software Engineering-A Practitioner's Approach—Required. Mechanical industry press, Beijing (2002)
2. Wang, Q.: A Comparison of effort estimation techniques in software development projects. J. Zhejiang **3**(04), 121–125 (2005)
3. Li, M., He, M., Yang, D.: Software cost estimation method and application. J. Softw. **4**(18), 775–795 (2007)
4. Molokken, K., Jorgensen, M.: A Review of Software Surveys on Software Effort Estimation. In: International Proceedings of Empirical Software Engineering, pp. 223–230 (2003)
5. Tao, Y., Jin, H., Wu, S.: An optimistic checkpoint mechanism based on job characteristics and resource availability for dynamic grids. Wuhan Univ. J. Nat. Sci. **3**(36), 621–625 (2011)
6. Ye, B., Wen, Y.: Survey of gait-based human identity recognition techniques. Comput. Appl. **11**(25), 2577–2580

Chapter 83
Acquisition Time Performance of Initial Cell Search in 3GPP LTE System

You Zhou, Fei Qi and Hanying Hu

Abstract This paper focuses on shortening the mean time to acquire for initial cell search in the 3rd Generation Partnership Project (3GPP) Long Term Evolution (LTE) system. Exact result of mean time of the cell search is obtained by use of a state transition diagram of discrete-time Markov process for the acquisition process. From the deduced mean time formula, it can be seen that the false alarm probability plays a more important role in the acquisition time performance. In order to perform the initial cell search quickly, a new PSS detection structure which includes a verification module to suppress the false alarm is developed. Simulations show that the new structure can decrease the false alarm probability, and reduce the total acquisition time especially in the low SNR circumstance.

Keywords LTE · OFDM · Cell search · Acquisition time · Verification module

83.1 Introduction

3GPP LTE is recognized as the key technology for the next generation wireless communication system [1]. The LTE employs some advanced technologies which include Orthogonal Frequency Division Multiplexing (OFDM), Multiple Input Multiple Output (MIMO) and so on [1]. Demodulation of OFDM signal is vulnerable to timing offset and frequency offset. A User Equipment (UE) wishing to access an LTE cell must first undertake a cell search procedure, which is tightly connected to the Primary Synchronization Signal (PSS) and Secondary Synchronization Signal (SSS) transmitted by an LTE cell [2].

Y. Zhou (✉) · F. Qi · H. Hu
Zhengzhou Information Technology Institute, Zhengzhou, Henan, China
e-mail: emailzhouyou@sina.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_83,

The synchronization is a continuous and periodic process, which must be always active. So it is of significance to acquire it in a short time and the mean acquisition time of the cell search is a very important parameter to reflect the performance of the cell search structure. Currently, PSS detection in LTE has been studied considerably, and the topic lies in increasing the detection probability or simplifying the detection process. The classic maximum likelihood detector takes advantage of good autocorrelation and cross-correlation performance of PSS sequence in the time domain [3]. A normalized detection is proposed in order to avoid the confused detection of the peak sequence [4]. It firstly detects the cyclic prefix (CP) type which leads to the detection of symbol synchronization, and then acquires the PSS in the frequency domain which greatly simplifies the acquisition process [5]. It achieves a reliable PSS detection with a much lower complexity by exploiting the central-symmetric property of Zadoff-Chuf (ZC) sequences in the time domain [6].

In this paper we will focus on the mean acquisition time performance of initial cell search via a state transition diagram of discrete-time Markov process. From the numerical result of mean acquisition time, we find that false alarm property plays a more important role in the acquisition time performance. In order to perform the cell search quickly a new structure is developed to suppress the false alarm. The performance of the proposed structure is compared with that of normal PSS detection structure.

The remainder of this paper is organized as follows. A system model of PSS transmission and receiving is provided in Sect. 83.2. Section 83.3 deduces the mean time of the acquisition process and Sect. 83.4 introduces a new PSS detection structure. In Sect. 83.5, based on the numerology provided in the physical layer specification for 3G LTE, the performance of the proposed structure is investigated and compared with the normal detection structure. Finally the conclusions are drawn in Sect. 83.6.

83.2 System Signal Model

Let us firstly introduce the notation that will be adopted throughout the paper. Superscripts * denotes complex conjugate. $[\cdot]^T$ is the transpose of a matrix. $\mathrm{diag}\{\cdot\}$ is the diagonal matrix. A circularly symmetric complex Gaussian random variable w with mean m and variance σ^2 is denoted by $w \sim \mathcal{CN}(m, \sigma^2)$. The operator $|\cdot|$ returns the modulus of a complex number. $\mathrm{mod}(x, y)$ is the remainder operator. $\max(\mathbf{x})$ gives the maximum value of vector $\mathbf{x}$.

In LTE system, each cell is identified by the cell identification information carried by PSS and SSS. A length of 63 ZC sequence is used to generate the PSS, which occupies the central 72 sub-carriers [7], as illustrated in Fig. 83.1. It is transmitted every 5 ms.

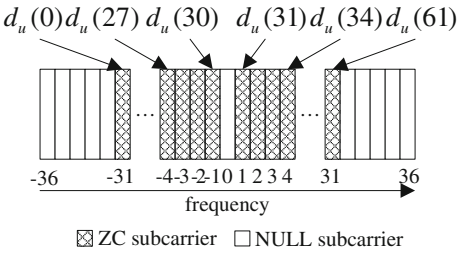


Fig. 83.1 PSS mapping of 3G LTE in the frequency domain

The sequence $d_u(n)$ is generated according to

$$d_u(n) = \begin{cases} \exp[-j\pi un(n+1)/63] & n = 0, 1, \ldots, 30 \\ \exp[-j\pi u(n+1)(n+2)/63] & n = 31, 32, \ldots, 61 \end{cases} \tag{83.1}$$

where the index u is depending on physical-layer identity $N_{ID}^{(2)}$ within the physical-layer cell-identity group. Without loss of generality, we use $\mathbf{x} = [0, d_u(31), \ldots, d_u(61), \mathbf{z}_1 \ldots \mathbf{z}_2, d_u(0), \ldots, d_u(30)]^T$ to represent the transmitted PSS vector, $\mathbf{z}_1$ and $\mathbf{z}_2$ are random transmitted symbols whose size depends on the bandwidth of the LTE system. The received PSS base-band signal on the antenna i is given by

$$\mathbf{r}_i = \mathbf{E}_i \mathbf{F} \mathbf{H}_i \mathbf{x} + \mathbf{w}_i \tag{83.2}$$

where $\mathbf{E}_i = \text{diag}\{1, e^{j2\pi\varepsilon_i/N_{fft}}, \ldots, e^{j2\pi\varepsilon_i(N_{fft}-1)/N_{fft}}\}$ with ε_i representing the normalized frequency offset (frequency offset normalized to a subcarrier spacing of OFDM symbols), N_{fft} is the FFT size, $\mathbf{F}$ is the IFFT matrix with $[\mathbf{F}]_{nk} = e^{\frac{j2\pi nk}{N_{fft}}}/\sqrt{N_{fft}}$, $\mathbf{H}_i = \text{diag}\{H_i(0), H_i(1), \ldots, H_i(N_{fft}-1)\}$ represents the frequency-domain channel attenuation matrix. $\mathbf{w}_i$ is a vector of additive white Gaussian noise with $\mathbf{w}_i[k] \sim \mathcal{CN}(0, \sigma_w^2)$. $\mathbf{y}_i$ will be fed to the PSS detection structure to perform the acquisition process.

83.3 Analysis of Mean Acquisition Time

There is no time limit for the acquisition process, since PSS is periodically transmitted. Therefore, it is almost certain that the whole system will be in an acquisition state and the measurement of performance lies in the mean acquisition time $\text{E}(T_{acq})$. We use a state transition diagram of discrete-time Markov process to describe the process of the initial acquisition in the LTE system, as shown in Fig. 83.2.

The points on the outer circle represent the false alarm state, and ones on the inner circle are detection points whose number depends on the PSS transmission period, synchronization window size and sampling frequency. Assuming that there are V possible detected points, and each is labeled from 1 to V. Among those

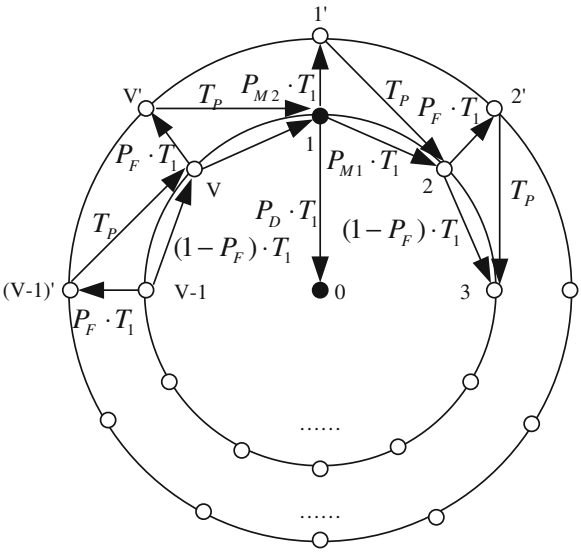


Fig. 83.2 Flow graph of 3G LTE initial cell search

points, there is only one point which corresponds to the collective state H_1 which is labeled as 1 with generalization, and the rest ones belong to false alarm state H_0. Each detected point corresponds to a false alarm state except for 1. Here a serial search strategy is adopted, any past time search information will not be used to change the search direction. The system will go to the next detected point after T_P seconds when a false alarm appears. The detection probability from 1 to 0 is P_D called detection probability. The missing probabilities from state 1 to $1'$ and 2 are P_{M2} and P_{M1}, which represents two kinds of missing scenarios, one is false detected PSS index, and the other is missing the timing point. The relationship is $P_{M1} + P_{M2} + P_D = 1$. The false alarm probability from i to i' $(i \neq 1)$ is P_F, and $1 - P_F$ indicates the probability from i to $i + 1$. Let z indicate the unit delay operator. Its value depends on the sampling rate. Furthermore, we can get the generating function from one point to the other,

$$H_{1\to 0}(z) = H_D(z) = P_D z \tag{83.3}$$

$$H_{i\to i+1}(z) = H_0(z) = (1 - P_F)z + P_F z^{K_p+1}, (i \neq 1) \tag{83.4}$$

$$H_{1\to 2}(z) = H_M(z) = P_{M1} z + P_{M2} z^{K_p+1} \tag{83.5}$$

Here we assume that $T_P = K_p T_1$. Then we can simplify Figs. 83.2–83.3. According to Mason formula [8], the generating function from i to 0 is

$$H_i(z) = \frac{H_D(z) H_0^{\mathrm{mod}(V-i+1,V)}(z)}{1 - H_M(z) H_0^{V-1}(z)} \tag{83.6}$$

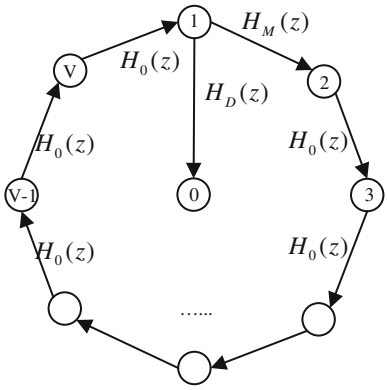


Fig. 83.3 Simplified flow graph of 3G LTE initial cell search

Since the search process can start at any one of the detection points, we assume that the probability follows uniform distribution with probability$1/V$. Then the generating function of the whole acquisition process follows that

$$\begin{aligned} H(z) &= \sum_{i=1}^{V} \frac{1}{V} H_i(z) = \frac{H_D(z)}{V\left[1 - H_M(z)H_0^{V-1}(z)\right]} \sum_{i=1}^{V} H_0^{\mathrm{mod}(V-i+1,V)}(z) \\ &= \frac{H_D(z)\left[1 - H_0^V(z)\right]}{V\left[1 - H_M(z)H_0^{V-1}(z)\right]\left[1 - H_0(z)\right]} \end{aligned} \tag{83.7}$$

The mean acquisition time can be expressed as [9, 10]

$$\begin{aligned} \bar{T}_{acq} &= \left.\frac{\partial H(z)}{\partial z}\right|_{z=1} \cdot T_1 \\ &= \frac{T_1}{H_D(1)}\left[H_D'(1) + H_M'(1) + (V-1)H_0'(1)\left[1 - \frac{H_D(1)}{2}\right]\right] \\ &\approx \frac{T_1}{P_D}\left[2 - P_D + (V-1)(K_P P_F + 1)(1 - \frac{P_D}{2})\right] \end{aligned} \tag{83.8}$$

Because of the PSS near perfect cross-correlation properties, it is reasonable to assume that $P_{M2} \ll 1$ during the derivation process.

Figure 83.4 shows $\bar{T}_{acq}$ trend with the change of different $P_D$and P_F according to Eq. (83.8), where $V = 1/153600$, $K_P = 307200$. $\bar{T}_{acq}$ decreases with the increase of P_D under certain P_F, while it increases with the increase of P_F under certain P_D. The relationship between $\bar{T}_{acq}$ and P_D, P_F is basically linear. $\bar{T}_{acq}$ equals to 2.5 ms with limited condition of $P_D = 1$ and $P_F = 0$. Furthermore, it is shown that P_F has a greater influence on $\bar{T}_{acq}$ than P_D, to which should be paid great attention.

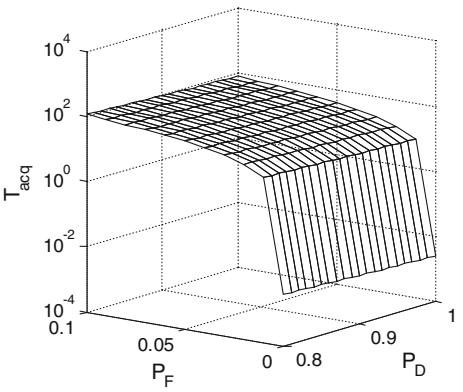


Fig. 83.4 Mean acquisition time with different P_F and P_D

83.4 PSS Detection Structure

From the previous discussion we can see that even a little increase of false alarm probability will lead to a great increase of mean acquisition time. Taking acquisition performance and complexity into account, we design a PSS acquisition structure shown in Fig. 83.5. At the first stage, it matches the received data from single antenna in the time domain to simplify the acquisition process. The sampled signal from one of the antennas is fed to a Low Pass Filter (LPF) with a pass bandwidth of 1.08 MHz. The maximum likelihood detector can be expressed as

$$(m, M) = \arg\max_{m,M} |\sum_{i=0}^{N-1} r'(i+m) S_M^*(i)|^2 \tag{83.9}$$

where $r'(i)$ represents the filtered signal, i is the time index, m is the timing offset, N is the FFT size, $S_M(i)$ is the local PSS replica with index M in the time domain. Having finished the first stage, received data from different antennas is performed with FFT. The corresponding PSS data is extracted to correlate with the local PSS replica in frequency domain which can be expressed as

$$A_u = \sum_{k=1}^{N_{ant}} |\sum_{i=0}^{61} R_k(i) d_{M'}^*(i)|^2 \tag{83.10}$$

where $R_k(i)$ is the received PSS data of antenna k in the frequency domain, N_{ant} is the number of antennas in the receiving system, and $d_{M'}(i)$ is local PSS replica with index M' in the frequency domain. During the second stage, the detector selects the index with the largest correlation value in the frequency domain.

$$M' = \arg\max_{M'} A_{M'} \tag{83.11}$$

If M' is identical with M, the whole PSS detection process is over, or go back to the first stage.

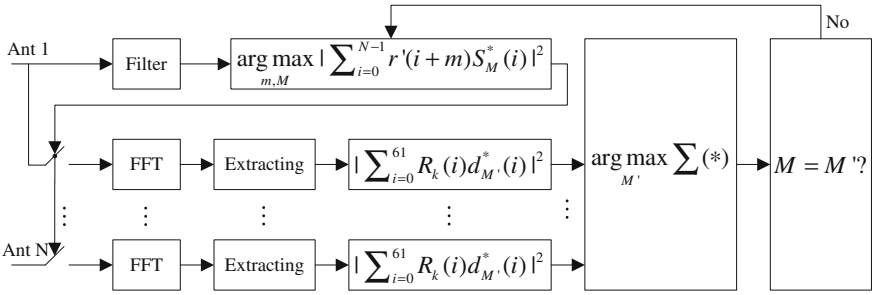


Fig. 83.5 PSS detection structure with verification module

The detection FFT module corresponding to each antenna can be multiplexed with FFT module used for OFDM demodulation. Also there are only 62 subcarriers containing PSS, so the increase of computation of the proposed structure is very small.

83.5 Simulations

Simulation parameters are based on the numerology in the 3G LTE physical layer specification [7]. Maximum frequency offset is 2 kHz and extended typical urban channel model (ETU) is used [11].

Figure 83.6 shows the false alarm performance comparison between normal structure (NS) and proposed structure (PS) with different number of antennas. Normal structure is detection structure without the verification module in the frequency domain. From the figure, we can see that PS has a better performance than that of NS, while PS with 4 antennas is better than PS with 2 antennas, especially for low SNR. The difference is becoming smaller and smaller along

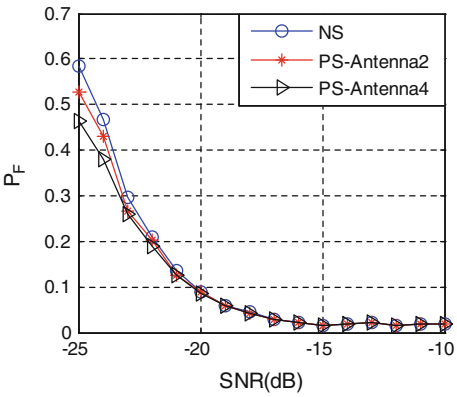


Fig. 83.6 P_F comparison between the two structures

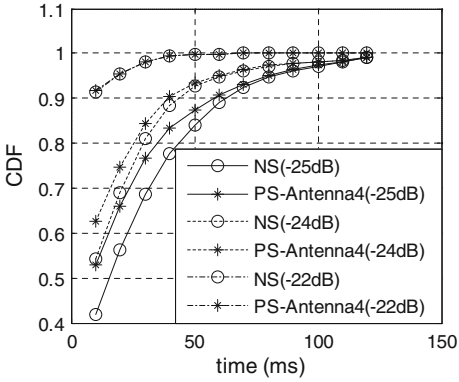


Fig. 83.7 CDF of cell search time comparison between the two structures

with the increase of SNR. The reason is that large SNR means a high credibility of the first stage, under which scenario the effect of verification module is weakened.

Figure 83.7 shows the CDF of cell search time of the two structures. It can be seen that the CDF of both structures is increasing along with the time while PS has a better performance than that of NS. Scenarios with different SNR have different kinds of performance. Lower SNR means larger difference in CDF performance between the two structures. The superiority of PS over NS is weakened with the increase of SNR. Cell search time also have an impact on the CDF performance. The advantage of PS finally disappears over the time.

83.6 Conclusion

In this paper, the mean time of initial cell search in 3GPP LTE system is studied, from which it can be found that the false alarm probability plays a more important role in the acquisition time performance. A novel PSS detection structure is proposed to suppress the false alarm by adding a verification module. Simulation results show that the proposed structure can decrease the false alarm probability, and reduce the total acquisition time especially in the low SNR circumstance.

References

1. Zhang, X., Tian, T., Zhou, X., Wen, Z.: LTE air interface technology and performance, pp. 1–6. Posts & Telecom press, Beijing (2009)
2. Shen, J., Suo, S., Quan, H., Zhao, X., Hu, H., Jiang, Y.: 3GPP long term evolution: principle and system design, pp. 280–300. Posts & Telecom press, Beijing (2009)
3. Sesia, S., Toufik, I., Baker, M.: LTE—the UMTS long term evolution, from theory to practice, pp.148–150. Wiley, New York (2009)

4. Silva, E.M., Dolecek, G.J., Harris, F.J.: Cell search in long term evolution systems: Primary and secondary synchronization. In: IEEE Third Latin American Symposium on Circuits and Systems, pp. 1–4 (2012)
5. Kim, Y.B., Chang, K.H.: Complexity optimized CP length pre-decision metric for cell searcher in the downlink of 3GPP LTE system. In: IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications, pp.895–899 (2009)
6. Zhang, Z., Liu, J., Long, K.: Low-complexity cell search with fast PSS identification in LTE. IEEE Trans. Veh. Technol. **61**(4), 1719–1729 (2012)
7. 3GPP TS 36.211 v10.4.0, Technical specification group radio access network: Evolved universal terrestrial radio access, physical channels and modulation (Release 10) (2011)
8. Wu, D., Yang, L., Zhang, Y.: Signal and linear system analysis, pp. 350–357. Higher Education Press, Beijing (2006)
9. Polydoros, A., Weber, C.L.: A unified approach to serial search spread-spectrum code acquisition-part I: General theory. IEEE Trans. Commun. **32**(5), 542–549 (1984)
10. Polydoros, A., Weber, C.L.: A unified approach to serial search spread-spectrum code acquisition-part II: A matched-filter receiver. IEEE Trans. Commun. **32**(5), 550–560 (1984)
11. 3GPP TS 36.101 v10.6.0, Technical specification group radio access network: Evolved universal terrestrial radio access, user equipment radio transmission and reception (Release 10) (2012)

Chapter 84
Multi-Agent System Set for Software Maintainability Design

Xiaowei Wang, Wenhong Chen, Luping Pan, Yanping Cui, Xinxin Tian and Si Wu

Abstract In order to search out the useful resources about software maintainability requirement from the huge contents in the rapid development of Internet, an intelligent decision-making model in software maintainability design is used in this paper. By introducing the Agent technique, intelligent decision-making model can rationally evaluate the affections between all indexes to select the best design scheme. The model mainly consists of two steps: The first one is Data-Mining and building information warehouse about software maintainability analysis. The second one is reasoning based on cases and evaluates the maintainability index. Giving the task to main-Agent, the main Agent communicates with the sub-Agents and allocates the task to all of them. The precedent warehouse not only instructs Agent to cooperate with other Agents, but also supplies it with the datum of intelligent actions. The Agent technology applied in software maintainability design. It's proved that it's an effective way to evaluate the scheme of maintainability design.

Keywords Software maintainability · Agent · Data-mining · Intelligent decision-making

84.1 Introduction

Most software projects in society are completed by the components cooperatively in special environment.To ensure the efficiency of the system, it's essential to control the process of a software project. The traditional methods have been applied extensively, such as the technology of working process in software system,

X. Wang (✉) · W. Chen · L. Pan · Y. Cui · X. Tian · S. Wu
Wuhan Ordnance N.C.O Academy of PLA, Wuhan, China
e-mail: tangsengwang163@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_84,

the theory of enterprise's recombining and the theory of standardization of management. But the disadvantages of the theories are obvious day by day [1].

Firstly, they can't evaluate the procedure, mode and method effectively. Secondly, the phantasmagoric recourses can't be utilized efficiently. Thirdly, the components who are distributed in a system can't supervise the system's changing intelligently just because they are always controlled in a concentric way.

At present, the same disadvantages are existed in software maintainability design. It's known that maintainability is an important attribution of design and fundamental characteristic of quality. With the development of the science and technology, the object of maintain is more complex and contain more technology. So, new method must be applied in maintainability design [2].

Agent is very popular in the field of artificial intelligence recently, especially with the boom of Internet. The definition of Agent can be found in most papers, so there is no need to repeat in this paper. In different fields, different Agent model is established. However, the Agent function model includes apperceive module, communication module, reasoning module, execution module, inner-station warehouse and information warehouse and so on. Especially in multi-Agent system, cooperation is an important characteristic, owning to the distribution of resources and the limit of ability. Interaction is the base of cooperation and communication is an important means to interact.

Agent is applied to software maintainability design in this paper and an intelligent decision-making model is set up to evaluate the affections between all indexes and to select the best design scheme.

84.2 The Model of a Software Intelligent Decision-Making

The model consists of two steps mainly: The first one is Data-Mining and building information warehouse about software maintainability analysis. The second one is reasoning based on cases and evaluates the maintainability index [3].

The user gives the task to main-Agent. The main Agent communicates with the sub-Agents and allocates the task to all of them. For example: the following twelve properties are informed to each sub-Agent: (1) usability; (2) clarity; (3) modularization; (4) portability; (5) testability; (6) technical review; (7) reparability; (8) quality assurance; (9) dubugging; (10) simplified parallel process; (11) software component; (12) majorization and compromise. After this, all the sub-Agents start to self-study and Data-Mining in the light of the main-Agent. Some of the sub-Agents maybe are mobile-Agents. The mobile-Agents work by the Net [4]. They seek the useful information and services in the net. All the Agents cooperate and balance the loads. When the mobile-Agent finds the information warehouses, it moves to the server and then search for the useful information. After completed the task, the results are fed back to the main-Agent. The process of it can be illustrated by the Fig. 84.1.

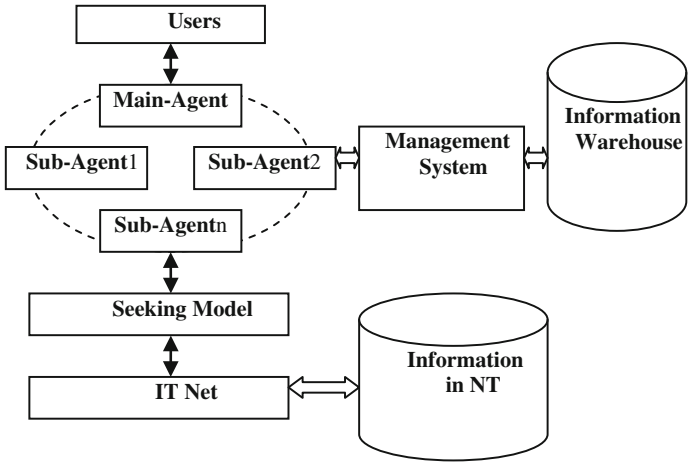


Fig. 84.1 The frame of software maintainability design model

The adaptive optimization is based on the intelligent analysis of the indexes. An important step is to predict the maintainability indexes, including mean-time-to-restore-system (MTTRS), mission-time-to–restore-function (MTTRF), maximum time to repair, mean-maintenance-time, mean-time-to-repair (MTTR), and maintenance ratio. All the designing indexes are interrelated. The pivotal work is to evaluate the effects to other indexes when one index is changed and to estimate whether or not the capability of the whole system is improved. All the calculating work is complex and the workload is tremendous. The Agent system has the capability of distributing decision-making and mobile calculation. When an Agent's task exceeds its capability, the main-Agent redistributes the tasks. This redistribution is also called mission transferring which can be processing between the sub-Agents in different time and position asynchronously. When the task is completed, all the sub-Agents give the feedback to main-Agent. In the Agent system, the subjection Agent also can be the main Agent, and the former main Agent can be one of the sub-Agent when it's needed. The relation is determined by the environment and the resource's distribution [5]. Especially when the main-Agent can not work, the role of main-Agent can be transferred to the optimum Agent. And when a sub-Agent can not work, the task of this Agent will be transferred to other Agents. In a word, the system works in a steady way. The relations among Agents can be illustrated by the Fig. 84.2.

When it comes to a single Agent, each one can deal with the task of maintainability design and is independent in some time. The base frame of sub-Agent based on information is made up of property, method, information, reasoning, language, function wizard, information transferring, the information warehouse and the server. The base frame can be illustrated by the Fig. 84.3.

The communication between Agent and user bases on the social ability of Agent, and Agent language is used. The self-adapted ability enables the Agent not

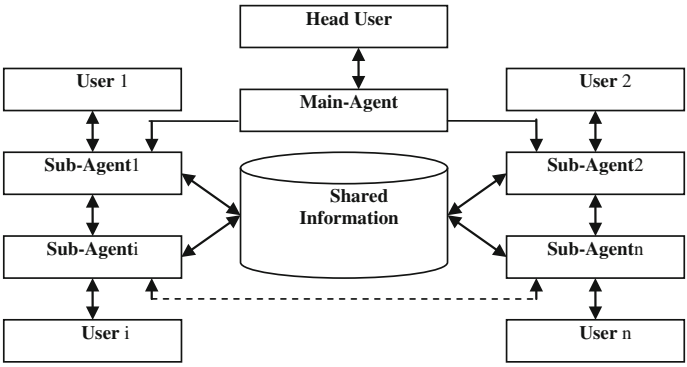
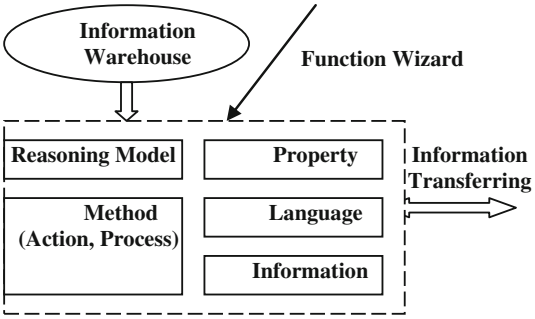


Fig. 84.2 The frame of cooperation among agents

Fig. 84.3 The frame of agent



only to react to the environment simply but also to analyze the fuzzy and developmental information from the users, which embodies the restriction to the task as a whole.

84.3 The Information's Application

The precedent warehouse not only instructs Agent to cooperate with other Agents, but also supplies it the datum of intelligent actions. There are three levels in the information warehouse: The first one is professional knowledge about software maintainability design, including the information about the structure of task. It maybe includes the information about the eight properties about software maintainability, such as predigesting the design and repair. The second one is the information for reasoning. It's about communication, cooperation, decision, and task's management. There are some rules for reasoning. The rule is: If subtask ('task', subtask) and state ('subtask', started) then state ('task', started). The third one is the information for controlling, which is the bridge between the reasoning knowledge and software maintainability information. The information about

maintainability is conversed to professional information through reasoning. The rule is: If scheme (conditions, conclusion) and all_true(conditions), then add(conclusions) [6].

The first step is to access the information warehouses and experience warehouses to seek for the similar solutions. And then apply them in the task. When the method is combined with Agent, the method can be faster and more precision. The following is an example.

Based on the properties of software maintainability and the newly developments in this field, twelve properties of maintainability are listed in this paper. All the twelve properties are not existed alone. There are relations among them. When one property is optimized, it affects other properties. In another word, when one Agent optimizes himself, it affects other Agents. The effects among properties are called "facilitations", marked as f_{ij}. The value of facilitation is between -1 and 1. The relations among the twelve properties can be expressed by the Table 84.1.

After analyzing the facilitations, the relations among the properties can be marked as vector diagram, and the vector diagram will be showed to users. The vector diagram facilitates the users understanding the relations. The vector diagram of maintainability design can be marked as G' = (N', E). N' = {A1, A2, …, An} means a series of nodes, which delegate Agents; E = {e12, e13, …, eij, …}means the arcs among the nodes, which delegate the facilitations among the sub-Agents. For example, e12 means the arc between A1 and A2. In the following vector diagram, five properties are chose. It's Fig. 84.4.

If considered the facilitations among the properties, the vector diagram can be expressed as matrix F. Because the facilitation can't take place itself, elements in the main diagonal all are 0, marking as fij = 0; the matrix F is called as facilitation matrix. If only considered the optimizations of each property, the matrix H can be educed. The value of Vi is between 0 and 1. In the end, the two matrixes can be merged as one matrix M.

Table 84.1 The relations among the twelve properties

Properties of software maintainability	i	$f_{i,1}$	$f_{i,2}$	$f_{i,3}$	$f_{i,4}$	…	$f_{i,12}$
Usability	1	$f_{1,1}$	$f_{1,2}$	$f_{1,3}$	$f_{1,4}$	…	$f_{1,12}$
Clarity	2	$f_{2,1}$	$f_{2,2}$	$f_{2,3}$	$f_{2,4}$	…	$f_{2,12}$
Modularization	3	$f_{3,1}$	$f_{3,2}$	$f_{3,3}$	$f_{3,4}$	…	$f_{3,12}$
Portability	4	$f_{4,1}$	$f_{4,2}$	$f_{4,3}$	$f_{4,4}$	…	$f_{4,12}$
Testability	5	$f_{5,1}$	$f_{5,2}$	$f_{5,3}$	$f_{5,4}$	…	$f_{5,12}$
Technical review	6	$f_{6,1}$	$f_{6,2}$	$f_{6,3}$	$f_{6,4}$	…	$f_{6,12}$
Reparability	7	$f_{7,1}$	$f_{7,2}$	$f_{7,3}$	$f_{7,4}$	…	$f_{7,12}$
Quality assurance	8	$f_{8,1}$	$f_{8,2}$	$f_{8,3}$	$f_{8,4}$	…	$f_{8,12}$
Debugging	9	$f_{9,1}$	$f_{9,2}$	$f_{9,3}$	$f_{9,4}$	…	$f_{9,12}$
Simplified parallel process	10	$f_{10,1}$	$f_{10,2}$	$f_{10,3}$	$f_{10,4}$	…	$f_{10,12}$
Software component	11	$f_{11,1}$	$f_{11,2}$	$f_{11,3}$	$f_{11,4}$	…	$f_{11,12}$
Majorization and compromise	12	$f_{12,1}$	$f_{12,2}$	$f_{12,3}$	$f_{12,4}$	…	$f_{12,12}$

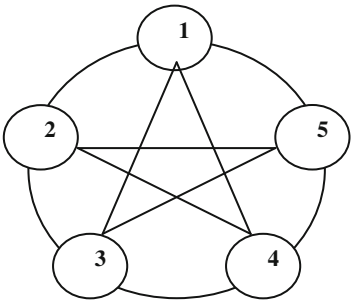


Fig. 84.4 The facilitations among the sub-agents

$$F = \begin{bmatrix} 0 & f_{12} & f_{13} & f_{14} & f_{15} \\ f_{21} & 0 & f_{23} & f_{24} & f_{25} \\ f_{31} & f_{32} & 0 & f_{34} & f_{35} \\ f_{41} & f_{42} & f_{43} & 0 & f_{45} \\ f_{51} & f_{52} & f_{53} & f_{54} & 0 \end{bmatrix} H = \begin{bmatrix} v_1 & 0 & 0 & 0 & 0 \\ 0 & v_2 & 0 & 0 & 0 \\ 0 & 0 & v_3 & 0 & 0 \\ 0 & 0 & 0 & v_4 & 0 \\ 0 & 0 & 0 & 0 & v_5 \end{bmatrix} M = [H + F]$$

$$= \begin{bmatrix} v_1 & f_{12} & f_{13} & f_{14} & f_{15} \\ f_{21} & v_2 & f_{23} & f_{24} & f_{25} \\ f_{31} & f_{32} & v_3 & f_{34} & f_{35} \\ f_{41} & f_{42} & f_{43} & v_4 & f_{45} \\ f_{51} & f_{52} & f_{53} & f_{54} & v_5 \end{bmatrix}$$

The matrix M is the matrix of software maintainability design indexes. All the effects among the sub-Agents and all the interaction relations can be expressed quantitatively. The dimension of matrix M is decided by the amount of properties. The value of the matrix means the optimization of the software maintainability design. Given all the facilitations are fixed, each scheme has the different property index Vi. After compared all the values, the Agent system choose the best scheme.

84.4 Conclusion

By introducing the Agent technique, intelligent decision-making model can rationally evaluate the affections between all indexes to select the best design scheme. The model mainly consists of two steps: The first one is Data-Mining and building information warehouse about software maintainability analysis. The second one is reasoning based on cases and evaluates the maintainability index. Giving the task to main-Agent, the main Agent communicates with the sub-Agents and allocates the task to all of them. The precedent warehouse not only instructs Agent to cooperate with other Agents, but also supplies it the datum of intelligent actions. The Agent technology applied in software maintainability design. It's proved that it's an effective way to evaluate the scheme of maintainability design.

The Agent technology is applied in software maintainability design. It's proved that it's an effective way to evaluate the scheme of maintainability design. With the development of Agent and net, the intelligent decision model based on Agent has the expansive prospect.

References

1. Humphrey, V.S.: Managing the Software Process [M]. Addison-Wesley, Reading (2009)
2. Zhou, J., Du, L.: The cost budgeting means in software project management by COCOMOII. Comput. Appl. Study **5**, 12–15 (2010)
3. Wang, X., Long, C.: The software maintenance measurement research based on ambiguous syntheses judge method. Fire Contr. Command Contr. **1**, 7–9 (2008)
4. Wang, X., Zhu, X.: Software maintenance organization management in SIS. Comput. Eng. Appl. **3**, 25–27 (2008)
5. Sharon, D.: Meeting the challenge of software maintenance. IEEE Softw. **1**, 19–22 (2006)
6. Zhou, J., Du, L.: Computer application study. The cost budgeting means in software project management by COCOMOII. **3**, 12–15 (2010)

Chapter 85
Wireless Video Transmission System Based on WIFI

Shouhuan Jiang and Zhikao Ren

Abstract Reliable video transmission is very challenging over wireless network. Efficient coding and transmission control techniques are required to meet the requirements of Qos. This paper designs the model of video transmission system based on WIFI, which achieves a real-time wireless video transmission system. The RTP encapsulation and real-time control methods of wireless video transmission based on H.264 encoding have been focused on. An adaptive control model of video transmission based on WIFI is presented in this paper. Experiment results shows that the proposed the adaptive control method of video transmission can gain good Qos so that the stability and good qualities of video transmission are ensured.

Keywords WI-FI · Streaming media · RTP/RTCP · Wireless video transmission · H.264

85.1 Introduction

Since the twentieth century, the computer network technology, modern communications technology and the artificial intelligence technology have developed greatly. The range of applications of real-time video transmission are increasingly widespread, not only requires the adaption to a known environment and the more important work environment of the future such as military reconnaissance, rescue

The Work is Supported by Shandong Provincial Natural Science Foundation, China: ZR2009BL021

S. Jiang (✉) · Z. Ren
College of Information, Qingdao University of Science and Technology, Qingdao, China
e-mail: qust008@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_85,

and relief, but also requires the detection of high pollution of the environment, anti-terrorism blasting, mine safety accident rescue in many areas.

But remote monitoring, remote education, remote medical diagnosis, remote shopping, remote access, television and telephone conference, etc. applications urgently need a network video transmission. Therefore, the research and design of a reliable wireless video transmission system is an urgent task.

Wireless video transmission based on the IEEE802.11 protocol [1], transmission control use the TCP and UDP protocols, but to some extent, it cannot meet the requirements of the wireless video transmission. So the system uses the RTP protocol [2–4] based on the application layer extended in this paper, the RTP protocol is a transmission control protocol used for real-time audio, video and other multimedia data which includes RTP and RTCP two sub-protocols: RTP is used for end-to-end transmission of real-time data, RTCP for quality of service monitoring and network diagnostics. RTP runs on top of UDP generally, both of them works together to accomplish the functions of the transport layer protocol. Its upper layer is the application layer, which mainly includes sound, video and data. RTP receives multimedia information streams (such as H.264 video) from the application layer, encapsulated into RTP packets, and then sends to the underlying UDP. In addition, RTP is an open application protocol, and the characteristic is that the basic functions can be clearly reflected in the user's program.

When the application starts the RTP session, the sender will use two adjacent UDP port number, even-numbered is used to send RTP packets, and odd numbers is used to send RTCP packets, during RTP session, participants periodically send packets, shown in Fig. 4.1. RTP itself neither provides a reliable delivery mechanism for RTP packets in order, nor provides flow control or congestion control; it relies on supporting the use of RTCP to provide these services. RTCP packets contain the number of data packets that have been sent, the situation of the missing data packets, etc. RTCP packets sent in the receive direction, which is responsible for monitoring network quality of service, communication bandwidth and transmission over the Internet, and send the information to the sender. Transmit end can use the information provided by RTCP dynamically to change the transmission rate, and even change the payload type. Use RTP and RTCP together can provide flow control and congestion control service, and it can optimize the transmission efficiency by effective feedback and minimal spending, so they are particularly suitable for real-time data transfer.

85.2 Wireless Video Transmission System Design

85.2.1 Wireless Video Transmission System

In order to meet the video transmission requirements that the occasion is not fixed or unknown environment, it is necessary to consider from two aspects: one aspect is to design a wireless video transmission link and another aspect is video capture

and encoding transmission. Wireless video transmission link comprises three categories of nodes, the first is a video capture terminal which is ended by a mobile terminal, embedded the WIFI modules and video capture encoding module with the node which is responsible for video capture, encoding compression and sub-contracting transfer; the second is a wireless AP points with embedded WIFI module, which is responsible for the wireless link repeaters and packet forwarding, the node requires several according to the actual situation; the third one is a remote control center which can be a wireless network PC, which is responsible for receiving the video data packets, sort reorganization, video decoding and display, and the control of the wireless transmission network.

In an unknown environment, in order to meet the needs of wireless network video transmission by setting wireless AP point in advance, or setting the AP point on the way according to the needs by a mobile device to set up the wireless transmission link. The structure of the wireless video transmission system is shown in Fig. 85.1.

The wireless video transmission system captures video through the video capture terminal site while doing the H.264 video compression coding, and then do the RTP sub-transmission through AP point, the control terminal receiving the video packets transmitted and decoded to restore an image display.

85.2.2 Wireless Video Transmission Scheme Analysis

Wireless video transmission is different from the wire environment, the bandwidth is stability in wire network, and change of the signal is very small, loss or disorder of data packets occur rarely, transmitting a video sequence uses I frames, B frames and P frames,as IBBBPIBBBPIBBBP… or IPPPIPPPIPPP… sequence. For each I-frame or P-frame can be carried out in accordance with the fixed-size the RTP subcontractors transmission. In the wireless network environment, the design requirement of real-time video transmission system is strict, while transmitting a video sequence with a sequence of I frames and P frames can be ok by using the RTP/RTCP protocol based on application layer extension. For each I-frame or p-frame can be transported according to a fixed RTP packets size, and also need to adjust the quantization parameter or other network-oriented video coding techniques to adapt to changes of the wireless network bandwidth and flow through the "frame skip". This may cause the out-of-order packet loss and the emergence of the "chaotic" situation, and therefore need shuffle rearrangement and RTCP feedback control messages dynamically adjust the flow and timely recovery of a frame video data, then decode and display.

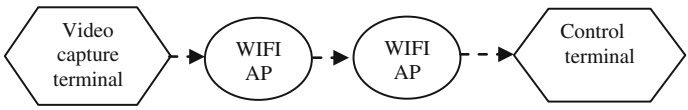


Fig. 85.1 Wireless video transmission system structure

85.2.3 Wireless Video Transmission System Structure

The video transmission system in this article is designed for wireless network environment, and is oriented towards the application., which can be designed for different structural forms facing different needs and different occasions in the practical application and development process, Here are the structure of the point-to-point video transmission system, shown in Fig. 85.2.

85.3 The Key Technologies of Wireless Video Transmission

85.3.1 RTP Packet and Its Package

Wireless real-time video transmission system uses H.264 compression coding [5–8]. The transmission control subsystem is constructed based on the RTP/RTCP protocol, completed by the UDP used by the transport layer communication. The compressed video stream does the RTP subcontract package, and then transmits to the remote through the network, separates the received RTP packet at the receiving end and remove valid data, then decodes and displays, and Receiving end sends RTCP control packet to the sender, feedbacks the situation that the network is sent, the sender end analyses and adjust send hairdo according to the network conditions. A video stream RTP packet generally consists of a fixed head of RTP, RTP payload header and RTP payload.

The RTP data packet format has some differences for the different systems under normal circumstances, for a particular system, the RTP data packet format need only two parts, namely the RTP fixed header and the payload of two parts.

Payload that is, to form the video source collected to the stream by compression coding, encapsulate the stream into RTP data packets by a cyclic process, and then

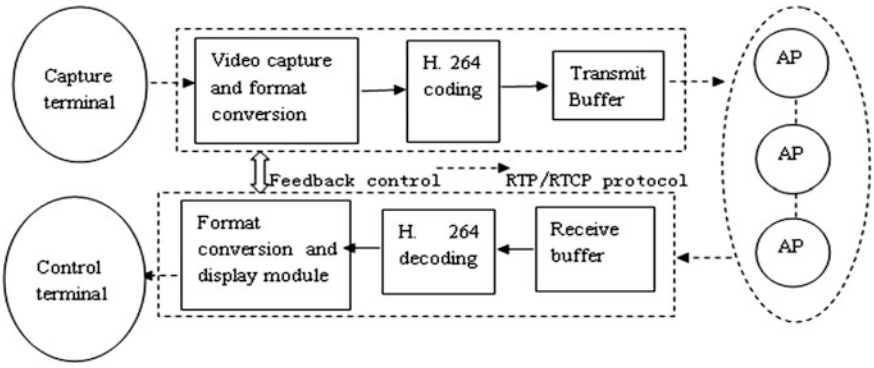


Fig. 85.2 Diagram of Transmission system

transmit to the remote. The video payload of the RTP packets can be differentiated according to the different network subcontract. The system uses IPPP4 frame loop stream format to send video sequence, a simple and effective subcontracting is to divide a frame data (I and P frames) separately into several RTP packets. A commonly divided data packets method is to divide an image of good coding compression into two packets in accordance with the parity number of the macro-block, the first packet includes all of the odd macro-block, the second packet includes all of the even macro-block, and each RTP packet must contain the header information. Complex subcontracting and stream structure combined with each other.

85.3.2 Wireless Video Transmission Control Method

In order to meet the quality requirement of real-time transmission system, uses one kind of 2 levels comprehensive control methods, video coding is divided to basic layer and expand layer in level 1, when network load is large only transmits the basic level, the basic level may meet the basic display of video, in level 2, through the feedback of RTCP and appraising the network load situations, carries on the dynamic flows control, includes several methods of "the selective sending", "adjusts the sizes of RTP data packets automatically","realigning disordered data packages".

A transmission control model is established on the second level in order to improve the stabilities of the wireless video transmission system, as shows in Fig. 85.3.

The sampling frequency is amounts of collected images of camera each second, here, presumes camera collects 25 images each second. There is not enough time to encode and transport each image in the actual transmission process, which leads to the occurring of the delay and instability as transporting images, therefore, we must associate actual situations of the network to decide the sampling frequency (i.e. jump frames processing). The control method is that: suppose gathering the image frames each second is F(x, y), the network situation weight is Q(x, y), F(x, y) initialization of F is 25, as follows:

$$F(x,y) = \begin{cases} \frac{25}{1+1/Q(x,y)} & x \neq 0 \\ 25 & x = 0 \end{cases} \tag{85.1}$$

where x represents the transmission delay of the network condition, $x = 0$ indicates that the network transmission delay is small. $x = 1$ represents the transmission delay and $x = 2$ indicates that the network propagation delays large. Where y represents disorder and packet loss, $y = 0$ represents that the network transmission disorder or packet loss did not occur or little.

$Q(x, y) > 0$, $F(x, y) \in [10, 25]$, the network real-time effect is poor if it is less than 10, the transmission quality is low, you can stop the transfer, and perform maintenance or re-commissioning solve.

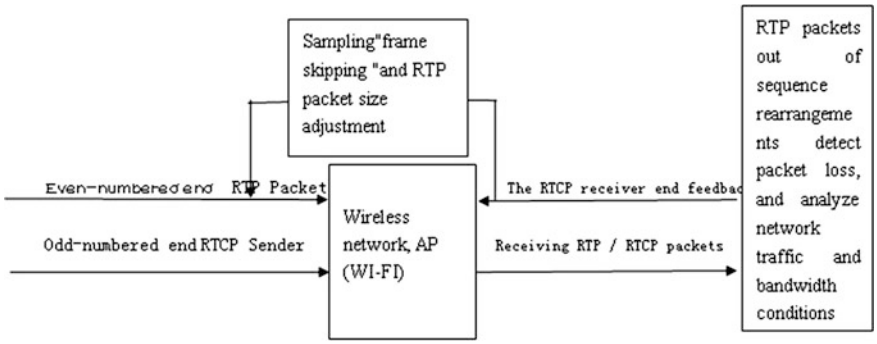


Fig. 85.3 Model diagram of the control of transmission

So we can adjust the sampling frequency timely to adapt to the current network transmission requirements through the feedback network conditions.

The system encapsulates individually each I-frame or P-frame into RTP packets in accordance with a predetermined size; Set the initial value for each RTP payload to 1024 K, and then change flexible according to the effect of the actual network transmission when the network is inefficient. If it is found that there is a serious disorder and frequent packet loss occurring at receiving end, need to control the sampling frequency (skip frames), at the same time change the packet size of the RTP data packet to achieve good implementation of real-time transmission by adjusting. Set RTP data packet size as S, then:

$$S(x) = \begin{cases} \frac{2048}{1-F(x,y)/Q(x,y)} & x \neq 0 \\ 2048 & x = 0 \end{cases} \tag{85.2}$$

where Q(x, y) reflects the degree of congestion in the network. S(x) ∈ [2048, 4096].

The general single RTP packet does not exceed 4096 K according to the experimental analysis, so you can have a better quality of transmission in different network environments.

The key of the problem is how to identify a suitable network transmission quality reference weight by analyzing the receiver situation at the receiving end. And then fed back the reception conditions and analytical results to the transmitting side by the RTCP packet of the sending and receiving end to adjust the way of the sending end. This process is a dynamic equilibrium process is also the quality control process of the entire transmission system.

$$Q(x,y) = \sum_{c=1}^{n} \frac{\sqrt{x+y}}{\sqrt{x2+y2}} \quad x = 0, 1, 2, \ldots \tag{85.3}$$

$$y = \textit{packet loss or the number of out of order}$$

where n may have different values according to the different requirements, the meaning is to examine the receiving condition of n times of the receiving end n times, the present system n = 5.

85.4 The Experimental Results and Analysis

When the wireless video transmission starts, the collection terminal and the control center connect through the wireless AP link firstly. After connection, the transmission side starts the video capture, while compressing encoding the video, outputting the I frame eligible for a P-frame image, RTP sub-packaging the compressed image information, and is formed to the RTCP sender report, transmitted to the receiving end, and then send RTP subcontractors one by one until an information transmission is completed, and then proceed to the next cycle.

The control terminal receipts that the RTP packets are transmitted by the collection side and stores to the corresponding location of the cache according to the rearrangement algorithm, calculates the situation of the disorder and the packet loss combined with analysis of previously-received RTCP sender report, and then on one hand deals with the missing data packet through interpolation treatment methods,recoveries to form a code stream, and decodes the output and displays; on the other hand, forming a receiving side report and feedback so that it is easy to control the traffic or other measures for the sender.

The rearrangement algorithm as follows:

Firstly, the smallest serial number minsequence of an image subcontracted is obtained by the formula:

(Current packet sequence number - the smallest serial number) * the length of the packet;

receivestru [i]. sequ- minsequence.

For (i = 0; i < jjj; i + +).

*{memcpy (cc + (1024 * (receivestru [i]. sequ-minsequence)), receivestru [i]. receivedata, eceivestru [i]. length);} receivestru[i].sequ is the current serial number.*

Determine the address, the receiving data would be write into the corresponding memory space, and be decoded to restore an image data, and then display. Control the transmission quality timely, such as frame skipping and control packet-divided size according to the RTCP packets feedback by the wireless network receiving terminal. The achievement of skipping is: when initialization, *F (0,0) = 25*, corresponding *sleep (0)*; control skipping by the changes of the parameter t of *sleep (t)*. When the network changes, it gets the control parameters Q through the feedback, and dynamically adjust skipping based on network conditions Q in the receiver. In addition, a simple control is fault-tolerant can skip and receive another I frame again from the next cycle if an error occurs when the receiver receives I frames, this fault-tolerance is very convenient, and also very practical.

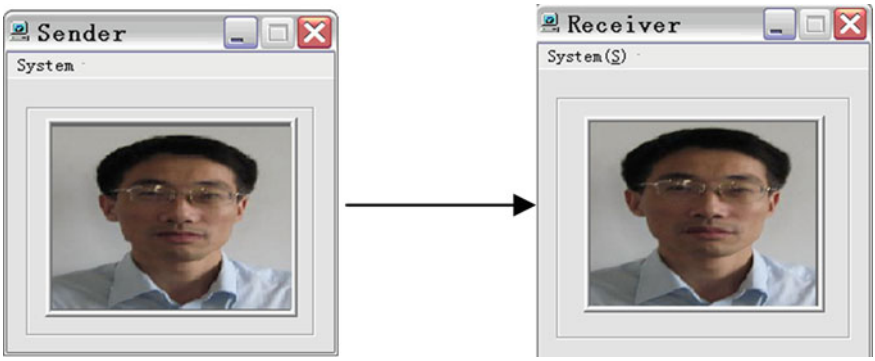


Fig. 85.4 The interface of real time transmission system

Table 85.1 Comparsion of real time transmission system

Performance control	Frame rate	Mosaic	Delay	Video background mobile condition
No quality control	3 ~ 20 frame/s	Often appears	Often happen	The image is very unstable
A quality control	10 ~ 25 frame/s	Basic no	Seldom occurs	The image stability, good adaptability

Wireless video real-time transmission system interface as shown in Fig. 85.4, through wireless network in the experiment many times, some conclusion shown in Table 85.1, jump frame phenomenon is obvious in the wireless network environment, transmission frame rates generally keeps in 10 ~ 15 frame/s. The effect is very good when the video background fast moving, operation complexity increases, the transmission frame rates need be reduced, and adjusts frame rates in time to ensure the stable quality of transmission.

85.5 Conclusion

This paper presents a model of real time video transmission system based on WIFI. It is concluded that the demand of the transmission performance is higher and transmission quality control is important in wireless network environment by several experiments,. It is very difficult to guarantee the quality of the transmission if there is no flexible control method. RTP packets size divided method can adapt to the dynamic changes of the wireless network. This system has practical applications mean, such as in unknown environment detection, pollution of the environment live video transmission, etc., and can be developed and applied under DSP environment, such as ROV and exploration robot. In a word, the design of the system adopted some new methods: (1) the fault-tolerant; (2) the two levels

treatment structure; (3) comprehensive control system, including the "frame skipping", RTP packets size appropriate adjustment, etc. These can meet requirements of the wireless video transmission based on WIFI basically.

Acknowledgments The work is supported by Shandong Provincial Natural Science Foundation, China:ZR2009BL021

References

1. Pan, T., Liu, X.: Hybrid wireless communication system using ZigBee and WiFi technology in the coalmine tunnels. In: 2011 Third International Conference on Measuring Technology and Mechatronics Automation. 978-0-7695-4296-6/11(2011)
2. Shi, B., Wang, Z., Liu, J.: Research and realization of real-time remote image transmission based on RTP, Microcompute information[J], **21**(2), 178–179 (2005)
3. Schulzrinne, H.: RTP Profile for Audio and Video Conferences with Miniimal Control[S]. Internet RFC 1890(01) (2006)
4. Peng, W.L The Realization and Study of the Real time transmission system Based on RTP, the paper of master'degree (2005)
5. Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG. Draft ITU-T Recommendation and Final Draft international Standard of Joint Video Specification (ITU-T Rec. H.264 I ISO/IEC 14496-10 AVC). JVT-G050r1 (2003)
6. Wiegand, T., Schwarz, H., Joch, A., Kossentini, F., Sullivan, G.: Rate-constrained coder control and comparison of video coding standards. IEEE Trans. CAS Video Technol. **7**(13), 688–703 (2003)
7. Wiegand, T., Sullivan, G., Bjontegaard, G., Luthra, A.: Overview of the H.264/AVC video coding standard. IEEE Trans. CAS Video Technol. **7**(13), 560–576 (2003)
8. Pan, F., Lin, X. et. al.: Fast Mode Decision for Intra Prediction. Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT-G013 (2003)

Chapter 86
The Self-Adapted Taxi Dispatch Platform Based on Geographic Information System

Yi-ren Ding, Jing Xiong and Heng-jian Liu

Abstract In order to improve the efficiency of taxi dispatching, we build a center-controlled, real-time management system. With the help of the real-time data collected by recorders in taxis and the wavelet neural network utilized to predict passenger current, the whole system can work more precisely. Besides, the exceptional situations are also taken into consideration in this system. Thus, the whole system is able to distribute taxis efficiently in any situation. Simulation results indicate that the wavelet neural network could make more accurate prediction than former methods and the self-adapting distribution strategy can increase load rate effectively.

Keywords Taxi dispatching · Passenger flow prediction · Dynamic model

86.1 Introduction

Taxi is and still will be an important part of city transportation system; it is a supplement of public transportation [1]. The improvement of the efficiency of taxi has taken paces with the help of vehicle GPS and center GIS control system. Take Taipei as an example, the taxi information sharing system established several years ago provides a platform for strangers who have the same destination to take one cab together, which improves the efficiency of the city transportation [2]. The system introduced in this passage, committed to analyzing the data from recorder in taxis, establishing self-adapted dynamic model in consideration of emergency, and predicting passenger flow in each road. The application will be given in passage.

Y. Ding (✉) · J. Xiong · H. Liu
Faculty of Information Engineering, China University of Geosciences, Wuhan, Hubei, China
e-mail: junecug@gmail.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_86,

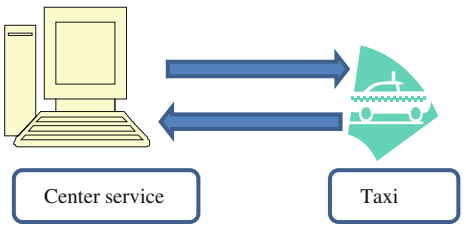


Fig. 86.1 System configuration

86.2 Geographic Information Management System

The real time information (route, amount of passenger) which comes from recorders in taxi is transmitted to the center. The management center not only receives the information of current passenger flow but also receives the location and state of every taxi. All information comes from taxi will be saved in server, which will be invoked as original data for real-time analysis, then updating dynamic model and giving guidance. It is an archetype of intelligent transportation cloud [3]. The FLEX[1] is used for building the interface and Java is applied for calculating. Besides, the MapGIS (see footnote 1) platform is utilized for map service (Fig. 86.1).

86.3 The Wavelet Neural Network

The wavelet neural network is one kind of the former dyke type models for prediction. The wavelet neural network, with the ability of self—learning and self-revising, is the most comprehensive and typical model of neural network. The learning processes of neural network are made up of forward propagating and counter propagating. Only if the result of first forward process didn‘t meet the expectation will the counter propagating interfere and revise weight in each neural cell, then the new results could be more precise. The advantage of wavelet in information process could be fully developed if it is used as transmit formula in neural network system. And it could overcome the nonlinearity and the interactive feedback attribute of taxi [4]. Based on published experiment, the accuracy of wavelet neural network model is better than GNF and NBRR model which are most common in use today [5] (Fig. 86.2).

The number x formula of wavelet transmitting is showed:

---

[1] Notifications

1 FLEX: A free and effective open source software platform produced by Adobe™ for application development.
2 MapGIS: The first erect-style data center development platform in the world produced by ZONDY CYBER.

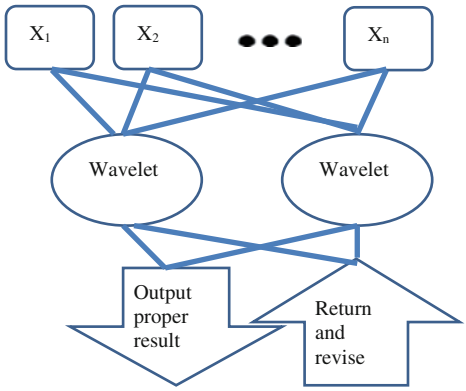


Fig. 86.2 Wavelet neural network

$$M_j = M_j\left(\frac{\sum_{i=1}^{k}\omega_{ij}x_i - a_j}{b_j}\right) \mathrm{j} = 1, 2, 3\ldots \tag{86.1}$$

$\mathrm{j} = 1, 2, 3\ldots$ are indicate factors in the wavelet formula. These factors are supposed to be modified only if the first output value exceeds the tolerable difference of expectation.

The method of modifying is called gradient-corrected.
Weights modifying process:
Modify weight of each level: $\omega\mathrm{k} + 1\mathrm{nj} = \omega\mathrm{knj} + \Delta\omega\mathrm{k} + 1\mathrm{nj}$
Modify the factors in the number I neural cell:

$$\mathrm{ai} + \mathrm{lj} = \mathrm{aji} + \Delta\mathrm{aji} + \mathrm{l};\ \mathrm{bji} + 1 = \Delta\mathrm{bij} + 1 \tag{86.2}$$

In this formula $\Delta\omega\mathrm{k} + 1\mathrm{nj}, \Delta\mathrm{aji} + 1, \Delta\mathrm{bij} + 1$ are obtained by gradient-corrected method [3]:

$$\Delta\omega_{nj}^{k+1} = -\sigma\frac{\partial e}{\partial\omega_{nj}^{k}}; \Delta a_j^{i+1} = -\sigma\frac{\partial e}{\partial a_j^{i}}; \Delta b_j^{i+1} = -\sigma\frac{\partial e}{\partial b_j^{i}} \tag{86.3}$$

After the model is established, to increase the speed and efficiency of learning of neural network, the pretreatment is necessary. We use linearization method to narrow the range of data, then the possibility of successful prediction is likely to be improved [4, 6]. Linearization formula:

$$\mathrm{Xi} = (\mathrm{x} - \min(\mathrm{X}))/\max(\mathrm{X}) - \min(\mathrm{x}) \tag{86.4}$$

Table 86.1 Exceptional situations

Name	Anomalies	Example
One-time output anomaly	A place with unusual output passenger for one time	A train station right after long holiday
One-time input anomaly	A place with unusual input passenger for one time	An international auto show
Long time anomaly	Unusual passenger flow for long time	New shopping mall

86.4 Exception Handing

It is an effective method to find regular fluctuation of passenger flow through the result of wavelet neural model. However, the exceptions caused by unexpected events cannot be predicted by the model. Besides, due to the accumulation of data, the abnormal data may influence the short-time prediction of passenger flow. Therefore, manual adjustment is required when the abnormal data is detected, and the cause of anomaly should be investigated and analyzed [7].

We classify exception into three categories (Table 86.1).

86.4.1 The One-time Output Anomaly and The One-time Input Anomaly

In these two situations, the anomalies would last for one day or several days, usually accompanied with the abrupt appearance and fast disappearance. Holidays, festivals, celebrations and grand activities always contribute to the situations. The method to detect the anomaly is called growing-rate threshold. In general, the cycles of statistics are set each 6 h, one day, and three days. If the growth rate exceeds the threshold of each time cycle, the system will inform managers to change the auto-dispatch to manual-dispatch and then experts are informed to investigate the causes.

86.4.2 Long-time Anomaly

The changing rate of data in long-time anomaly situation is not as sharp as one-time anomaly. But the difference is that the passenger flow will never come back to the former level. Like the new shopping mall put into use, it will lead the long-standing change in passenger flow which is different from one-time anomaly. In this case, the cycles of statistics are 15 days, one month, three months and one year. When the long-term anomaly is detected, the old data will be cleared and

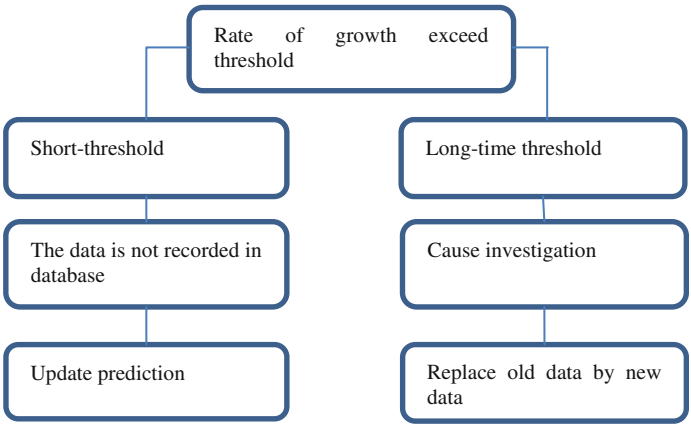


Fig. 86.3 Exceptional handling process

train the newly collecting data in model. And this data replacement makes the prediction more accurate (Fig. 86.3).

86.5 Real-time Self-adapting Dispaatching

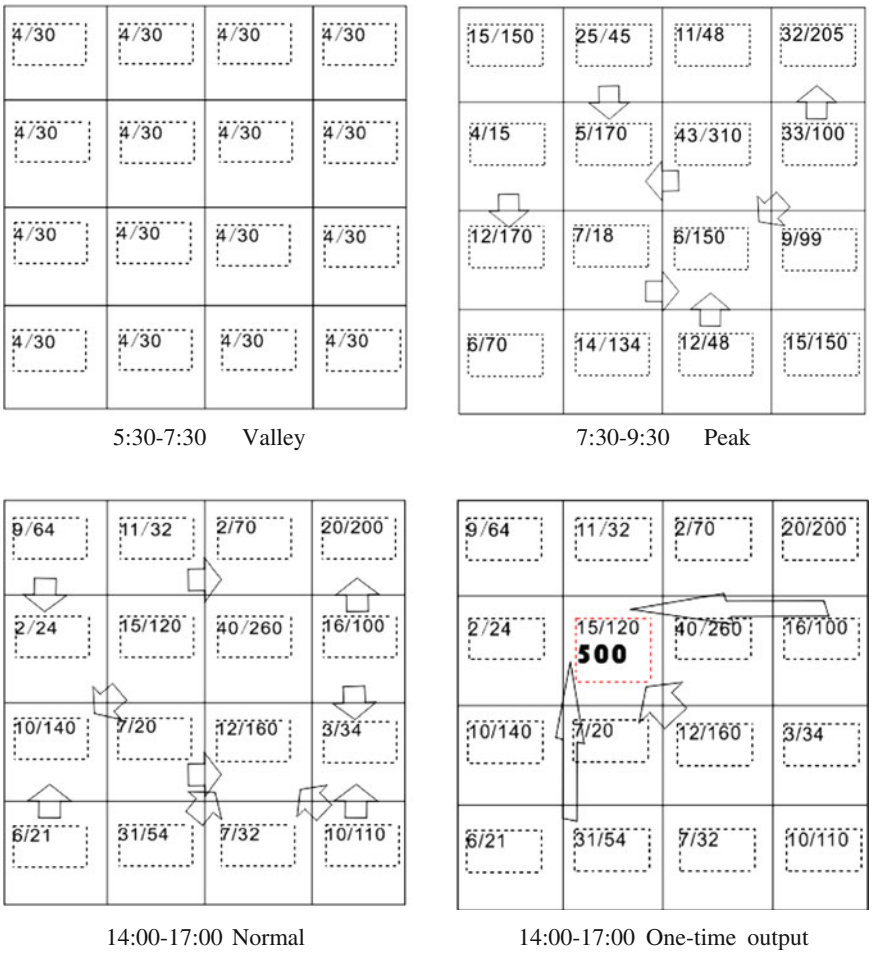
The passage above simply introduces the system configuration, prediction formula and exception handling, the key part of system is to dispatch taxi by the result that is calculated by the model and the real-time data. Real-time dispatching is made up of three core parts: the model of passenger flow prediction, the real-time statistic data, and the "back much better repair" policy. This policy means to predict potential amount of passengers according to the difference between statistic number and prediction number.

Because of the uncertainty of the passengers amount, and the wavelet neural model can only predict the total amount of passengers, the system obtains many unsatisfactory results during tests. We find that even the amount of total passengers has been predicted precisely by wavelet network model, the passenger flow fluctuate in every period—the surge in rush hour. And the increase of passengers in rush hours makes counteraction on steady phase. So the prediction should be adjusted by the data we collected in 1 day.

Assuming Pn is the passenger prediction in one period, T means the total amount of passenger before time Pn, S means total prediction number, Xn means the impact factor.

$$P_n = S * X_n * \left(\sum_{i=1}^{n-1} P_{n-1} \right) / T \tag{86.5}$$

After the passenger flow prediction in current period is obtained by using GPS locating the position of every taxi, then we can give guidance to empty taxis. As the suggestion given by construction bureau, the rational rate of utilization of taxi is 70 %. And the rate of utilization approximately equals to the ratio of the amount of passengers and the number of taxis in one time period [8, 9]. Because the system has not been put into use, in the example followed, we assume that the rational ratio of passenger and taxi is 10.



5:30-7:30 Valley

7:30-9:30 Peak

14:00-17:00 Normal

14:00-17:00 One-time output

Fig. 86.4 Passenger, taxi and dispatching situation in city one

Table 86.2 Rate of load in three cities

Rate of Load	Valley time (%)	Peak time (%)	Normal time (%)	Anomaly in normal time (%)
City 1	65	94	75	86
City 2	60	90	68	55
City 3	39	82	50	45

86.6 Example of Application and Contrast

86.6.1 One Particular Simulation

In order to test the superiority of our system, we create three simulated cities and each city is tested for three times with different original status. Take city A for example, to simulate the rush hour and normal hour, we suppose there are 16 fixed blocks in the city and the numbers of passenger range from 500 to 2,500. The data is recorded in representative time, and it includes the number of potential passenger, the number of taxi, the number of people who took taxi and people who failed to take taxi in each block. To compare the efficiency of rate of load, city A will simulated under three different management modes.

Mode1: Using self-adapted dispatching system.
Mode2: Using normal dispatching system.
Mode3: Without using dispatching system.
The X/Y in the graph means: amount of taxi in the area/passenger prediction in the area.
The arrow means the guidance given to the taxi.
The bold number in fourth picture means a one-time output anomaly; the number of real passenger is over 4 times larger than prediction.The following pictures show the dispatching process in self-adapted management system (mode 1) (Fig. 86.4), (Table 86.2).

Table 86.3 Rate of load in each simulation

	Mode 1 (%)	Mode 2 (%)	Mode3 (%)
City A	80	68	54
	83	69	60
	80	74	62
City B	86	74	62
	81	69	65
	82	70	59
City C	78	69	58
	78	63	53
	77	65	53
Average	80.6	69	59.6
Standard deviation	2.83	3.60	4.33

Through this simulation, we obtain the results of rate of load in different management modes. In mode 3, the rate of load is always lower than that of mode 1 and mode 2. In mode 2, though the rate of load surpasses that in mode 3 a lot, it decreased sharply when an anomaly appears. In city 1, with the help of self-adapted management system, not only the rate of load is the highest, but also the fluctuation is the tiniest, and the advantage of stable is particular significant when anomaly happens.

86.6.2 All Simulations

City B is supposed to have 25 blocks and City C is supposed to have 36 blocks. Each city is tested three times with different original status.

The following table shows the average rate of load in each simulation that we have done (Table 86.3).

86.7 Conclusion

With the help of our system, three goals are achieved. Firstly, managing taxi cab in center controlled method. Secondly, predicting the amount of potential passenger precisely. Thirdly, providing proper suggestions to taxi. Simulation result shows that the self-adapted management improves the rate of load by approximately 10 %, it provides better Stability. Center controlled system and cloud system is the tendency of management system. The communication technique and GIS platform contribute a lot to establishing this system. However, this system only predicts and monitors the passengers who take taxis. The next step of studying should refer to the system which could control all public transportation system. In that way, the efficiency of city transportation could improve much more.

Acknowledgments The project was supported by the Fundamental Research Founds for National University, China University of Geosciences (Wuhan) 1210491B08.

References

1. An, S., Qi, L.: Assessment and prediction of the number of taxicab. Technol. Econ. Areas Commun. **3**, 34–37 (2010)
2. Tao, C.C.: Dynamic taxi-sharing service using intelligent transportation system technologies. Wirel. Commun. Netw. Mobile Comput. Conf. **5**, 3209–3212 (2007)
3. Qian, S., Zhang, Y., Huang, S.: Intelligent transport clouds: ITS based on cloud computing. Comput. Modern. **183**, 168–171 (2010)

4. Yi, L., Wang, L.-z, Lu, X.: Study on the simulation and prediction model for urban Taxi's demand. J. Changsha Commun. Univ. **23**(4), 23–27 (2007)
5. Ren, C., Cao, C., Li, J., Shi, W.-W.: Research for short-term passenger flow forecasting based on wavelet neural network. Sci. Technol. Engi. **11**(21), 5099–5103 (2011)
6. Li, S., Jiang, H., Wang, F. Application of heredetary neural net in public transport passenger volume forecast. Commun. Stand. **160**, 161–165 (2006)
7. Hang, J., Han, B.: Comprehensive post-evaluation for the passenger prediction of urban rail based onPSR. Urban Mass Transit. **8**, 31–35 (2011)
8. Wang, J., Lu, X., Li, J.: Multi-Layer BP neural network comprehensive evalution on the ability ot interated passenger transportation on the urban and rural roads. Dystems Eng. **25**(3), 83–86 (2007)
9. Zhang, Y.: Road transportation systems engineering. China Commun. Press **5**, 101–105 (2004)

Chapter 87
Contract-Based Combined Component Interaction Graph Generation

Haiqiang Li, Min Cao and Jian Cao

Abstract Component interaction graph is used to describe the interrelation of components, which provides an objective basis and technology to test component composition. However, the traditional component interaction graph cannot serve as a basis to test a component itself and the state transition between components for lacking of description of states of individual component. Therefore, a novel model, named Contract-Based Combined Component Interaction Graph (CBCCIG) is put forward in this paper. CBCCIG combined the thought of contract test with the UML state diagram which is introduced in the paper. The proposed model can not only support the quick assembly of the software system depending on developer's own willing, but also the automatic or semi-automatic generation of test cases which are the state transition and information interaction between components. Thus, CBCCIG improves the efficiency of development and testing.

Keywords CBSD · Component composition · CIG · UML state diagram · Testing by contracts

87.1 Introduction

Component-based Software Development (CBSD) is an effective and efficient approach to improve the productivity and quality of software development [1]. In CBSD, the most important thing is how to obtain suitable components and integrate them to product a reliable software system. Component composition is a

H. Li (✉) · M. Cao
School of Computer Engineering and Science, Shanghai University, Shanghai, China
e-mail: lhqmaillove@163.com

J. Cao
School of Physics, Nankai University, Tianjin, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_87,

critical process which determines whether your CBSD can acquire, reuse, or build a component. Due to different component versions, different component technologies, and different integrated environments, there is no mature technical standard and feasible method to capture the mistakes in the integration testing [2]. CBSD improves the efficiency of software development, but it brings the testing difficulties at the same time.

Component Interaction Graph (CIG) is used to describe the interrelation of components, which provides an objective basis and technology for the implementation of component composition testing [3]. Nowadays, there are many researches on the generation of CIG. Ye Wu et al. [4] presented a method to construct the CIG in which the interactions and the dependency relationships among components are illustrated. By utilizing the CIG, they propose a family of test adequacy criteria which allow optimization of the balancing among budget, schedule, and quality requirements typically necessary in software development. Based on the direct and indirect correlation analysis, CIG was established [5]. By using the component specification structure and the established CIG, the component interactions can be modeled to provide support for testing component-based software. Lun [6] represented software architecture possessing C2 style through CIG, and abstracted the behavior of interactive between components and connectors, then they defined three testing criteria and introduced algorithms to generate testing coverage set according to edge types of CIG.

The above methods provide theoretical and experimental basis for the generation of CIG. However, these traditional CIGs do not describe the state of individual component. They cannot serve as a basis for testing a component itself and the state transition between components. In this paper, a component is firstly represented in the form of UML state diagram. Then, synthesizing CIG and UML state diagram, we propose a novel model, named (CBCCIG), to generate test cases of the state transition between components and information interaction.

87.2 Related Concepts

87.2.1 Component and Component Composition

Component interfaces are the access points of components, through which a client component can request a service declared in an interface of the service providing component. Each interface is identified by an interface name and a unique interface ID.

Definition 1 A component is a 2-tuple C = (P, R), where:

- $P = \{P_1, P_2, \ldots, P_n\}$ is the set of providing services interface.
- $R = \{r_1, r_2, \ldots, r_n\}$ is the set of required services interface.

In this case, collections and the elements in the collection are represented by capital letters and small letters respectively.

Definition 2 Component composition: The composition of two component means that the required services of one component are provided by another partly or fully.

87.2.2 Testing by Contracts

A contract is a stipulation between two parties, containing benefits and obligations for each part. Design by Contract (DBC) is an object oriented design technique that ensures high-quality software by guaranteeing that every component of a system lives up to its expectations [7]. Under the design by contract theory, a software system is viewed as a set of communicating components whose interaction are based on precisely defined specifications of the mutual obligations–contracts. Every good contract entails obligations as well as benefits for three parties: (1) the precondition; (2) the post-condition; (3) the invariant.

87.3 Contract-Based Combined Component Interaction Graph

87.3.1 CIG

87.3.1.1 Semantic Analysis

A CIG is a directed graph which is used to depict interaction scenarios among components. The major elements related to interactive feature are interfaces, events, context dependence and content dependence [4].

- Interfaces:
  Interfaces are the basic access means via which components are activated.
- Events:
  We define an event as an incident in which an interface is invoked in response to the incident. The interface events defined in the CIG are usually methods.
- Context dependence:
  One event has a context dependence relationship with the other event if there exists an execution path which triggers one event directly or indirectly.
- Content dependence:
  The content dependence relationship is defined as follows: a function (named functions 2) depends on another function (named function 1) if the value of a variable defined in function 1 is used in function 2.

87.3.1.2 Mathematical Definition

Definition 3 A CIG is a 2-tuple CIG= < V, E > , where:

- V = VCUVE is a set of nodes. Accordingly to definition 1, VC = (P, R) is the set of component interface nodes, VE is the set of event caused by component.
- E = ECUED represents a set of directed edges. EC represents context dependence, ED represents content dependence.

If there is an existing edge form C1.P1 to C2.R1 in the CIG, it means the required service R1 of C2 has been satisfied by the providing service P1 of C1, namely, C2.R1 = C1.P1. We denote an interface with an ellipse and a component with square, and the interfaces belong to one component that was drawn in the same square. Then the CIG is built as follows shown in Fig. 87.1.

87.3.1.3 Effects and Problems of CIG

Component-based software is often built through component composition. Component interfaces are the access points of components and define all content of interaction with the external. The only way that a component communicates with the outside is component interface. We modeled component interaction by establishing CIG, which can describe the interaction semantic better and also provide support for testing component-based software. At the same time, the test model can be useful to explain interactive and dependent relationship between components. Both the direct and indirect interaction relationship between components, based on which the test cases are chosen, by traversing the CIG. However, in the traditional component interaction graph, components are presented in the form of interfaces, which does not describe the state of individual component. CIG cannot serve as a basis for testing a component itself and the state transition between components. Therefore, we introduce UML state diagram to represent the state of components. The combination of CIG and UML state diagram can be used to generate the test cases of the state transition and interaction between components.

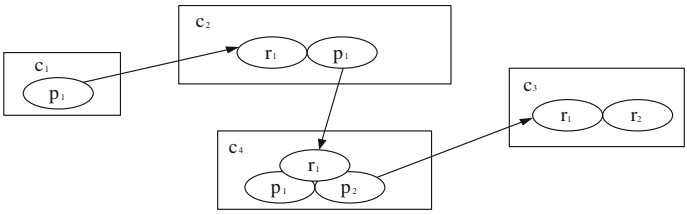


Fig. 87.1 Component interaction graph

87.3.2 UML State Diagram

87.3.2.1 Semantic Analysis

UML state diagrams depict the various states that a specific component may be in and the transitions between those states, which will be modified by events. UML state diagram consists of states, transitions, events and actions [8].

- States:
  States are defined as a condition in which a component is in. State will change when some event is triggered.
- Transitions:
  A transition is a progression from one state to another, triggered by an event which is either internal or external to the component. It also causes an important change of state.
- Events:
  Events will cause some actions and the transitions between states. Generally, they are method invocations.
- Actions:
  An action is an operation of an active component. When an event is dispatched, the component responds with performing actions, which cannot be interrupted.

87.3.2.2 Mathematical Definition

Definition 4 A state chart diagram is a 5-tuple $\mathrm{SD} = (\mathrm{S}, \mathrm{E},\ \mathrm{F}, \mathrm{s}_0, \mathrm{s}_\mathrm{F})$ where:

- $\mathrm{S} = \{\mathrm{s}_0, \mathrm{s}_1, \ldots \mathrm{s}_\mathrm{i}, \ldots, \mathrm{s}_\mathrm{n}\}$ is a finite set of states, where $\mathrm{i} \in (0, \mathrm{n})$.
- E is a finite set of event driven of state chart diagram.
- F is a finite set of transitions. $f : \mathrm{S} \times \mathrm{E} \rightarrow \mathrm{S}, f(q, e) = p$, where $\forall \mathrm{e} \in \mathrm{E}$.
- $\mathrm{s}_0 \in \mathrm{S}$ is an initial state. A SD must have one and only one initial state.
- $\mathrm{s}_\mathrm{F} \subseteq \mathrm{S}$ is a nonempty set of final states.

87.3.2.3 Effects

In our model, we represent the components of CIG in form of UML state diagrams. The proposed approach indicates not only the state transition of a component itself, but also the state transition between components. Therefore, Combined Component Interaction Graph (CCIG) is putting forward by combining CIG and state chart diagram. CBCCIG introduces the thought of contract in CCIG.

87.3.3 CBCCIG

87.3.3.1 Thought and Semantic Analysis

In CIG, the interaction among components can be described as directional arrows, with the providing service interface points to required services interface. We define an event as an incident in which an interface is invoked in response to it. They are usually methods. Also, the state transition of UML state diagram can be described as directional arrows, with initial state points to final state. The arrows with state input information are methods too. Therefore, CIG and UML state diagram in the same system can be combined. We represent the components of CIG in form of UML state diagram, and combine the special features of these two models.

In our model, there is a one-to-one mapping between the interfaces of CIG and the states of UML state diagram. At the same time, we analyze the transition arrows with method name and finally form the CBCCIG.

87.3.3.2 Mathematical Definition

Definition 5 A CBCCIG is a 5-tuple CBCCIG = (C, CS, CE, CF, CG), where:

- C is a finite set of all components.
- CS is a finite set of states of C. $c_i s_j$ represents the sate j of component i.
- CE = (precondition, E, postcondition) represents a set of events of C. We have described the concepts of event in the previous chapters in the component interface and state transition manner. However, there are no restraint conditions to guarantee the proper operation of the method, such as the accuracy of input parameters and return results. Therefore, in order to ensure the accuracy of the interaction and connection among components, we introduce the basic idea of contract testing, and add some constraint rules like precondition and postcondition to event.
- $s_0 \in S$ is an initial state. A SD must have one and only one initial state.
- CF is a finite set of transitions between component states. $f: CS \times CE \rightarrow CS, cf(c_i s_j, e) = c_p s_q$ represents the transition from $c_i s_j$ to $c_p s_q$, where $\forall e \in CE$.
- $CG \subseteq CS$ called intermediate state. These states neither cause an event to interact with another component actively nor need the service provided by component itself or other components.

87.3.3.3 Generation Algorithm

Definition 6 The abstract mapping between interface and state: let a component c_1 be at state s_1, and a component c_2 be at $states_2$. If c_1 is triggered by an event and

it reaches at state s_2 automatically, we definite s_1 as an interface which provides service and in reverse s_2 as an interface which requires service.

According to above definition, we present a way to generate the CBCCIG.

Algorithm:
Components $C=\{c_1, c_2, \cdots, c_i, \cdots, c_n\}$
States $S=\{s_0, s_1, \cdots, s_j, \cdots, s_m\}$
1. Begin
2. Remove the states which switch between the components.
3. For i=1 to n
4. For j=0 to m
Search a state s_j from component c_i from which, occurring of an event activate another component.
5. Set s_j as p_j.
6. EndFor
7. EndFor
8. For i=1 to n
9. For j=0 to m
10. Search a state s_j from the component set which require services.
11. Set s_j as r_j.
12. EndFor
13. EndFor
14. If a state is not $p_j \| r_j$
15. Then set it as an intermediate state G_j.
16. EndIf
17. End

Figure 87.2 is an incorporative CIG, in which a rectangle, a circle and a square denotes component, component state and component interface respectively. In addition, the solid line represents the state transition of individual components, and

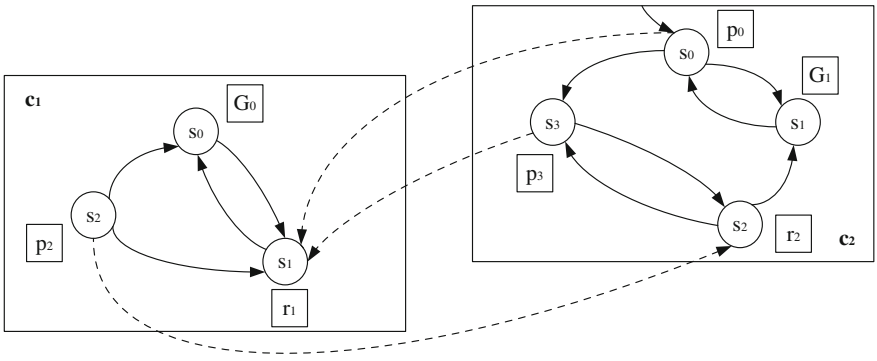


Fig. 87.2 Combined component interaction graph

the dotted line represents the state transition among components. p is the providing service interface, while r is the requiring service interface.

87.3.4 Research Significance of the CBCCIG

The core technology of CBSD is component composition. A large number of components need to be assembled together for a complex system. Most of the component composition testing is realized with the help of the combination of the component models, for example the CIG which provides an objective basis and technology for the implementation of the component composition testing. In this paper, the proposed model, named CBCCIG, not only contains the assembly and interactive elements that the development model needed, but also increases the testing elements, such as state, contract and so on. The model fully plays the role of testing in CBSD. Test driven component composition treats the test as the center, thus makes every step of the development process have the measure criteria. At the same time, with this model framework, we can not only assemble the software system according to own willing quickly, but also generate test cases of the state transition between components and information interaction automatically or semi-automatically. Both the development and testing efficiency are improved under this novel model framework.

87.4 Conclusion

Take CIG and UML state diagram together, and also consider about the thought of contract test, this paper proposes a new type of CBCCIG model, which will be much helpful with the generation of test cases of the state transition between components and information interaction. The future work includes the development of a tool to support automation of the CBCCIG generation, and automated generation of test cases from this model.

References

1. Shang, M., Wang, H., Jiang, L.: The development process of component-based application software. In: 2011 International Conference of Information Technology, Computer Engineering Management Sciences, pp. 11–14 (2011)
2. Fu, L., Sun, G., Chen, J.: An approach for component-based software development. In: International Forum on Information Technology and Applications, pp. 22–25 (2010)
3. Li, L., Wang, Z., Zhang, X.: An approach to testing based component composition. In: International Colloquium Computer Communication, Control Management, pp. 735–739 (2008)

4. Wu, Y., Pan, D., Chen, M-H.: Techniques for testing component-based software. In: Proceedings of the Seventh IEEE International Conference on Engineering of Complex Computer Systems, vol. **2**, pp. 222–232 (2001)
5. Cao, W., Zhang, W., A software test method based on CBD. Comput. Sci. **2,** 156–158 (2005) (Chinese)
6. Lun, L., Chi, X.: Software architecture testing in the C2 Style. In: 2001 3rd International Conference Advanced Computer Theory Engine, pp. 123–127 (2010)
7. Valentini, E., Fliess, G., Haselwanter, E.: A framework for efficient contract-based testing of software components. In: Proceeding of the 29th Annual International Computer Software and Applications Conference, pp. 219–222 (2005)
8. Mohanty, S., Acharya, A.A., Mohapatra, D.P.: A model based prioritization technique for component based software retesting using UML state diagram. In: International Conference Electronics Computer Technology, pp. 364–368 (2011)

Chapter 88
An Improved Weighted Averaging Method for Evidence Fusion

Ye Li, Li Xu, Yagang Wang and Xiaoming Xu

Abstract D-S evidence theory is an important mathematical tool for uncertainty reasoning. However, it may lead to counterintuitive conclusions when combining conflicting evidences. In order to overcome this disadvantage, one can modify the evidences before Dempster's rule of combination. One representative method is to assign a weight to each evidence according to its credibility degree based on the concept of distance (or similarity) between two evidences. This method can gain more robust fusion results than many other known methods. However, it may fail to correctly converge according to the cardinality of the sets in the evidence. When evidence conflicts with other evidences, the evidence may lose impact on the combination result. Moreover, the combined mass is nonmonotonic even though evidence varies monotonically. Therefore, the method still leads to counterintuitive or confusing results. This paper brings forward an improved weighted averaging method involving a new similarity measure between evidences and a new combination rule. The numerical examples show the proposed method well solves the above problems.

Keywords Data fusion · Evidence theory · Conflicting evidence · Combination rule · Evidence distance · Evidence similarity

Y. Li (✉) · Y. Wang · X. Xu
School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, China
e-mail: liye@usst.edu.cn

L. Xu
Department of Electronics and Information Systems, Akita Prefectural University, Akita, Japan
e-mail: xuli@akita-pu.ac.jp

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_88,

88.1 Introduction

D-S evidence theory is first proposed by Dempster [1] and later developed by Shafer [2]. It can be regarded as a general extension of Bayesian theory that can robustly deal with incomplete data. Due to the capability of uncertain reasoning, it is widely applied in many fields. When there are conflicts among the evidences, however, D-S evidence theory may draw a counterintuitive conclusion [3]. Generally, there are two types of methods for dealing with conflicting evidences. One is to modify Dempster's rule of combination [4–8], while the other is to modify the evidences before using Dempster's rule. Evidence-modifying methods can be further classified into two types, i.e. weighted averaging methods [9–11] and discounting techniques [12–14]. In this paper we study the fusion performance of weighted averaging methods. Murphy's simple averaging method [9] can be viewed as a special case of weighted averaging methods where all the weights of the evidences are identical.

As studied in our previous work [15], compared with rule-modifying methods, weighted averaging methods are more attractive in that they can not only deal with conflicting evidences but converge towards dominant opinion with higher convergence speed. Among the three weighted averaging methods, Deng et al.'s method [10], which is based on Jousselme's measure of distance between two evidences [16], outperform the other two [9, 11]. Since it takes the relationship among the evidences into account, reasonable combination results can be obtained even if some conflicting evidences are collected due to e.g. enemy's disguise or bad weather.

Nevertheless, the convergence of Deng et al.'s method is still imperfect. This paper analyzes the problems and then presents an improved fusion method based on a similarity measure between two evidences and a new combination rule.

88.2 Deng et al.'s Weighted Averaging Fusion Method

In a practical multi-sensors system, the signals may be interfered with by many factors and to different degrees. Besides, some sensors may also be more stable than others. Therefore, the evidences obtained from the sensors are of different credibility degrees and should have different impacts on the fusion result. A reasonable way to handle this problem is to assign a weight to each evidence. When there is no prior knowledge, the relative importance of an evidence can be evaluated by the similarities between it and the other evidences.

Given a finite set of mutually exclusive and exhaustive propositions, i.e. a frame of discernment $\Theta=\{A_1, A_2, \ldots, A_m\}$, where A_i denotes a proposition. All possible subsets of Θ form are a superset $P(\Theta)$ containing 2^N elements. Suppose m_i and m_j be two basic probability assignment functions under the same frame of discernment. Jousselme [16] propose a distance measure between two evidences as.

$$d_{ij} = \sqrt{\frac{1}{2}(m_i - m_j)^T D(m_i - m_j)} \tag{88.1}$$

where D is a $2^N \times 2^N$ matrix with elements $D(A,B) = \frac{|A \cap B|}{|A \cup B|}$, $A, B \in P(\Theta)$.

Then the similarity between two evidences can be defined as

$$sim_{ij} = \frac{1}{2}(\cos(\pi d_{ij}) + 1)$$

The degree of support of an evidence by all the other evidences is defined by.

$$sup_i = \sum_{j=1, j \neq i}^{n} sim_{ij}$$

The normalization of support degree leads to the following credibility degree of evidence

$$cred_i = sup_i / \sum_{j=1}^{n} sup_j \tag{88.2}$$

Accordingly, the weighted average of the evidences is given as

$$MAE(m) = \sum_{i=1}^{n} (cred_i \times m_i)$$

As done in Murphy's method [9], the new BPA is incorporated into Dempster's rule of combination for $n - 1$ times in order to offer convergence toward certainty, if there are n evidences.

88.3 Analysis on Deng et al.'s Method

We illustrate the problems of Deng et al.'s weighted averaging method by several numerical examples as follows.

Example 1 Consider the following two groups of evidences under the frame of discernment$\Theta = \{A_1, A_2, A_3, A_4\}$:

Group 1: $m_1(A_1) = 1$, $m_2(A_1) = 1$, $m_3(\{A_1, A_2\}) = 1$, $m_4(\{A_1, A_2\}) = 1$
Group 2: $m_1(A_1) = 1$, $m_2(A_1) = 1$, $m_3(\{A_1, A_2, A_3\}) = 1$, $m_4(\{A_1, A_2, A_3\}) = 1$

The combination results by Deng et al.'s method are shown in Table 88.1. When combining the former three evidences, $m_1 \oplus m_2 \oplus m_3(A_1)$ of Group 2 gains a bigger value than that of Group 1, which is unreasonable. Since the cardinality of

Table 88.1 Combination results of Deng et al.'s method for Example 1

Evidences	$m_1 \oplus m_2$			$m_1 \oplus m_2 \oplus m_3$			$m_1 \oplus m_2 \oplus m_3 \oplus m_4$		
	A_1	$\{A_1, A_2\}$	$\{A_1, A_2, A_3\}$	A_1	$\{A_1, A_2\}$	$\{A_1, A_2, A_3\}$	A_1	$\{A_1, A_2\}$	$\{A_1, A_2, A_3\}$
Group 1	1			0.9937	0.0063		0.9375	0.0625	
Group 2	1			0.9976		0.0024	0.9375		0.0625

$\{A_1, A_2\}$ is less than that of $\{A_1, A_2, A_3\}$, the third evidence of Group 1 contains more certainty information about A_1 than that of Group 2. Therefore, the combined mass on A_1 of Group 1 should be bigger. It is also unreasonable the combined mass on A_1 are equal for both groups when combining all the four evidences.

Example 2 Given the frame of discernment $\Theta = \{A_1, A_2, A_3, A_4\}$ and two groups of evidences, each comprising four conflicting evidences with the following BPAs.

Group 1: $m_1(A_1) = 1$, $m_2(A_1) = 1$, $m_3(\{A_2, A_3\}) = 1$, $m_4(\{A_2, A_3\}) = 1$.
Group 2: $m_1(A_1) = 1$, $m_2(A_1) = 1$, $m_3(\{A_2, A_3, A_4\}) = 1$, $m_4(\{A_2, A_3, A_4\}) = 1$.

Table 88.2 shows the fusion results. Obviously, combining the former three evidences produces counterintuitive results in both the groups. Since A_1 is not a focal element in the third evidence, $m_1 \oplus m_2 \oplus m_3(A_1)$ should be smaller than $m_1 \oplus m_2(A_1)$. In fact, the combination leads to the following distance matrix for both the groups

$$d = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

According to formula (2), there is $Cred = [0.5, 0.5, 0]$. Thus, the third evidence does not have any impact on the combination result.

Example 3 Suppose the first two evidences are the same as in Example 2 and the third one varies as follows.

$$m_3(A_1) = 1 - j/100,\ m_3(A_2) = j/100\ (j = 1, \cdots, 100)$$

As shown in Fig. 88.1, the combined mass on A_1 varies nonmonotonically, which is confusing since $m_3(A_1)$ decreases monotonically.

Table 88.2 Combination results of Deng et al.'s method for Example 2

Evidences	$m_1 \oplus m_2$			$m_1 \oplus m_2 \oplus m_3$			$m_1 \oplus m_2 \oplus m_3 \oplus m_4$		
	A_1	$\{A_2, A_3\}$	$\{A_2, A_3, A_4\}$	A_1	$\{A_2, A_3\}$	$\{A_2, A_3, A_4\}$	A_1	$\{A_2, A_3\}$	$\{A_2, A_3, A_4\}$
Group 1	1			1			0.5	0.5	
Group 2	1			1			0.5		0.5

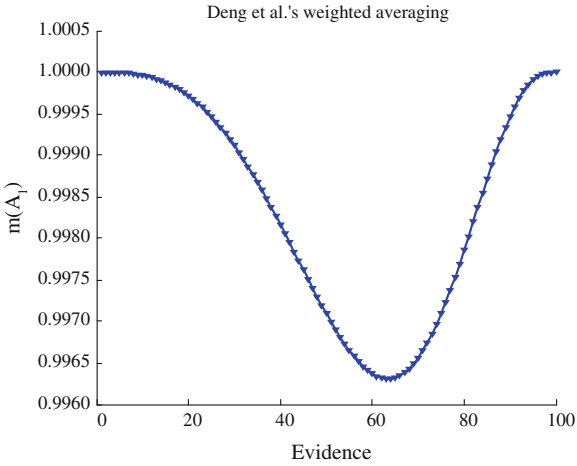


Fig. 88.1 Combined mass on A_1 by Deng et al's weighted averaging method when the third evidence varies

88.4 A New Fusion Method

Let m_i and m_j be two BPAs under the frame of discernment Θ containing N propositions. The similarity between m_i and m_j is defined as

$$sim(\mathbf{m}_i, \mathbf{m}_j) = \frac{\mathbf{m}_i' \mathbf{D} \mathbf{m}_j}{||\mathbf{m}_i||_{\mathbf{D}} ||\mathbf{m}_j||_{\mathbf{D}}}$$

where $\mathbf{D}$ is a $2^N \times 2^N$ matrix and $||\mathbf{m}||_{\mathbf{D}} = \sqrt{\mathbf{m}' \mathbf{D} \mathbf{m}}$.

The similarity measure *sim* is a cosine measure which can be categorized into the inner product family of similarity measures [17]. Wen et al. define a cosine similarity measure with $\mathbf{D} = \mathbf{I}$ [18], which does not satisfy any structural property [16]. For the proposed similarity measure, the $\mathbf{D}$ matrix would quantify the interaction between the focal elements of the BPAs. As can be seen from formula (1), Jousselme's distance also satisfied the strong structural property by constructing the matrix via Jaccard index. More choice of such indexes can be found in [19]. However, using any of these indexes still results in the problem described in Example 2. Therefore, a new index is needed.

Let s denote $|A \cap B|$, t refer to $|\Theta - (A \cup B)|$, and p to $|\Theta|$, where $A, B \in P(\Theta)$. The index is defined as

$$D(A, B) = \frac{s + t + p}{2p}$$

Then the degree of support of an evidence by other evidences is defined by

$$sup_i = \sum_{j=1, j \neq i}^{n} sim_{ij} + c$$

where c is a constant which has important influence on the monotonicity of combined mass. In this paper, c takes the value of 2.

Afterwards, the credibility degree of evidence and the weighted average of the evidences can be defined similar to Deng et al.'s method. In order to improve the converging performance, we also integrate structural information into Dempster's rule of combination as follows.

$$\left.\begin{aligned} m(A) &= \frac{1}{1-k}\sum_{A_i\cap B_j=A} m_1(A_i)m_2(B_j)\frac{|A|^2}{|A_i||B_j|} \\ k &= 1-\sum_{A\in P(\Theta),A\neq\emptyset} m(A) \end{aligned}\right\}$$

We apply the proposed fusion method to the examples discussed in Sect. 88.3 and the combination results are shown in Tables 88.3, 88.4 and Fig. 88.2, respectively.

For Example 1, the combined mass on A_1 of Group 1 gains a bigger value than that of Group 2, no matter when the former three evidences or all the four evidences are combined.

For both the groups in Example 2, the former two evidences are identical and therefore there is $m_1 \oplus m_2(A_1) = 1$. The combined mass on A_1 decreases when combining the former three evidences due to the high conflict among them. Obviously, the third evidence exerts an impact on the combination result as expected. Taking Group 2 as an illustration, when the former three evidences are considered, the credibility degree of the third evidence is 0.3. Besides, it is worthy of notice that the combination results of the former three evidences are different for the two groups. Similar to Example 1, the reason is also related to the cardinalities of focal elements. That is, though $m_3(\{A_2, A_3\})$ and $m_3(\{A_2, A_3, A_4\})$ are equal, the former contains more certainty information than the latter and therefore the combined mass on $\{A_2, A_3\}$ is bigger than that on $\{A_2, A_3, A_4\}$.

For Example 3, it can be observed from Fig. 88.2 that the combined mass on A_1 decreases monotonically when $m_3(A_1)$ decreases. The combination result is much reasonable than as shown in Fig. 88.1.

Table 88.3 Combination results of the proposed method for Example 1

Evidences	$m_1 \oplus m_2$			$m_1 \oplus m_2 \oplus m_3$			$m_1 \oplus m_2 \oplus m_3 \oplus m_4$		
	A_1	$\{A_1, A_2\}$	$\{A_1, A_2, A_3\}$	A_1	$\{A_1, A_2\}$	$\{A_1, A_2, A_3\}$	A_1	$\{A_1, A_2\}$	$\{A_1, A_2, A_3\}$
Group 1	1			0.9454	0.0546		0.8519	0.1481	
Group 2	1			0.9398		0.0602	0.7891		0.2109

Table 88.4 Combination results of the proposed method for Example 2

Evidences	$m_1 \oplus m_2$			$m_1 \oplus m_2 \oplus m_3$			$m_1 \oplus m_2 \oplus m_3 \oplus m_4$		
	A_1	$\{A_2, A_3\}$	$\{A_2, A_3, A_4\}$	A_1	$\{A_2, A_3\}$	$\{A_2, A_3, A_4\}$	A_1	$\{A_2, A_3\}$	$\{A_2, A_3, A_4\}$
Group 1	1			0.9174	0.0826		0.5	0.5	
Group 2	1			0.9270		0.0730	0.5		0.5

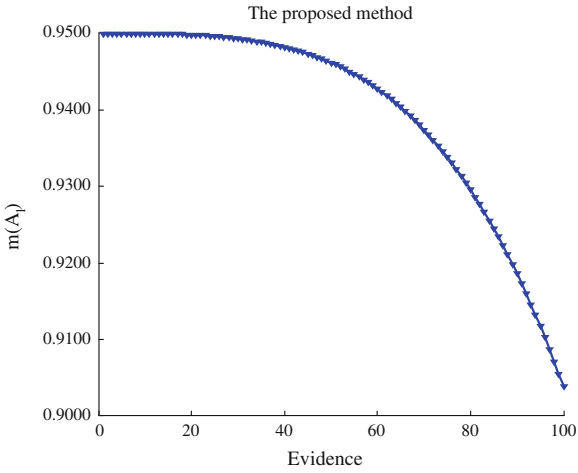


Fig. 88.2 Combined mass on A_1 by the proposed method when the third evidence varies

88.5 Conclusion

Though Deng et al.'s fusion method can gain more robust results than many other known methods, it may still lead to counterintuitive or confusing results. This paper brings forward an improved weighted averaging method involving a new similarity measure between evidences and a new combination rule. The numerical examples show the proposed method well solves the problems.

Acknowledgments This paper is supported by National Natural Science Foundation of China (61074087), Innovation Program of Shanghai Municipal Education Commission of China (12ZZ144), and Innovation Ability Construction Project for Teachers of School of Optical-Electrical and Computer Engineering of University of Shanghai Science and Technology (GDCX-Y1111).

References

1. Dempster, A.P.: Upper and lower probabilities induced by multivalued mapping. Ann. Math. Statist. **38**(3), 325–339 (1967)
2. Shafer, G.: A mathematical theory of evidence. Princeton University Press, Princeton (1976)

3. Zadeh, L.A.: A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination. AI mag. **7**(2), p. 85 (1986)
4. Yager, R.R.: On the Dempster-Shafer framework and new combination rules. Inf. Sci. **41**(2), 93–137 (1987)
5. Sun, Q., Ye, X.Q., Gu, W.K.: A new combination rule of evidence theory. Acta Electroni. Sinica **28**(8), 116–119 (2000) (Chinese)
6. Deng, Y., Shi, W.K.: A modified combination rule of evidence theory. J. Shanghai Jiaotong Univ. **37**(8), 1275–1278 (2003) (Chinese)
7. Du, F., Shi, W.K., Deng, Y.: Feature extraction of evidence and its application in modification of evidence theory. J. Shanghai Jiaotong Univ. z**1**, 164–168 (2004) (Chinese)
8. Xiang, Y., Shi, X.Z.: Modification on combination rules of evidence theory. J. Shanghai Jiaotong Univ. **33**(3), 357–360 (1999) (Chinese)
9. Murphy, C.K.: Combining belief functions when evidence conflicts. Decis. Support Syst. **29**(1), 1–9 (2000)
10. Deng, Y., Shi, W.K., Zhu, Z.F., Liu, Q.: Combining belief functions based on distance of evidence. Decis. Support Syst. **38**(3), 489–493 (2004)
11. Chen, L.Z., Shi, W.K., Deng, Y., Zhu, Z.F.: A new fusion approach based on distance of evidences. J. Zhejiang Univ. Sci. **6A**(5), 476–482 (2005)
12. Martin, A., Jousselme, A.L., Osswald, C.:Conflict measure for the discounting operation on belief functions. In: Proceedings of the 11th International Conference on Information Fusion. 1–8 (2008)
13. Browne, F., Bell, D., Liu, W., Jin, Y., Higgins, C., Rooney, N., Wang, H., Muller, J.: Application of evidence theory and discounting techniques to aerospace design. Adv. Comput. Intel. 543–553 (2012)
14. Zhou, Z., Xu, X.B., Wen, C.L., Lv, F.: An optimal method for combining conflicting evidences. Acta Automat. Sinica. **38**(6), 976–985 (2012) (Chinese)
15. Li, Y., Wang, Y.G., Xu, X.M.: A numerical cases study of evidence fusion methods. Chinese Automation Congress, Beijing (2011)
16. Jousselme, A., Grenier, D., Bossé, É.: A new distance between two bodies of evidence. Inf. Fusion **2**(1), 91–101 (2001)
17. Jousselme, A., Maupin, P.: Distances in evidence theory: comprehensive survey and generalizations. Int. J. Approx. Reason. **53**(2), 118–145 (2012)
18. Wen, C., Wang, Y., Xu, X.: Fuzzy information fusion algorithm of fault diagnosis based on similarity measure of evidence. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Beijing, pp. 506–515 (2008)
19. Diaz, J., Rifqi, M., Bouchon-Meunier, B.: A similarity measure between basic belief assignments. In: Proceedings of the 9th International Conference on Information Fusion. pp. 1–6 (2006)

Chapter 89
Optimization for Family Energy Consumption in Real-Time Pricing Environment

Weipo Wu, Genke Yang, Changchun Pan and Changjiang Ju

Abstract In order to help consumers adapt to electricity consumption in real-time electricity pricing environment, an energy consumption scheme is proposed in this paper. This scheme focuses on the prediction, modeling and optimization for family energy consumption. A method based on support-vector machine (SVM) is used to predict the real-time price (RTP) and the optimization model divides every hour into equal time slots and thus provides more opportunities to schedule household appliances in proper working time. Then the simulation results show that the proposed optimal control model reduces the daily electricity expenditures.

Keywords RTP · Electricity consumption scheduling · Price prediction · Time slot

89.1 Introduction

Nowadays, real-time pricing model has been proposed in order to reflect the real supply—demand relationship in the electricity market more accurately. This pricing strategy not only reflects the actual wholesale prices but also encourages consumers to shift high-load household appliances to off-peak hours so that it can reduce their electricity payments and peak-to-average ratio (PAR) in load demand simultaneously [1, 2].

However, recent studies showed that there are two major limitations to implement the RTP strategy. On one hand, most consumers do not want to choose the RTP electricity supply system due to lack of the knowledge about it. On the other hand, the absence of automatic family energy management system is the

W. Wu (✉) · G. Yang · C. Pan · C. Ju
Department of Automation, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai Jiao Tong University, Shanghai, China
e-mail: weipowu@126.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_89,

other element which limits the consumers to respond to the time-varying electricity prices more properly [3].

This paper focuses on the family energy consumption scheduling model which aims to solve the above problem. The second section explains details of an electricity prediction method based on SVM, and then gives the forecasting value based on the RTP data of Illinois Power Company (IPC) from January 2009 to December 2011 [4]. In the third section, it clearly describes how to schedule the consumption under different conditions. Then, it gives a model that could ensure the consumer spend the minimum payment but still finish the work in a comfort way. In the fourth section, this paper illustrates the simulation results. Finally, there is the conclusion.

89.2 Price PredictionModel

Clearly, electricity price mainly depends on the wholesale market prices, different time in the day and different weathers which determine the supply and demand of the electricity [5]. Since it has several input variables, the prediction model will be the non-linear mapping function. In order to be used in the Real-time pricing environment, this section will use the (SVM) Price Prediction Strategy [6].

Recent studies showed that hourly price of electricity is highly related with the historical price [7]. This part will analyze the RTP data by IPC from January 2009 to December 2011 [4].

The result has been showed in the Fig. 89.1, which plotted the correlation among the current hourly-prices with the same time in the past few days. Clearly indicated in the Fig. 89.1, the correlation coefficient is declining cyclically as it goes further back, and the prices have the highest correlation between two continuous days, e.g., today and yesterday. Additionally, the figure also represents a noticeable correlation between the prices today and those in the same day last week.

With consideration of these characteristics of the price series, the following vector of input features has been considered to forecast the price p_h at hour h

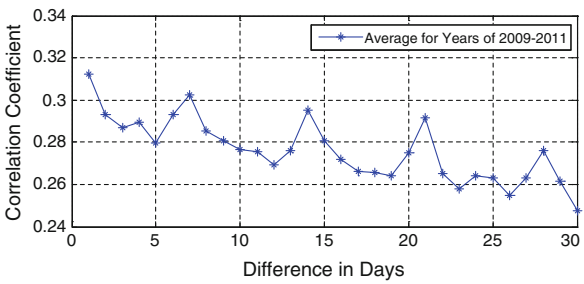


Fig. 89.1 The correlation coefficient between the same hour in different days

$$\begin{aligned} X_i = [&p_{h-1}, p_{h-2}, p_{h-22}, p_{h-23}, p_{h-24}, p_{h-25}, \\ &p_{h-26}, p_{h-167}, p_{h-168}, p_{h-169}, p_{h-192}, p_{h-193}] \end{aligned} \tag{89.1}$$

In (89.1), the first two terms that consist of price information of the two previous hours are used to model the trend of the price signal. The rest of the terms contain information about price in the previous period to model the multiple seasonality of the electricity price signal.

This paper uses LIBSVM software to perform experiments and choose Mean Squared Error (MSE) which is defined as follow to measure its prediction accuracy [8].

$$\text{MSE} = \frac{1}{l}\sum_{i=1}^{l}(f(X_i) - y_i)^2 \tag{89.2}$$

where l is the number of prediction prices, y_i is the real price data and $f(X_i)$ is the forecasting price data. So it is easy to know X_i is the input vector of prediction model, f is the prediction function.

This paper chooses the data from 1st May to 31st July in 2011 as training data, and chooses the data in August as testing data as well as chooses cross-validation and grid search method to determine the penalty parameter c and kernel parameter g in LIBSVM [8]. The result about parameters is shown in Fig. 89.2(a, b) where they find the best penalty parameter $c = 0.5$ and kernel parameter $g = 4$. At the same time, the forecasting electricity price is shown in Fig. 89.3 and the prediction result approximates to the real data, where the MSE = 0.0275 is far less than the result of Back-Propagation Neural Network model, which is 1.1230.

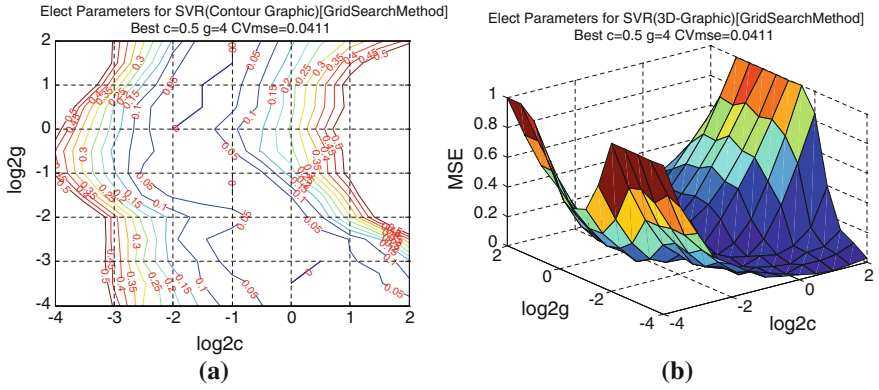


Fig. 89.2 Use cross-validation and grid search method to find the best parameters in SVM

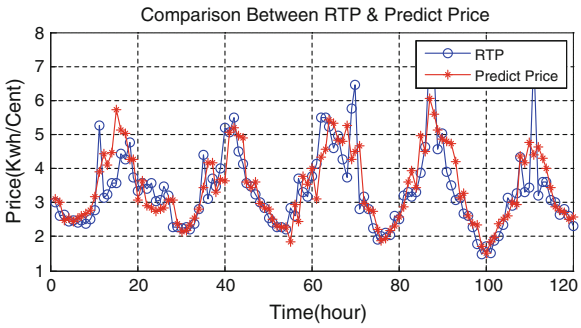


Fig. 89.3 Result when using SVM to predict the price from 5th to 9th August in 2011 of IPC

89.3 Model Formulations

This section will introduce the family energy consumption optimal control model, aiming to help each household to maximize the efficiency of electricity they are consumed, and meanwhile minimize the electricity payment they are supposed to spend.

89.3.1 Electricity Consumption Scheduling

This part will describe the energy consumption scheduling model including continuous electricity consumption and discrete electricity consumption. Additionally, the situation with uninterruptible electricity consumption will be discussed.

89.3.1.1 Continuous Electricity Consumption

Consider that each residential unit wants to optimize the electricity consumption in the next $H(H \geq 1)$ hours, where H represents the scheduling horizon and we define $\mathbf{H} = [1, \ldots, H]$. Let $\mathcal{A}$ denotes the set of appliances, which could include washing machine, refrigerator, air condition, etc. Because the working time of most household appliances does not occupy the whole hour, therefore, the time axis of each hour could be divided into equal time slots Δ. It has to ensure the number of time slots in each hour $N = 1/\Delta$ is an integer. Thus, the number of total time slots in scheduling horizon is $L = N * H$, where $\mathbf{L} = [1, \ldots, L]$.As a result, for each appliance $a \in \mathcal{A}$, we define an electricity consumption scheduling vector [7]

$$\mathbf{e}_a = \left[e_a^1, \ldots, e_a^n, \ldots, e_a^L\right] \tag{89.3}$$

where e_a^n means how much electricity the appliance a consumed in the n^{th} time slot. So it is easy to know $e_a^n \geq 0$ when $n \in \mathbf{L}$ and $a \in \mathcal{A}$.

Now, assume consumers set their own scheduling horizon for each household appliance. For example, consumers want the automatic clean machine start to clean the house at their working time. Hence, they set the machine's scheduling horizon from 8:00 A.M. to 17:00 P.M. Then, electricity consumption *Ea* is expressed as follow,

$$\sum_{n=\alpha_a}^{\beta_a} e_a^n = E_a \tag{89.4}$$

where $\alpha_a \geq 1$ is the beginning of time interval, and $\beta_a \geq \alpha_a$ is the ending of time interval of the scheduling horizon for appliance a.

However, as we know, the household appliance is working in a limited power. So the constraint could be expressed as

$$\gamma_a^{min}/N \leq e_a^n \leq \gamma_a^{max}/N, \forall n \in [\alpha_a,\ \beta_a] \tag{89.5}$$

which means the scheduled energy consumption of appliance a in hour h is bounded between γ_a^{min} and γ_a^{max}.

Due to the assigned electricity load for each family at each hour is limited, so the limited equation is

$$\sum_{a \in \mathcal{A}} e_a^n \leq \boldsymbol{E}_{max}/N, \quad \forall n \in \boldsymbol{L}, h \in \boldsymbol{H} \tag{89.6}$$

where $\mathrm{E}_{max} \geq 0$ is the upper limited power in hour h for a family.

89.3.1.2 Discrete Electricity Consumption

So far it considers the household appliances consume electricity in a continuous way. However, some households work with discrete electricity consumption level, which $\mathcal{A}_\mathrm{D}$ denotes. In other words, the scheduled electricity consumption for some appliance may only take the discrete values $\gamma_\mathrm{a}^{\min}/N$ and $\gamma_\mathrm{a}^{\max}/N$ when the appliance is "off" and "on".

In order to describe this kind of households, let $\mathrm{y_a^n}$ denote an auxiliary binary variable, when $\mathrm{y_a^n} = 1$ the appliance a is "on" and when $\mathrm{y_a^n} = 0$ the appliance a is "off". By definition, the former requires an energy consumption level of $\mathrm{e_a^n} = \gamma_\mathrm{a}^{\min}/N$ while the latter is $\mathrm{e_a^n} = \gamma_\mathrm{a}^{\max}/N$. Therefore, for each appliance $\mathrm{a} \in \mathcal{A}_\mathrm{D}$, the relationship between the energy consumption scheduling vector $\boldsymbol{e_a}$ and the auxiliary $\mathrm{y}_{\ \mathrm{a}} = [\mathrm{y_a^{\alpha_a}}, \ldots, \mathrm{y_a^n}, \ldots, \mathrm{y_a^{\beta_a}}]$ can be expressed as follows:

$$\mathrm{e_a^n} = \mathrm{y_a^n} * \gamma_\mathrm{a}^{\max}/\mathrm{N} \ + \left(1 - \mathrm{y_a^n}\right) * \gamma_\mathrm{a}^{\min}/\mathrm{N} \tag{89.7}$$

89.3.1.3 Uninterruptible Electricity Consumption

Under another circumstance, the household may have some appliances that have to work in uninterruptible electricity consumption condition. We call them uninterruptible loads which mean once the appliances start operation, their operation need to continue until they finish. This paper defines them as $\mathcal{A}_U$.

Consider an uninterruptible load $a \in \mathcal{A}_U$ working in discrete energy consumption level, let θ_a denote the duration of time, in number of time slots, the appliance a needs to operate at power level γ_a^{max}/N. Let's impose z_a^n as an auxiliary binary variable as well. When the uninterruptible load starts to operate, $z_a^n = 1$, otherwise $z_a^n = 0$. So equations are expressed as follow,

$$\begin{aligned} &\sum_{n=\alpha_a}^{\beta_a-\theta_a+1} z_a^n = 1, \qquad \forall a \in \mathcal{A}_U \\ &z_a^n = 0, \quad \forall n \in \mathbf{L}\backslash[\alpha_a, \ldots, \beta_a - \theta_a + 1], \end{aligned} \tag{89.8}$$

that is, the operation of appliance a is to begin working between time slot α_a and $\beta_a - \theta_a + 1$.We can relate the start time vector $\boldsymbol{z}_a = [z_a^{\alpha_a}, \ldots, z_a^n, \ldots, z_a^{\beta_a - \theta_a + 1}]$ with auxiliary vector $\mathrm{y_a}$ as

$$\mathrm{y_a^n} \geq \mathrm{z_a^n}, \ldots, \mathrm{y_a^{\beta_a-\theta_a+1}} \geq \mathrm{z_a^n}, \forall \mathrm{n} \in [\alpha_\mathrm{a}, \ldots, \beta_\mathrm{a} - \theta_\mathrm{a} + 1] \tag{89.9}$$

from (89.8), if $\mathrm{z_a^n} = 1$, $\mathrm{y_a^n} = \mathrm{y_a^{n+1}} = \ldots = \mathrm{y_a^{n+\theta_a-1}} = 1$. On the other hand, from (89.7) and (89.9), it is easy to get $\mathrm{e_a^n} = \mathrm{e_a^{n+1}} = \ldots = \mathrm{e_a^{n+\theta_a-1}} = \gamma_\mathrm{a}^{\max}/N$.

89.3.2 Problem Formulation

In this section, assume that each household is equipped with a smart meter with two-way communication and the real-time prices are provided by the utility company via local area network. The consumers choose their requirements by selecting parameters E_a, α_a, β_a, γ_a^{min} and γ_a^{max} as well as adjusted the appliance's energy consumption ways, such as continuous way, discrete way or interruptible way. Consequently, the electricity scheduler determines the optimal choice of electricity consumption scheduling vector $\boldsymbol{e}$. Then the resulting electricity consumption schedule is applied to all household appliances.

To minimize the energy payment, the optimal control model is

$$\min \sum_{h=1}^{H} p_h(E_h) * E_h \tag{89.10}$$

where H is the schedule horizon and $h \in \boldsymbol{H}$ as well as $p_h(E_h)$ is the electricity price of hour h. Additionally, Formulation (89.4–89.9) are the constraints of this model and how much energy is consumed in hour$h$ is calculated as follow:

$$E_h = \sum_{n=(h-1)*N+1}^{h*N} \sum_{a \in \mathcal{A}} e_a^n \tag{89.11}$$

89.4 Simulations

This section will present the simulation results and evaluate the performance of the proposed model with price prediction. Consider a single household with different appliances and assume that it has adopted the RTP program. The test period is one month from 1st August to 31st August in 2011, which includes 31 days in total. For the purpose of this paper, assume that the number of appliances used in this household each day varies from 10 to 15. They include certain appliances with fixed consumption schedules such as lighting, heating, refrigerator, etc., and appliances with flexible energy consumption schedules such as house clean machine, dishwasher, clothes washer, and PHEV, etc [9]. Here assume that the scheduling horizon $H = 24$. As the user has subscribed for the RTP program adopted by IPC, this would require price prediction as discussed in Sect. 89.3.

89.4.1 Gains with Control Model

This paper simulates the energy consumption in two ways. In the first way, consider the household consume energy with the proposed optimal model while the other way is to use power as usual. As indicated in Fig. 89.4, the payment with energy optimal control and the parameter $N = 1$ is less than the expenditures without control. In the August 2011, the user only need to pay 29.26 dollars for the electricity consumption with scheduling optimal control, while 32.41 dollars will be cost if there was not an energy optimal control scheme, which is nearly 10 % cheaper. However, there still have some exceptions that in 3rd and 26th of August, the electricity charges with control are higher than that without control. It is easy to understand as the error of price forecasting. Nevertheless, the differences between those payments are not significant; therefore, this control scheme could be seen as useful.

89.4.2 The Influence of the Parameter N

Here discuss the influence of parameter N. it uses the optimal control model with different value of N. Figure 89.5 shows that the electricity payment with 4 time slot per hour ($N = 4$) will pay 4.585 % less than that with only 1 time slot per hour

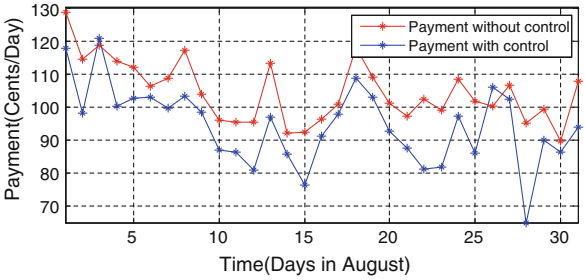


Fig. 89.4 Payment comparison between control and uncontrol

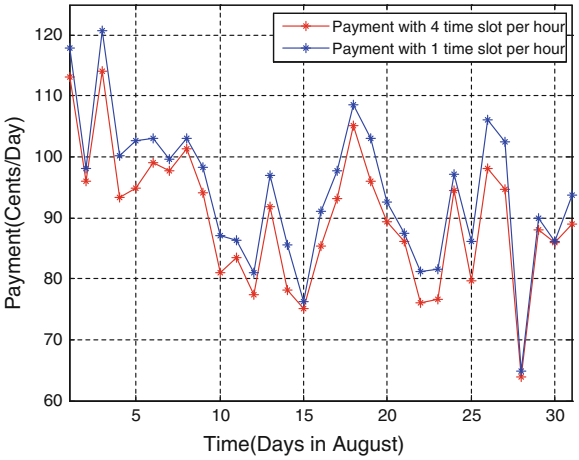


Fig. 89.5 Payment comparison between different numbers of time slots

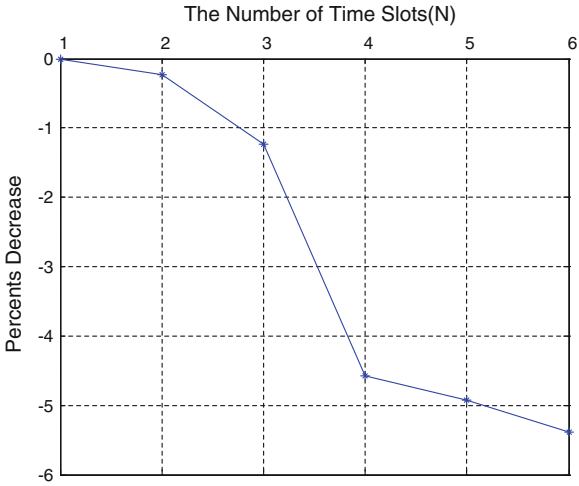


Fig. 89.6 Payment decrease with the increase of N

(parameter $N = 1$). This is mainly due to the fact that with more time slots, the control model will have more schedule range to ensure the appliances to work in more proper time, such as off-peak time.

According to Fig. 89.6, the payment is decreasing when parameter N is increasing. The reason is the same as above, i.e., with more time slot in an hour, the optimal control model will have more opportunities to schedule the household appliances in order to decrease the expenses. However, as presented in the Fig. 89.6, the decreasing rate of payment is lower when time slot s higher than 5. Thus, there must be a proper number of N could help the consumer to save maximum money. However, because different consumer will set their own parameters, therefore, the appropriate of N is hard to measure based on single benchmark. The graph here only describes the overall pattern.

89.5 Conclusion

This paper proposes a family energy consumption optimal control model which is applied in the environment installed smart meter and aims to minimize the electricity payment based on the needs declared by users. It argues that any load control in real-time electricity pricing environment essentially requires some price prediction capabilities to enable planning for the household energy consumption in advance. This paper uses SVM method with proper input values to forecast the hourly-based prices adopted by IPC from January 2009 to December 2011 and obtains the best parameters for the prediction model. Then it describes the electricity consumption scheduling model where it divides each hour into equal time slots. In the end, it makes a simulation whose results show that the optimal control model reduces the daily electricity expenditures, which will encourage the users to participate in the proposed control model.

Acknowledgments This research is supported by National Nature Science Foundation of China Grant 61074150 and Grant 61203178.

References

1. Kim, T., Poor, H.V.: Scheduling power consumption with price uncertainty. IEEE Trans. Smart Grid **2**(3), 519–527 (2011)
2. Ipakchi, A., Albuyeh, F.: Grid of the future. IEEE Power Energy Msg **8**(4), 52–62 (2009)
3. Ann-Piette, M., Ghatikar, G., Kiliccote, S., Watson, D., Koch, E., Hennage, D.: Design and operation of an open, interoperable automated demand response infrastructure for commercial buildings. J. Comput. Inf. Sci. Eng. **9**, 1–9 (2009)
4. Real-time pricing for residential customers, Ameren Illinois Power Co. http://www.ameren.com/sites/aiu/ElectricChoice/Pages/Home.aspx (2012)
5. Zhang, X., Wang, X.: Review of the short-term electricity price forecasting. Autom. Electr. Power Sys. **30**(3), 92–101 (2006)

6. Vapnik, V.: The nature of statistical learning theory. Springer, New York (1995)
7. Mohsenian-Rad, A.H., Leon-Garcia, A.: Optimal residential load control with price prediction in real-time electricity pricing environments. IEEE Trans. Smart Grid. **1**(2), 120–133 (2010)
8. Chang, C., Lin, C.: LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/cjlin/libsvm (2012)
9. Natural Resources Canada.: Energy consumption of major household appliances shipped in Canada. http://oee.nrcan.gc.ca/Publications/statistics/cama07/pdf/cama07.pdf (2007)

Chapter 90
The Implementation with the Network Data Security on the Secure Desktop

Yi Liao and Xiao-Ting Li

Abstract Nowadays, the electronic power industry is improving the level of information, and the data security of the internal network is arousing more and more attention. The domestic units and enterprises implemented relatively comprehensively. Most of them deployed firewall, intrusion detection and vulnerability scanning systems which allow them to reduce the risk of network boundaries greatly. However, as to the network terminal management within the network of the enterprise, the current risk prevention measures are far from enough because the internal network security issues come out easily. Combined with the extensive practical application of the electric power enterprise, this paper puts forward the secure desktop idea so as to solve the terminal security of the internal network computers. This measure can achieve the aim of a further control of the terminal security of the internal network computers. Thereby, it realizes a better controlled data security within the internal network and provides a good security network solution of the protection with the electric power business.

Keywords Secure desktop · Internal network security · Computer terminals

90.1 Introduction

With the technology of computer network, digital communication and virtualization technology are developing fast, the level of the application of information network technology is extending, the applied business system of the electric power enterprise is keeping increasing, and the applied system of enterprise has been networked and infomationization. As the large enterprise may be related to the

Y. Liao (✉) · X.-T. Li
Information Technology Center, China Nuclear Power Technology Research Institute, China
Guangdong Nuclear Power Holding Co. Ltd, Shenzhen, China
e-mail: liaoyi2010@cgnpc.com.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_90,

national secret-related system, the data transmitted may involve the information that related to the safe of country and the life of people so that it should be kept secure and strictly prevent from letting out. As we can see that the security of the internal network is influencing on more and more extend field, the security of the computer terminal is attaching much more attention [1].

The secure desktop idea provides a relatively good solution method which is being used more and more aboard in the electric power enterprise. This paper will introduce the technology of the secure desktop and explain how to apply the secure desktop technology on the computer terminal of enterprise so to achieve the aim of keeping the security of the internal network of enterprise.

The structure of this paper is as follows: Sect. 90.2 will explain the main threat faced by the current internal network security ; Sect. 90.3 will introduce the basic technology and the concept of secure desktop and the principle of how to execute the control of internal network security while using the secure desktop technology; Sect. 90.4 will introduce the advantage brought by the technology's controlling of internal network; Sect. 90.5 make a conclusion.

90.2 Analysis

The security assurance is always the primary content that should be firstly schemed and established by the builders during the establishment of informationization. Especially, when the government, army, large-scale enterprise develop the informationization construction, they build special internal network that physically separated from the internet, adopt many security assurance system, build corresponding institution of security assurance, even caring nothing about how much it will cost. Meanwhile, the security of enterprise internal network is getting more and more attention. No one hopes that the confidential sensitive information can be leaked out or stolen by others, no matter the nation, government, or individual. However, as the informationization of enterprise carries on, various kinds of accident emerge in endlessly.

As to the internal network, the ordinary security accident can be divided into two aspects. One occurs on the borderline of network, that is, the intrusion came from the outside of the network like the vicious attack, long-distance intrusion, worm virus and so on. The other occurs within the network, that is, the voluntary or passive divulgence by the internal network users. Presently, the domestic units and enterprises are implementing much more countermeasures in the aspect of borderline management. Most of the units and enterprises have deployed firewalls, intrusion detection and vulnerability scanning systems which highly reduced the risk of network borderline in the enterprise. However, these measures are far from enough on keeping security. It is suggested by investigation that the proportion of security accident happened at the borderline and internal network reach 42.86 %, among which more than 50 % information risk come from the inside of enterprise. We can say that the threat brought by the abundant computer terminal is much

more terrible than that come from the outside as the scale of terminal computer is becoming bigger and bigger [2].

By analyzing the current security accidents, we can find that the main reason is the erroneous operation by the terminal users, especially the cross usage of removable mass storage device among the internet, internal network, even secret-related network, which resulted that the secret-related information should be restricted strictly in the internal secret-related system may be leaked by internet. A mass of facts suggest that, for the general shortage of knowledge and skill among the ordinary terminal users, the document is processed and stored directly on the personal computer terminal. Some temporary file and information of buffer will be produced, which become the source of leakage. What's worse is that it can hardly be controlled for the broad distribution of the terminal. For those reasons, there is no time to delay to strengthen the protection of terminal.

90.3 Principles and Models

90.3.1 The Concept of Secure Desktop

The secure desktop is a kind of technology based on internal network security management of enterprise computer desktop. Its function can be divided as application policy management of desktop, long-distance desktop maintenance, network access policy, removable mass storage device management, uniform distribution of system patch, capital management and so on.

The secure desktop technology roots in sandbox model, which primarily be invented by GreenBorder Company. And it was purchased by Google in May 2007. After that, this patent was applied in the development of Chrome browser. Currently, the application of the secure desktop technology is mainly used to keep the security of operation system, and prevent from the infringement of virus program [3].

90.3.2 The Principle of the Secure Desktop Technology

Using the sandbox technology and the virtualization technology, the secure desktop technology establishes a virtual "container" and makes the users stay in the "container" so that the users' operations of document and registry database are virtualized. When users exit from the "container", the modification made in it will all be reduced. The design mode is equal to running the program in the virtual container, protecting the bottom data by loading the drive of its own, which belongs to the drive level protection.

Besides, the secure desktop technology combines a serial of regulations (network regulations, authorization regulations and so on) to control the authority of users' operation in the "container".

90.3.3 How the Secure Desktop Technology Realize the Security of Internal Network

Establish another desktop under the system, namely the secure desktop. The operations on this desktop are virtualized, that is to say, the secure desktop is equal to sandbox container. During the usage, only one desktop can be see, the default desktop or the secure desktop.

Basing on the deep analysis of users' requirement, the secure desktop will be started-up as long as the users log on the VPN and access important resource. This moment, a closed and virtual working environment, namely the secure desktop, will be created automatically by using the virtual technology. All the operations are virtualized and the process within the secure desktop and out the secure desktop are separated. The operations, temporary usage, data received are all redirected (virtualized) and high strength encrypted, because all data are concentrated stored in the server and nothing will be exist on the client. It efficiently avoided the risk brought by the misusage of removable mass storage device, the network attack, and Trojan program.

The secure desktop aim at intercepting and redirecting the aspect (operations may lead to the leakage of data) as follows [4]:

- Forbid using peripheral copy output: include USB, print, COM, CD-RW and so on, so as to prevent from leaking out important data.
- Virtual file operation: the modification of files and system executed by process under the secure desktop will all be redirected and encrypted. The files redirected will all be deleted after exit from VPN and the secure desktop. That's to say, no change or mark will be kept on the default desktop while operating files under the secure desktop.
- Protect the operations of registry database: redirect the operation of registry database belonged to HKEY_CURRENT_USER branch and the key value of registry database belonged to other branches can be read but not wrote.
- Constrain network communication: under the secure desktop, the communication connect outward will be strictly controlled, only the access of VPN network can be permitted, so as to prevent from leaking out the material downloaded by VPN. This constraint contains the follows:
  - Forbid communicating with the local computer: forbid the communication between the secure desktop and the physical desktop, so to avoid saving the important data on the local server.
  - Forbid communicating with local network: in the secure desktop, communicating with other computers within the internal network is forbidden, so as to prevent from transmitting and restoring the important data through LAN.
  - Forbid communicating with internet: in the secure desktop, the usage of network application and leaking out data through network communication are forbidden, so as to prevent from leaking out the important data by internet transmission.

- internal process communication (IPC) filtration: IPC filtration is mainly aimed at cutting board and message. The process within the secure desktop can't send message to the process outside the secure desktop.
- Trace cleanup: after exiting from the secure desktop, all the trace will be cleaned up compulsorily and all the operations will be reduced. Even if the secure desktop get breakdown for power off, the secure desktop will automatically detect and clean up the trace which left behind when started up next time.

90.3.4 The Application Model of the Secure Desktop in the Internal Network Security Management

For example, when a staff needs to log on the business system (like the financial system) to operate business data, he/she can only download the data through sandbox and edit it in sandbox B.

If he/she needs to search related material by Google, he/she can switch to secure desktop A of sandbox A and access the internet in sandbox A.

When he/she want to transmit the data to personal mailbox of internet and treat them at home, he/she can't do it, because accessing the internal network is not permitted in sandbox A, and the data can't be copy to sandbox A, so he/she can't see the data if in the sandbox A.

He/she can't access the network under the default desktop, but can do the daily document maintenance. If he/she needs to access the internet, he/she should log on the virtual desktop A. If he/she needs to handle official business, he/she should log on the secure desktop B. These secure desktop are separated, and data can't be transmitted mutually. We can see from it that the logic isolation among the official network, internet, and the critical business network by using the secure desktop technology [5] (Fig. 90.1).

Fig. 90.1 The application model of the secure desktop in the internal network security management

90.4 Advantage

Generally speaking, the secure desktop management system improved the level of security management by using security management information technology, took full advantage of the extant advanced network management tool, strengthened the control of network computer terminal, realized the real time security monitor system, and can make the security-integration with other network security equipment at the same time, which made the internal network become the high speed, secure official network system [6, 7].

The advantage in internal network security management brought by the secure desktop contains the aspects as follows:

- Security: the critical data can only be accessed under the secure desktop, which prevents from leaking out through internet, and keeps the security of terminal.
- Low cost: only virtual security gateway device and authorization and authentication of terminal access should be purchased, and no need to buy another set of network device and official computer.
- Quick implementation: for the original network, there is no need to reconfigure the hardware on a large scale and change the network structure.
- Easy to maintain and upgrade: only virtual security gateway needs to be maintained, only authorization and authentication tool is needed to be accessed when upgrading, and only extra virtual security gateway device is needed to become a cluster when the existent device can't satisfy the support for the fast development of network [8] (Fig. 90.2).

In view of evaluations on the method, which realizes intranet data security by the secure desktop, and from the respective of practical effect on establishment, maintenance, management, secure desktop to realize data security satisfies the actual requirement on enterprise informationization establishment, and it is suitable for application [9, 10].

90.5 Conclusion

The security protection of personal terminal has aroused general attention among the constructor and IT service providing company. For example, some of the enterprises deployed strong audit system, the double-use monitor system, and the security detection system, etc. In the special network of industry, which somewhat relieved the hidden danger of terminal security. However, these measures were almost realized by installing clients (the principle of it is similar to Trojan program) on the personal terminal forcibly, which brought the terminal users a little worry (it might bring inside secret leakage or exposure of individual privacy) of the administrator department. Thus, it was rejected at different degrees by terminal users when implemented and had limited effect.

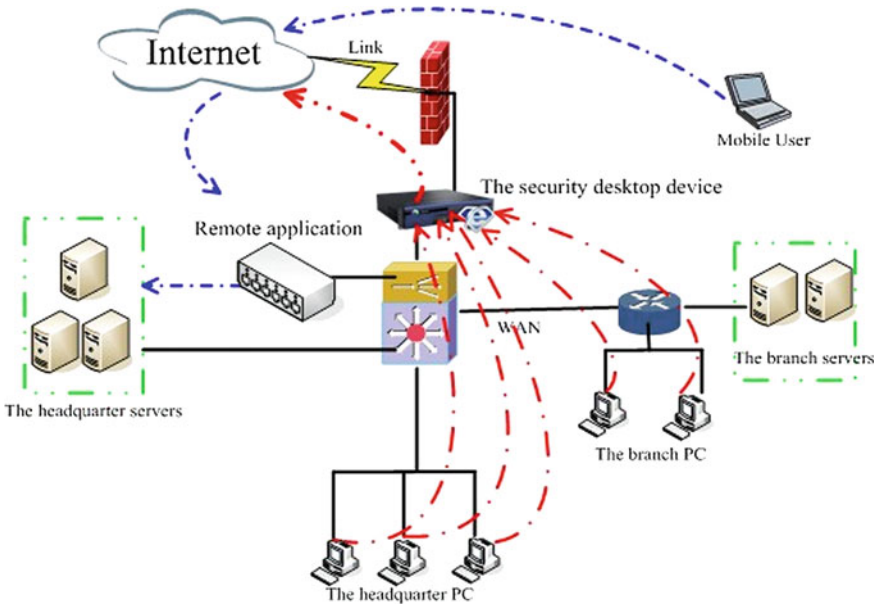


Fig. 90.2 The structure of the secure desktop in the network

As the secure desktop is specially used to protect the access of important resource, the original use habit of terminal users will not be influenced. No additional device or software will be installed in users' terminals. The users can access other resources in the special network as usual, meanwhile, they can access the important inside resources (like the inside official system) securely and conveniently without worrying about leaking out secret by his/her miss-operation and without worrying that the individual information (like personal directory and photo) stored in computer will be collected. Therefore, using the secure desktop technology to realize the security management of enterprise internal network will become necessary in the field of enterprise information security management.

References

1. Zhang, S., Liu, W., Yang, S.M.: Discuss about the green protection strategy of enterprise internal network. Electr. Power Inf. Technol. 3–8 (2008)
2. Pan, Z.Y., Pan, J.Y.: Analysis on internal network safety threat and prevention. Ordnance Ind. Autom. (2008)
3. Pan, Y.X.: The actuality and development trend of desktop security management technology. Inf. secur. Technol. (2010)
4. Hu, W., Zhang, C.H., Liao, W.: Security management function design of computer security defence system. Comput. Engine. Des. 5–30 (2010)

5. Liu, H.Y.: Research and design of security support model in the intranet. Mod. Electron. Tech. 3–20 (2007)
6. Cabuk, S., Dalton, C.I., Eriksson, K., Kuhlmann, D., Ramasamy, H.V., Ramunno, G., Sadeghi, A-R., Schunter, M., Stüble, C.: Towards automated security policy enforcement in multi-tenant virtual data centers. J. Comput. Secur. **18**(1), 80–121 (2010)
7. Berger, S., Caceres, R., Pendarakis, D.E., Sailer, R., Valdez, E., Perez, R., Schildhauer, W., Srinivasan, D.: Managing security in the trusted virtual datacenter. Oper. Syst. Rev. **42**(1), 40–47 (2008)
8. Catuogno, L., Dmitrienko, A., Eriksson, K., Kuhlmann, D., Sadeghi, A.-R., Schulz, S., Schunter, M., Winandy, M., Zhan, J.: Trusted virtual domains—design, implementation and lessons learned. In: International Conference on Trusted Systems. Spring (2009)
9. Adeyinka, O.: Internet attack methods and internet security technology. In: Second Asia international conference on modeling and simulation, 2008, AICMS 08, pp. 13–82 (2008)
10. Kartalopoulos, S.V.: Differentiating data security and network security. IEEE international conference on communications, 2008, ICC'08, pp. 1469–1473 (2008)

Chapter 91
The Integration of Temporal Database and Vague Set

Qifang Li and Chuanjuan Yin

Abstract In order to make the fuzzy information more accurate, Vague Set is used to represent the fuzzy attribute of the database in this paper. Each data and tuple in the database based on Vague Set can be reflected from the three sides of true, false and the unknown extent. For illustration, a vague relationship instance teacher1 is utilized to show the basic theory of vague relation. Results show that if the database model based on Vague Set and the fuzzy relational data model are compared, the former has a more fuzzy ability to express. The further expansion of vague database model which is temporal database based on Vague Set is proposed. In order to better explain vague temporal data, a time dimension 'duty period' is added into table teacher1, which constitutes vague temporal relation teacher2. The study lays the foundation for the further research into data model, data theory, and database management system of vague temporal database.

Keywords Fuzzy · Vague set · The temporal database · Model

91.1 Introduction

Time is nature's ubiquitous objective attributes, all the information have the corresponding temporal attributes. With in-depth development of database and information technology, information systems are facing many new applications

Q. Li (✉) · C. Yin
Institute of Information, Yunnan University of Finance and Economics, Kunming, Yunnan, China
e-mail: lym1213@163.com

C. Yin
e-mail: ycj_1980@yahoo.com.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_91,

and new requirements. More and more pressing demand for temporal information is processing.

Traditional database based on two-valued logic and accurate data can not represent a fuzzy object. With the establishment of the theory of fuzzy mathematics system, people can use relationship between the number to describe the fuzzy object and fuzzy computing. Fuzzy database is an important research direction of the database field. It provides an important way for a fuzzy information management, and in the management the information is inaccurate and incomplete.

Therefore, fuzzy database has become one of the forefront topics of the database field. Fuzzy Set is the theoretical basis of fuzzy database. Fuzzy Set to describe uncertain information is flawed. Vague Set is to promote the concept of Fuzzy Set; it has the ability of powerful processing fuzzy information. Vague Set theory was proposed in 1993 by Taiwanese scholars WLGau and DJBuehrer, after nearly 10 years of research and development some achievements have received, but they are mainly used in the fuzzy decision-making, rarely used in the research field of fuzzy relational database. The ability to express the fuzzy information of Vague Set is stronger than the Fuzzy Set. Temporal database research is based on Vague Set theory. Then a new research direction for fuzzy temporal database is opened.

91.2 Vague Relation

91.2.1 The Theory of Vague Relation

In Vague Set, the relationship of between the elements of the domain and a collection of the domain is 'belonging to a certain extent within the scope of', which uses interval to express its membership degree. The interval gives the extent of supporting evidence and opposing evidence at the same time, and also represent and process fuzzy information that can not be represented and processed by the Fuzzy Set.

Definition 1 Let U be the space of an object, x represents random one element, 'A ' is a Vague Set which is in U, that 'A' is expressed a true membership function t_A and a false membership function f_A. $t_A(x)$ expresses the lower bound of the degree of membership that support x ∈ A, and $f_A(x)$ expresses the lower bound of the degree of membership that opposes x ∈ A. Each point in X and a real number in interval [0,1] are linked by two functions $(t_A(x), f_A(x))$, that is, t_A: X → [0,1] f_A: X → [0,1]. A (x) means the degree of membership which X belongs to A, and A (x) is expressed as $[t_A(x), 1 - f_A(x)]$, and $t_A(x) + f_A(x) \leq 1$ [1].

Definition 2 Let A be a Vague Set, when X is a discrete space, there is $A = \sum_{i=1}^{n} [t_A(x_i), 1 - f_A(x_i)]/x_i$, $x_i \in X$; When X is a continuous space, there is

$A = \int [t_A(x), 1 - f_A(x)]/xd_x$, $x \in X$. By definition of Vague Set, the membership function of x is limited on [0, 1]. And $t_A(x)$ is the true membership function of the Vague Set A, it indicates the extent necessary to support the evidence of x ∈ A. $f_A(x)$ is the false membership function of the Vague Set A, it indicates the extent necessary to against the evidence of x ∈ A. $1 - f_A(x)$ indicates the possible extent to support the evidence of x ∈ A [1].

Vague value at the same time expresses the membership degree of support, against and the unknown extent of x ∈ A. Such as vague value of A at x is $[t_A(x), 1 - f_A(x)] = [0.5, 0.8]$, there is $t_A(x) = 0.5, 1 - f_A(x) = 0.8, f_A(x) = 0.2, 1 - t_A(x) - f_A(x) = 0.3$. It can be interpreted as: the degree which the element x belong to A is 0.5. The degree which the element x doesn't belong to A is 0.2. The unknown degree is 0.3. We use a voting model to explain: out of 10 votes, five people were in favor, two against and three abstaining. It shows that the connotation of Vague Set is much richer than the connotation of Fuzzy Set.

Definition 3 Vague relation data modelLet $A_i (1 \leq i \leq m)$ be the attributes which are defined in the domain U_i. Vague relation(r) which is defined in the relational schema $R(A_1, A_2, \ldots, A_m)$ can be seen as a vague subset of the Cartesian product of these attributes domain, namely:

$$r \subseteq V(U_1) \times V(U_2) \times \cdots \times V(U_m)$$

$V(U_i)$ represents the collection of all the vague subset in domain of U. Each tuple in the relationship r is composed by the Cartesian product of the vague subset, $t[A_i] = \pi[A_i]$, $\pi[A_i]$ is a vague subset of the attributes defined in the domain of U_i. In this way, the relationship (r) can be expressed as a relational table with m attributes. It should be noted that: different with classical relation and fuzzy relation, vague relation is a collection of Vague Set. Vague relation is the extension of the classical relations and fuzzy relations [2].

91.2.2 Example of Vague Relation

Classic database corresponds to a state in the real world. Due to the vagueness of the information, vague database may correspond to various states in the real world. Vague relation teacher1 as shown in Table 91.1 can be interpreted as 16 ($2 \times 2 \times 2 \times 2$) states of the possibilities. If we set length of service in the first tuple of Table 91.1 as 9, set length of service in the second tuple as eight, set

Table 91.1 Vague relation _teacher1

Name	Length of service	Professional titles
Zhang Sang	[0.7,0.9]/9 + [0.6,0.9]/10	Professor
Li Si	[0.5,0.6]/8 + [0.7,0.8]/9	[0.7,0.8]/Professor + [0.4,0.5]/associate professor
Wang Wu	[0.4,0.6]/6 + [0.7,0.9]/5	Lecturer

professional titles as associate professor, and set length of service in the third tuple as 5, then one possibility of state corresponding to the possible degree is [0.7, 0.9] ∧ [0.5, 0.6] ∧ [0.4, 0.5] ∧ [0. 7, 0.9] = [0.4, 0.5].

91.3 Vague Temporal Data

A variety of events and potential relationships between entities are often implicit in the temporal information. Temporal modeling is an important issue for many applications. However, the information of many events in real life is imprecise. In particular, due to the uncertainty of the inherent time of the historical events, the historical temporal information is described frequently in uncertain and ambiguous language.

91.3.1 Temporal Data Schema

For more intuitive to express the interval relationship of historical events, the real axis of mathematics corresponds to the timeline. R represents timeline. According to actual needs, milliseconds, seconds, minutes, days, months can be the time granularity. The timeline is divided into many small segments, each of length equal to the granularity needed to express the smallest unit of time, known as the time point. The time point of temporal object can be determined or uncertain. The determined time point is a special case of the uncertain time point. Types of relationship between the time points include the quantitative relationships and qualitative relationships [3].

The time interval is the time period divided by two time points on the timeline, T = < Tb, Te > and T* = < Tb, Te > represent determined and uncertain time interval. According to uncertainty interval endpoints, T* = < Tb, Te > is divided into three kinds: T* = < Tb*, Te > , T* = < Tb, Te* > , T* = < Tb*, Te* >. '*' is called fuzzy operator, which represent the uncertainty of the time interval and time endpoint. Tb and Te represent the time interval endpoint.

Now there are dozens of representation schemes of temporal data, but there is no unified schema [4]. Temporal data model is generated on the basis of the relational model, which adds temporal information. This is a very complicated process. Development of temporal database model should be combined with the practical application of the following four aspects to consider:

- What time scale is supported

According to what time scale is supported and the views of spipada's and sno dgrass', temporal database can be divided into three categories according to the functions: transactional database, historical database, temporal database [5].

- Tuple time scale or attribute time scale

Tuple time scale maintained the simplicity of the relational model, and easy to achieve high query efficiency. Attribute time scale will remain the value of an object in a tuple. Therefore, operational performance cannot compare with tuple. Tuple time scale may cause the storage redundancy, so it should be need to vertical decompose the attributes.

- Event-based time scale storage or state-based time scale storage

Event-based time scale storage is a record of the incident time point, and state-based time scale storage is the whole process of the state record with a time period. Expressions of these two forms have the corresponding advantages and disadvantages.

- Time granularity and time model

There is a variety of time granularity. Some transaction time may be accurate to the second. Some effective time may be accurate to the day. And some systems' effective time is accurate to month. Transaction time and valid time in temporal database is multi-granularity.

Considering the above four factors, a subject of tuple-based representation scheme can be expressed as:

$$R = (a1 \ldots a_n, T\ s, Te, Vs, Ve)$$

R is the snapshot relational model; a1,…, an expressed explicitly attributes; Ts, Te, Vs, Ve is the time scale properties of the atomic values; The meaning of the expression is: Ts expresses the transaction start time; Te expresses the transaction end time, and Vs expresses the state start time; Ve expresses the state end time [6].

91.3.2 Vague Temporal Database

Vague temporal database is based on vague technology. Vague temporal Not only describes a moment of fuzzy data, but also reflects its history and reveals its future. Similar to general temporal database, vague temporal database can be divided into historical database, transaction database and bitemporal database.

- Transaction Database

Transaction database supports transaction time, and addresses according to the transaction time, and saves all the states of past evolution. The relationship of the transaction is a three-dimensional structure, constituted by the tuples, attributes, and transaction time.

- Historical Database

The historical database is similar to transaction database, but time used in the historical database is the valid time, not the transaction time. The relationship of

Table 91.2 Vague temporal relation_teacher2

Name	Length of service	Professional titles	Duty period
Zhang Sang	[0.7,0.9]/9 + [0.6,0.9]/10	Professor	1997–2000
Li Si	[0.5,0.6]/8 + [0.7,0.8]/9	[0.7,0.8]/Professor + [0.4,0.5]/associate professor	1998–2003
Wang Wu	[0.4,0.6]/6 + [0.7,0.9]/5	Lecturer	1996–1997

the historical database is a three-dimensional structure, constituted by the tuples, attributes, and valid time.

- Temporal Database

Temporal database not only manages the history of object, but also manages the history of the database itself, so the temporal database also known as bitemporal database. Temporal database has the advantages of the former two, which support transaction time and valid time. The temporal relationship is a four-dimensional structure, composed of tuples, attributes, transaction time and valid time [7].

In order to make the fuzzy information more accurate, Vague Set is used to represent the fuzzy attribute of the database. The objective world is a four-dimensional world, every thing has its time dimension, and so a time dimension 'duty period' is added in vague temporal relation_teacher2. The database which is a historical database only uses effective time, Table 91.2:

The value of attribute 'duty period' is an interval or a collection of intervals, called tuple valid time or life cycle. In vague temporal database management system, the value should be managed by the system, rather than as a normal field to be managed.

When a tuple is filled into the vague temporal relation, the life cycle begins. When a tuple is deleted, a moment or the current time is specified as the end point of the life cycle of the tuple, the tuple does not delete physical.

In general, the history in the vague temporal database can not be deleted and modified, only can be inserted and queried. Comparison of Tables 91.1 and 91.2, the introduction of the life cycle has the following advantages: reduce redundancy, improve accuracy, and Table 91.2 can be derived from Table 91.1.

Attributes in Vague temporal database can be extended, which became temporal attributes. A temporal attribute can be two parts of an ordinary attribute and a time interval. The life cycle is a subset of the system time domain, and after set operations, the value of the life cycle is still the life cycle [8].

91.4 Conclusion

The database model based on Vague Set and the database model based on Fuzzy Set are compared; the former has a more fuzzy ability to express. Each data and tuple in the database can be reflected from the three sides of true, false and the

unknown extent. In the field of database system theory, the study on the temporal database based on Vague Set is in its infancy. Now there is no forming model of the temporal database based on Vague Set, and there are few articles in this field. The theory of temporal database based on Vague Set has not formed a system.

The further expansion of vague database model is proposed which is temporal database based on Vague Set. This research has laid the foundation for further research into data model, relational algebra, database management system of temporal database based on Vague Set, Etc.

The combination of Vague Set and temporal database technology can get more realistic description of the fuzzy temporal data in real-world. There will be a broad application prospect. The combination of Vague Set and temporal database technology is an important development direction in information systems. Belonging to a new area of research, it is also an important area of research and it will draw more attention. A lot of work needs to be studied in depth.

References

1. Zhou, X., Tan, C., Zhang, Q.: Decision theory and methods based on vague set. Science Press, Beijing, pp. 3–5 (1994)
2. Zhao, F., Ma, Z.: Aggregate operations in vague relational data model. J. Northeast. Univ. (Natural Science), **27**(12), 1331-1334 (2006)
3. Deng, L., Yang, S., Bian, L.Y.: 1NF Fuzzy temporal database model. J. Shenyang Univ. Constr. (Natural Science), 24(3), 503–507 (2008)
4. Peng, H.: The study on time of uncertainty of temporal data model. Yanshan University, Qinhuangdao (2007)
5. Shi, S., Dong, R., Yang, Y.: Temporal databases in e-commerce. Comput. Inf. Technol. **3**, 25–27 (2008)
6. Tang, Y., Xiao, F., Wei, J.: Construction of temporal GIS in the double-query mode. Inf. Technol. Inf. **1**, 35–45 (2008)
7. Yao, C., Hao, Z.: A T3NF decomposition algorithm of many temporal granularity temporal models. Mini-Micro Comput. Syst. **26**, 1530–1535 (2005)
8. Jiang, X., Jiang, X., Zhou, Y.: Research progress on temporal database. Comput. Eng. Appl. 27–31(2005)

Chapter 92
New Regional Investors Discovery by Web Mining

Ting Chen, Jian He and Quanyin Zhu

Abstract In order to promote micro business and sell their products and services, a new proposed system by web mining technology is used to discover new regional investment project. Project information published on the government websites of Huaian, Jiangsu province is extracted to utilize the proposed method. Python language, MySQL database and Django web framework are used to develop the application system, and the multi-factor matching algorithm is provided to collect key information of the project name, time, address, contact, domain and URL by web mining. Furthermore, the location and statistics functions are accomplished in the proposed system which can meet application requirements of micro business.

Keywords Investors information discovery · Micro business · Web mining · Python language · MySQL database · Django web framework

92.1 Introduction

In recent years, the rapid development of Internet brings geometric growth to web information, and the vast volume of information means it has the characteristic of polynary and redundant as well. Web pages cannot be directly made use of by traditional database systems for its semi-structured characteristic [1]. How to use the information better becomes the focus of attention. At present, most web pages are given to the semi-structured document in HTML form. Web documents can be represented as unstructured documents, semi-structured document and structured document. For the correct extraction of web information, a lot of work has been done at home and abroad. References [2] proposed that results of participle

T. Chen (✉) · J. He · Q. Zhu
Faculty of Computer Engineering, Huaiyin Institute of Technology,
Huaian, China
e-mail: apple_ting@126.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_92,

algorithm could touch the shopkeepers' minds, and it can support the originality data for the commodities markets and dynamic trend analysis. Reference [3] describe the new API for data mining proposed by Microsoft as extensions to OLE DB standard. Reference [4] describe data mining and brainstorms on its application to power systems. Reference [5] supported some actions to promote information and communications technologies in the autonomous community of the region of Murcia of Spain. Reference [6] analyzed the impact of urbanization on regional flood risk in the Qinhuai River Basin of China. Reference [7] applied research of the ash connection analysis in highway's influence of regional economies. But all of them not reported the useful method to discover new regional investment project using web mining technologies. So our proposed is how to find those new regional investment project and help the modern micro business to solve their concerned at the recent development.

Based on our past work [8–12], we select investor's discovery and web mining pages from the well-known colleges and universities in China in order to build an efficient system of new regional investor's discovery by web mining and study the Multi-factor matching method for information ex-traction. The second part gives the system architecture for new regional investor's discovery. In the third part we introduced web crawler strategy of webpage search. Multi-factor matching algorithm is proposed in the fourth part. In the last two parts we gave application system structure and result of system running.

92.2 System Design

92.2.1 System Framework and Development of Language

How to seek for specific investment projects investors from the web, mining data, and store in the database is serious problem for micro businesses. Fast and accurately discovering relevant project information and contacting information is very conducive for micro businesses to publicize and sell their products or services. To build a data mining system, the system should conclude the following functions:

1. Automatically mining the release date, URL and related data information. Such as, project overview, project introduction, project approval document and so on.
2. According to their own requirements and keyword, querying the data in the database easily through software interface.
3. Study of the basic knowledge of Python language.
4. Study of re-module in the regular expression matching.
5. Study of Django web development framework.
6. Study of MySQL data construction and query.

Django's advantage is simply and rapidly developed database-driven website. It emphasizes the importance of code reuse. Multiple components can be very convenient serving for the whole frame as plug-in form. Django has many powerful the third party plug-in. It makes Django Strong expansibility. It also stressed the rapid development and the principle of DRY.

With Python class form defining data model, ORM related model and database together. You will get a database API using very easily, and also you can use the original SQL statement in the Django language.URL assignment use regular expression matching URL. You can design URL random without specific limited framework.

Using Django powerful and extensible template language, the template system can be separated from design, content and Python code. And it also has inheritance.

92.2.2 Web Crawler

Web crawler crawls web information automatically following certain rules. Webpage search strategy can be divided into depth-first, breadth-first and best-first. Depth-first in many cases will lead to crawler trapped problems. And breadth-first and best-first are popular methods.

Breadth-first search strategy means that next level search must be after the completion of the current level in the crawl process. The algorithm design and implementation is relatively simple. Algorithm flow is shown as Fig. 92.1.

Best-first search strategy according to the web analytics algorithms to predict the similarity of the candidate URL and landing page, or correlation with the theme, and select the best rated one or several URLs to crawl. The only access through web analysis algorithm predicted "useful" pages. One problem is that

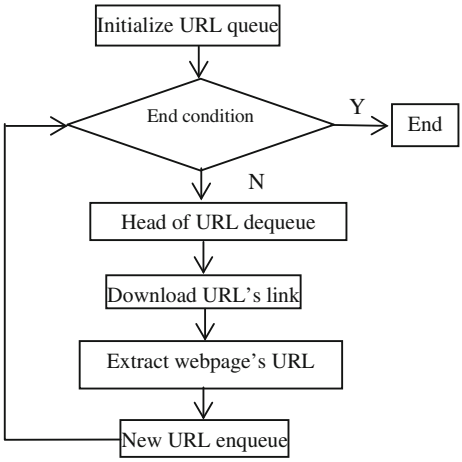


Fig. 92.1 Breadth-first search algorithm flow

reptiles crawl path on many relevant pages may be ignored, because the best-first strategy is a local optimum search algorithm. Therefore it needs to be combined with the best-first specific application to improve, to jump out of local minima.

92.2.3 Multi-Factor Matching Algorithm

This paper focuses on the example of chart showing for science expert information extraction.

Let the view-source of webpage be defined as D, body of the page be defined as S:

$$S \subset D \tag{92.1}$$

Let the normalized text be defined as $\hat{S}$:

$$\hat{S} \subseteq S \tag{92.2}$$

Keywords corpus consists of some self-learned keyword. Let the keywords corpus be defined as F, then F be presented as

$$F = \{f_1, f_2, \ldots, f_n\} \tag{92.3}$$

In fact, the expert information field included in F. Let it defined as f_n, the position of f_n in $\hat{S}$ is defined as k:

$$k = (f_n, \hat{S}) \tag{92.4}$$

Let the potential expert information field be defined as t, define a constant as con then t is equals to the normalized text range between k and $k + con$:

$$t = \hat{S}(k, k + con) \tag{92.5}$$

Finally, because of other fields affect, we must remove the affect. The positions of F except f_n is defined as K:

$$K = \{k_1, k_2, \ldots, k_n\} \tag{92.6}$$

Let define the min position in K as $k_{\min}$, define the real expert information as t_{real}, we can conclude that the real expert information filed:

$$t_{real} = t(0, k_{\min}) \tag{92.7}$$

92.2.4 Data Table

System data sheet is mainly used for storage government network project information and credit information publicly shared on all project information in the column. And it could be called and inquired by the system (Table 92.1).

92.2.5 System Data Structure

Application system design includes eight sheets, respectively as the auth_message sheet, django_conten_type sheet, auth_user sheet, auth_permission sheet, auth_user_user_permission sheet, auth_group_permission sheet, auth_user_group sheet and auth_group sheet. Its structure is shown in Fig. 92.2.

The data structure of application system is succinct than the other system. Our aim is adequate to the boss of micro business for application requirements. If using the web service technology, we can get the encapsulation web service [11, 12] for other application system and get the advantage of platform irrespective.

92.3 System Implementations

Each hyperlink function in the interface of the application system as follows: URL: Mining the demand URL in the webpage, and store in the URL sheet of the database. Data: Release date of mining project, and store in the date sheet of the database. Info: mining project summary, approval documents, posting date, and store in the Huaian sheet of the database. Query: query the project data and URL in the database. Project Analysis: Building materials can be divided into structural materials, decoration materials and some special materials. Structural materials include wood, bamboo, stone, cement, concrete, metal, brick, ceramics, glass, engineering plastics, composite materials, etc. Decorative materials include a variety of coatings, paints, coatings, veneer, colored tile, glass and other special effects. Special material includes waterproof, moisture-proof, anti-corrosion, fire,

Table 92.1 System data table

Name	Type	Length
id	int	10
doc	varchar	500
proview	varchar	1000
date	varchar	20
url	varchar	50
date2	varchar	50

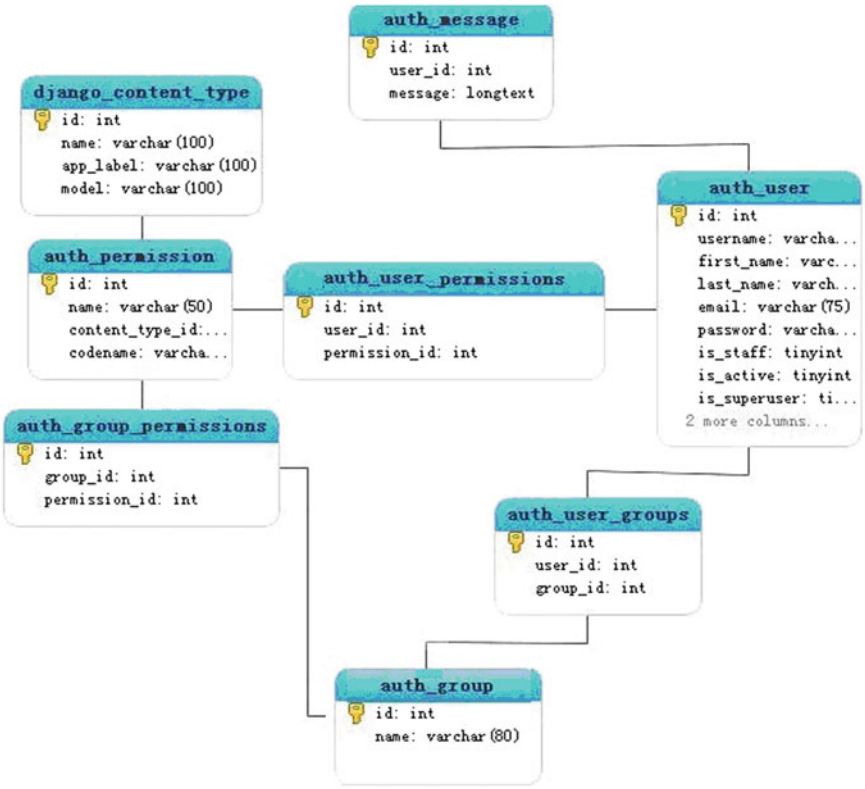


Fig. 92.2 Data structure of the application system

flame retardant, sound insulation, heat insulation, thermal insulation, sealing and so on. Figure 92.3 illustrates the connection information in a part of the URL database.

With the using of software, micro business based on building materials can find associated with their investment projects, promote and sell their products and services, thus derive profits to promote their own development.

Take Huaian of Jiangsu province for example, with the investment project comparison of nine counties in Huaian, micro business get an overall Huaian investment regional distribution map. With it, micro business promotes and market their product or service through its own geographical advantages. Figure 92.4 shows the investment project numbers of each region.

Take the industry for example, culture construction is the development of education, science, literature and art, the press and publishing, radio, television, sports and public health, library, museums and other cultural undertakings of activities. Micro business related with the cultural construction, find suitable for their own projects according to the search results, thus promote and sell their

id	url2
1	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=15257
2	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=15255
3	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=15260
4	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=15256
5	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=15101
6	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14747
7	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14694
8	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14418
9	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14051
10	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14112
11	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14116
12	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14120
13	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14108
14	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14106
15	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14100
16	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14098
17	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14118
18	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14104
19	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14053
20	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14102
21	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14122
22	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14114
23	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14110
24	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14124
25	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=7247
26	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=6879
27	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=6824
28	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=6629
29	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=6569
30	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=6570
31	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=6571

Fig. 92.3 Portion of URL extracted from the database

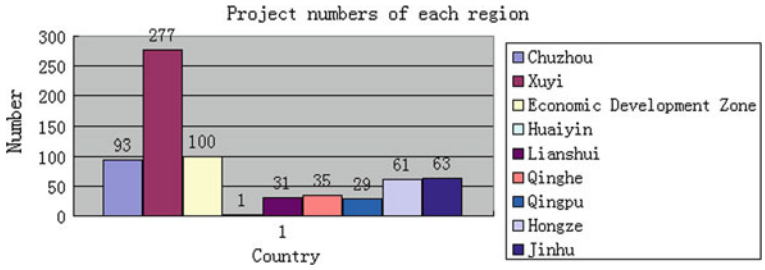


Fig. 92.4 Statistical data of nine counties

products or services. Figure 92.5 shows the investment project status according to the industry classification.

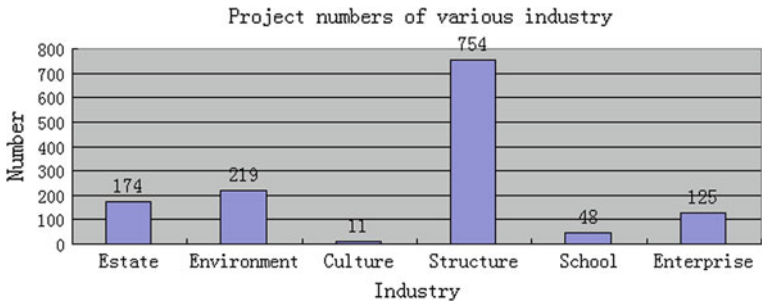


Fig. 92.5 The investment project status of various industry

92.4 Conclusion

Seeking for specific investment projects investors from the web, mining data, and storing in the database are serious problems for micro businesses. With the technology of data mining, discovering the new regional investment projects targeted the promotion of enterprise products or services can reduce the operation cost of small and micro businesses. The proposed systems can satisfy small and micro businesses for informatization and can meet market requirements.

Python language, MySQL database and Django web framework are used to develop the proposed system which can extract the region's investment project information from government websites. All the information of investment project can be stored in the database.

References

1. Li, Y.G., Sun, H.Y., Lin, S., et al.: Web information extraction based on hidden Markov model. In: Proceedings of the14th International Conference on Computer Supported Cooperative Work in Design, pp. 234–238 (2010)
2. Hoffman, P., Grinstein, G., Marx, K., et al.: DNA visual and analytic data mining. In: Procedings of the Visualization 1997, pp. 437–441 (1997)
3. Netz, A., Chaudhuri, S., Fayyad, U., et al.: Integrating data mining with SQL databases: OLE DB for data mining. In: Proceedings of 17th International Conference on Data Engineering, 2001, pp. 379–387 (2001)
4. Madan, S., Won-Kuk, S., Bollinge, K.E.: Applications of data mining for power systems. In: Proceedings of the IEEE 1997 Canadian Conference on Electrical and Computer Engineering, pp. 403–406 (1997)
5. Escudero-Sanchez, M., Pavn-Mario, P., Fernandez-Caceres, J.L.: Some actions to promote information and communications technologies in the autonomous community of the region of Murcia (Spain). In: Proceedings of the 11th Mediterranean Electrotechnical Conference, pp. 163–167 (2002)
6. Shi, Y., Xu, Y.P., Cai, J.: Analysis of the impact of urbanization on regional flood risk: A case study in the Qinhuai River Basin, China. In: Proceedings of the 19th International Conference on Geoinformatics, pp. 1–6 (2011)

7. Zhao, Q.W., An, Y.H.: The applied research of the ash connection analysis in highway's influence of regional economies. In: Proceedings of the Seventh International Conference on Fuzzy Systems and Knowledge Discovery 2010, pp. 1073–1076 (2010)
8. Zhu, Q.Y., Zhou, P., Cao, S.Q., et al.: A novel RDB-SW approach for commodities price dynamic trend analysis based on web extracting. J. Digit. Inf. Manag. **10**(4), 230–235 (2012)
9. Zhu, Q.Y., Yan, Y.Y., Ding, J., et al.: The case study for price extracting of mobile phone sell online. ICSESS 2011, pp. 282–285 (2011)
10. Ding, J., Zhu, Q.Y., Zhou, L.J., et al.: Research on the new products discovery based on web mining. MINES 2011, pp. 528–532 (2011)
11. Ding, J., Wu, B., Ding, T.T., Zhu, Q.Y.: The case study on service encapsulation for web-based application system. CSSS 2012, pp. 2684–2687 (2012)
12. Zhu, Q.Y., Zhou, H.Y., Yan, Y.Y., et al.: Research on the service encapsulation for web-based system. ICCDA **2011**(4), 535–538 (2011)

Part V
System Identification

Chapter 93
Enhancing Ability of Fault Detection for Component Systems Based on Object Interactions Graph

Fuzhen Sun, Lejian Liao, Jianguang Du and Guoqiang Li

Abstract Test case prioritization is a technique to schedule the test case in order to maximize some objective function. Early fault detection can provide a faster feedback generating a scope for debuggers to carry out their task at an early stage. In this paper, a method is proposed to prioritize the test cases for testing component dependency in a Component Based Software Development (CBSD) environment using Greedy Approach. The OIG (Object Interaction Graph) is traversed to calculate the total number of inter component object interactions and intra component object interactions. Depending upon the number of interactions, the objective function is calculated and the test cases are ordered accordingly. This technique is applied to the components developed in Java for a software system and found to be very effective in early fault detection as compared with non-prioritize approach.

Keywords Test case prioritization · Fault detection · Object interaction graph (OIG) · Component based software development (CBSD)

93.1 Introduction

A technique like test case prioritization has to be devised, which will lead to early fault detection. Test case prioritization aims at finding an execution order for the test cases which maximizes a given objective function. Among the others, the most

F. Sun (✉) · L. Liao · J. Du · G. Li
Beijing Engineering Research Centre of High Volume Language Information Processing & Cloud Computing Applications, Beijing Key Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing, China
e-mail: 10907023@bit.edu.cn

F. Sun
School of Computer Science and Technology, Shandong University of Technology, Zibo, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_93,

important prioritization objective is probably discovering faults as early as possible, which refers to maximize the rate of fault detection.

The major challenges in Component Based Software Development (CBSD) are testing component dependency. CBSD uses the reusable components as the building blocks for constructing the complex software system (component based system). Component based system promotes the software quality and productive. This building block approach has been increasingly adopted for software development, especially for large-scale software systems.

Previous work on test case prioritization [1–5] is based on the computation of a prioritization index, which determines the ordering of the test cases (e.g., by decreasing values of the index) [6, 7]. Srivastava [8] suggested prioritizing test cases according to the criterion of increased Average percentage of Faults detected (APFD) value. Rothermel et al. [9] have described several techniques for test case prioritization and empirically examined their relative abilities to improve how quickly faults can be detected by those suites. More importance is given to coverage based prioritization here [10, 11].

In this paper for describing each component we have taken the help of sequence diagrams, then a Object Interaction Graph (OIG) from sequence diagrams is constructed which shows the interrelation among the components. A new test prioritization algorithm is presented which is applied on OIG to count the maximum number of inter component interactions and intra component interactions made by the test cases.

93.2 Proposed Test Case Prioritization Model

In CBSD Component interface is defined as the only way that a component communicates with the external environment. There are two kinds of interface: service providing and service required. When the services are provided by an interface, it is called service providing interface and when the interface of a component requiring a service it is called service required interface. All components should be plug-compatible i.e. a service required interface can be connected to a service providing interface. We have defined a Component as follows: Component C = (P, R), where P = $P1$, $P2$, Pn is the set of providing services interface, R = $R1$, $R2$, Rm is the set of required services interface. The providing and required services of a component C is denoted by C.P and C.R respectively and $C.P \cap C.R = \emptyset$.

In Fig. 93.1 the required services of $C1P$ C2 are the union of C1.R1 and C2.R2 with the remove of satisfied services in S. With the definition of composition the providing and required services are propagated to the interface of composed component, so the composition could be carried parallel. A Component interaction graph (OIG) is used to describe the interrelation of components. A complete component interaction graph (OIG) makes the testing quite easy.

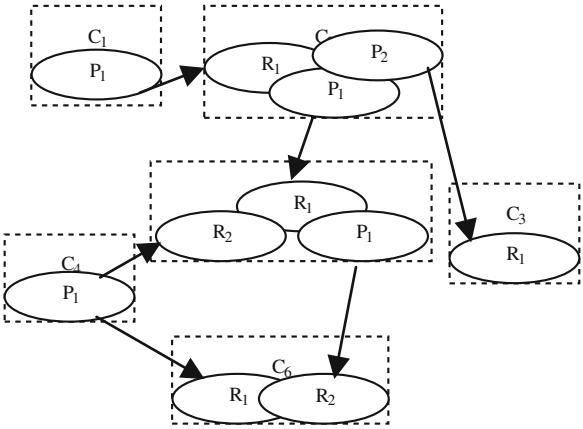


Fig. 93.1 Object interaction graph(OIG)

To facilitate regression testing by optimizing the time and cost, we propose a method to prioritize the test cases by using model based prioritization method by extracting the benefits of Unified Modelling Language (UML). UML provides lifecycle support in software development and is widely used to describe analysis and design specifications of software. It is a big challenge to study the test case generation from UML diagram (Fig. 93.2).

We have used sequence diagram from the set of diagrams present in UML 2.0. As Sequence diagram represents various object interactions through message passing, it can act as an input to the proposed model. We are generating an OIG from the sequence diagrams present. The methodology we have used for generating the graph has been discussed in Sect. 93.2.1 Further in Sect. 93.2.2 we have discussed how to traverse the OIG to calculate the number of inter component object interaction and intra component object interaction. Section 93.2.3 describes about objective function evaluation and the prioritization technique.

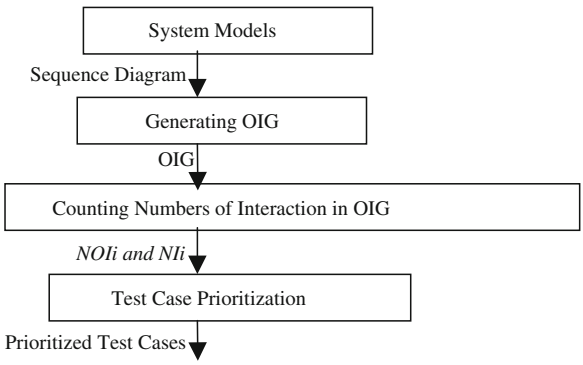


Fig. 93.2 A frame work for generating prioritized test cases

93.2.1 Generating OIG Form System Models

We have used sequence diagram for system modeling, and the object interactions can be very well identified by using a sequence diagram. During regression testing any modification in the code will have no effect on the sequence diagram. The object interaction can be categorized into two different types. One of them is intra component object interactions and the other one is inter component object interactions.

Sequence diagrams in UML are used to model how an object communicates with other objects in its life time. A complete object interaction graph (OIG) makes the testing quite easy. An OIG is a directed graph where OIG = (V, E), V represents a set of nodes. For generating Object Interaction Graph (OIG), each object present in the sequence diagram is represented as a node in the graph. The intra component object interactions form the edges of the graph and represented in Solid arrows. The inter component object interactions form the edges of the graph and represented in Dashed arrows.

Algorithm: GENERATE OIG

Input: Sequence Diagrams of various components of the system representing message passing between objects

Output: Object Interaction Graph (OIG)//It is a directed graph

1. Initialize OIG to be empty
2. For i = 1 to n//n is the total number of objects
3. Add a node N_i to OIG == N_i represents ith node.
   Object shared by different components treated as a single node.
4. For i = 1 to n
5. For j = 1 to n
6. For each incoming message from object O_i to O_j == All guard conditions are ignored
7. if (interaction types == intra) Establish an edge between O_i to O_j (i.e. N_i and N_i) and represent it as"Solid arrow" as well as append the pre and post conditions.
8. Else Establish an edge between Oi to Oj (i.e. Ni and Nj) and represent it as "Dashed arrow" as well as append the pre and post conditions.
9. The possible start and end of the scenario sequences are represented with solid arrows.

93.2.2 Traversing OIG

When the OIG is generated from the system models, it has to be traversed to count the number of inter component and intra component object interactions. NOI_i represents the number of Object Interactions discovered by test case t_i with in one component of the software and NI_i represents the number of Object Interactions

discovered by test case t_i between two different components of the software. We follow the depth first search (DFS) methodology for traversing the graph. The type of interaction is decided depending upon the color of the edge in the graph. If the edge color is found to be "Solid arrow", it represents an intra component object interaction, where as edges colored as "Dashed arrow" represents inter component object interaction

Algorithm: IN_CALCULATE
Input: Test case t_i & Object Interaction Graph (OIG)
Output: NOI_i and NI_i

1. Initialize both NOI_i and NI_i to 0.
2. Traverse each interaction in the OIG for *ti* in DFS
3. If (edge color ==‘VISITED’ && current edge is not visited already)
4. $NOI_i + +$ // Increment the value for intra component interaction
5. Else
6. $NI_i + +$ // Increment the value for inter component interaction
7. Return NOI_i and NI_i.

93.2.3 Generating Prioritized Test Cases

Once we get the value for NIi and NOIi by using the algorithm described in Sect. 93.2.3, prioritization process starts. For each test case ti, the value of NIi and NOIi are added. We have considered the total number of intra component interaction where as the total number of inter component object interactions is found out by multiplying it with RP i.e. total number of providing service interface and required service interface. If the faults due to component integration are detected early, it will give a better coverage. The added result is divided with unit time U to determine value of the objective function i.e. factor criteria FCi. We try to maximize the objective function using a Greedy approach.

Algorithm: TEST_PRI
Input: Regression Test Suite T
Output: Prioritize Test Suite T’

1. Traverse the test suite T, for each test case t_i present, call **IN_CALCULATE** (t_i) to calculate NOI_i and NI_i
2. Define some unit time U
3. Calculate objective function (FCi) for test case t_i as $FCi = (NOI_i + \boldsymbol{RP}*\mathrm{NI}_i)/$ U. (1)
   ***//RP* represents total number of providing service interface**.
4. Generate T’ by Sorting the test suit T in ascending order of FC_i for each t_i.
5. Store T’ in the test case repository for regression testing.

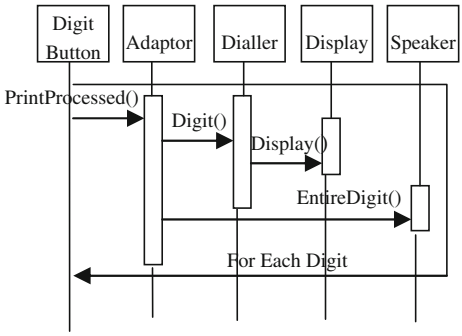


Fig. 93.3 Sequence diagram for dialing the number

93.3 Case Study: A Cellular Network Manager

We have taken the case study of a Cellular Network Manager to explain the proposed model. Two components i.e. "Dialing a Phone" and "Cellular Network Connection" have been presented.

From the sequence diagram of both the components given in Figs. 93.3 and 93.4, corresponding OIG are designed as given in Fig. 93.5.

Three test cases are considered to test the prioritization algorithm. The test cases are designed to test the Dialer Display (t1), to test the Speaker (t2) and to test the Cellular Radio Display (t3). Table 93.1 contains the value of NOIi, NIi and FCi. Here the unit time U is considered to be 1 unit.

From the Table 93.1 we conclude that the prioritized test sequence is: **t3, t2, t1 or t3, t1, t2** The proposed model found to be very effective as it increases the Average Percentage of Fault Detection (APFD) when it is compared with generalized model based method and few code based methods like LOC count and Function count. The comparison made is summarized in Table 93.2.

The cost and time required for regression testing can be minimized by using the prioritization technique discussed in this paper. Here we have proposed a model based prioritization method by considering the number of Object Interactions per unit time as the objective function. Here more importance is given to number of inter component object interactions present because maximum faults are expected

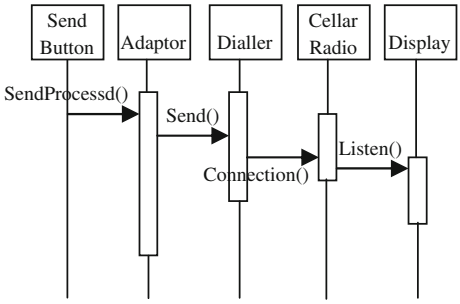


Fig. 93.4 Sequence diagram for cellular connection

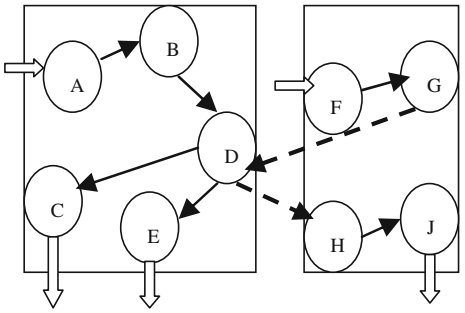


Fig. 93.5 OIG for a cellular network manager

Table 93.1 Objective function (FCi) evaluation

Test cases	NOI_i	NI_i	FC_i
t_1	3	0	3
t_2	3	0	3
t_3	2	4	6

Table 93.2 a comparative study

Name of prioritized technique	Approximate increase in APFD value (%)
Code based approach (LOC count, function count etc.)	30
General model based approach	35
Model based approach using the dependency criteria in CBSD	45

to be present when components interact with each other. The proposed model found to be very effective as it increases the Average Percentage of Fault Detection (APFD) when it is applied to few of the projects developed in Java by java 45–50 %.

93.4 Conclusion

In this paper, the authors have taken the help of sequence diagrams of describing each component to construct an Object Interaction Graph (OIG) which shows the interrelation among the components. Furthermore, A new test prioritization algorithm is presented which is applied on OIG to count the maximum number of inter and intra component interactions. The experiments show that this approach is mainly applicable to test the component composition in case of component based software maintenance.

The proposed method can further be extended to prioritize test cases to perform regression testing for real time systems and distributed systems. The authors are also working on adding new criterion like frequency of data base access number of state changes in UML state chart diagram.

References

1. Elbaum, Z.S., Malishevsky, A., Rothermel, G.: Test case prioritization: a family of empirical studies. IEEE Trans. Softw. Eng. **28**(2), 159–182 (2012)
2. Kim, J.M., Porter, A.A.: A history-based test prioritization technique for regression testing in resource constrained environments. In: Proceedings of the International Conference on Software Engineering (ICSE). ACM Press, pp. 119–129 (2011)
3. Rothermel, G., Untch, R., Chu, C., Harrold, M.J.: Test case prioritization. IEEE Trans. Softw. Eng. **27**(10), 929–948 (2012)
4. Srikanth, H., Williams, L., Osborne, J.: System test case prioritization of new and regression test cases. In: Proceedings of the 4th International Symposium on Empirical Software Engineering (ISESE). IEEE Computer Society, pp. 62–71(2005)
5. Thiagarajan, S.J.: Effectively prioritizing tests in development environment. In: Proceedings of the International Symposium on Software Testing and Analysis (ISSTA). ACM Press, pp. 97–106 (2009)
6. Kim, J., Porter, A.: A history-based test prioritization technique for regression testing in resource constraint environments. In: Proceeding of the 24th International Conference on Software Engineering, pp. 19–129 (2011)
7. Korel, B., Tahat, L., Harman, M.: Test prioritization using system models. 21st IEEE International Conference Software Maintenance (ICSM'05), pp. 559–568 (2005)
8. Korel, B., Koutsogiannakis, G., Tahat, L.: Application of system models in regression test suite prioritization. In: Proceeding of the 24th International Conference Software Maintenance (ICSM'08), pp. 247–256 (2008)
9. Li, Z., Harman, M., Hierons, R.: Search algorithms for regression test case prioritization. IEEE Trans. Softw. Eng. **33**(4), 225–237 (2007)
10. UML 2.0 Reference Manual, Object Management Group (2003)
11. Thiagarajan, S.J.: Effectively prioritizing tests in development environment. In: Proceeding ACM International Symposium on Software Testing and Analysis, ISSTA-02, pp. 97–106 (2002)

Chapter 94
A Method of Deploying Virtual Machine on Multi-core CPU in Decomposed Way

Qing-hua Guan

Abstract Nowadays, with the development of multi-core and cloud computing technology, the deployment of virtual machine faces opportunities as well as challenges in the process of virtualization. However, most virtualization deployment only considers the concept of combining single vCPUs with multi-core CPU. Aiming at solving those known problems based on experience, this paper proposes a new method of deployment of virtual machine in a decomposed way. The result shows that optimized method is more reasonable for resource allocation. It can provide a good principle to expand future datacenter virtualization.

Keywords Virtualization · Multi-core · Virtual machine · Decomposed way

94.1 Introduction

With the further development of the information process, more and more physical machines have been installed and allocated in datacenter. The traditional deployment method brings us not only the worse management and the inefficient deployment, but also the waste of resources. The problem of unreasonable allocation and integration needs to be solved. Technical virtualization provides a good solution of combining computing resource which has already been applied widely. But most of the method only considers the concept of combining single vCPUs [1]. Such methods just increase the numbers of vCPUs for high load applications without considering a decomposed solution.

At the same time, multi-core CPU has been rapidly developed by chip manufacturers. INTEL\AMD\IBM have already pushed out their multi-core chips or

Q. Guan (✉)
Information Technology Center, China Guangdong Nuclear Power Holding Co., Ltd, Shenzhen, China
e-mail: guanqinghua@cgnpc.com.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_94,

multi-core architecture. The complexity of the CPU scheduling has increased. However, current operation system can't support the multi-core hardware very well. Programmers have to use parallel programming skills to improve the CPU utilization ratio. Obviously, such skills have not been widely used except for professional parallel computation. So Jin Hai concluded [2] some researches in multi-core scheduling field. Some researches give good inspiration for narrowing the semantic gap between virtual machine and host machine. Other researchers try to modify the kernel code of the operation system for multi-core CPU [3]. This solution can be realized on self-developed operation system, but for general operation systems, e.g. Windows, changing their kernel codes is somehow very difficult. So if we keep using traditional ways without any changes, the advantage of virtualization technology can't be utilized completely.

Given this situation, this paper proposes the concept of deploying virtual machine on multi-core CPU in a decomposed way. The experiment shows this method can make full use of its host physical CPU and accord with the multi-core CPU virtualization mechanism.

This paper is organized as follows: Section 94.2 introduces related work about the architecture of virtualization and the two-layer scheduling based on multi-core CPU. Section 94.3 gives the method of deploying virtual machine in a decomposed way. Section 94.4 presents the experiment design and result analysis. Section 94.5 presents the conclusion and future work.

94.2 Related Works

94.2.1 Architecture of Virtualization

Virtualization has been defined as a system which can isolate physical/virtual machine. Virtual machine monitor (VMM) is used to schedule and monitor virtual machine. All servers can be integrated by VMM and used by all virtual machines in the cluster of resource pool. This resource pool can support HA and dynamic motion characteristics for virtual machine. Like VMware, the classic architecture of virtualization is represented in Fig. 94.1.

As presented in Fig. 94.1, physical machines become resource pool for virtual machines. These virtual machines not only meet all functional and operational requirements like traditional servers, but also improve the efficiency of resource utilization, and consequently the operational cost is reduced as well.

94.2.2 Two Layers Scheduling Based on Multi-core CPU

The classical virtualization architecture can be divided into three parts: multi-core CPU, VMM and virtual machines. The operation system of virtual machine and

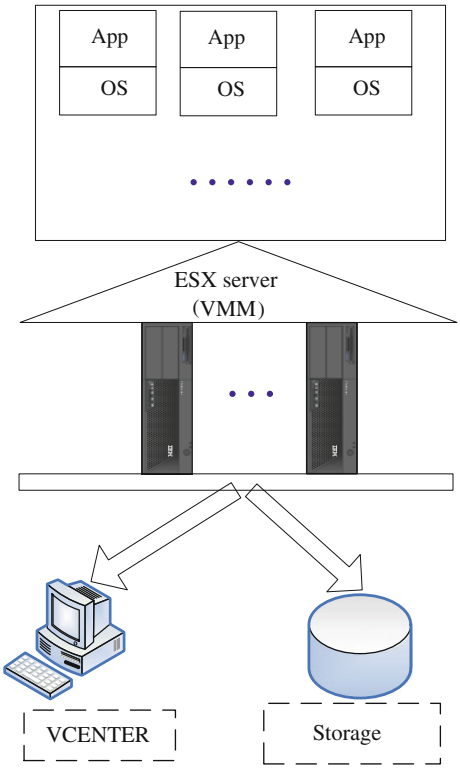


Fig. 94.1 Classical architecture of virtualization

host machine construct the two-stage scheduling framework. There is no connection between application threads and physical CPU scheduling in this framework (Figure 94.2).

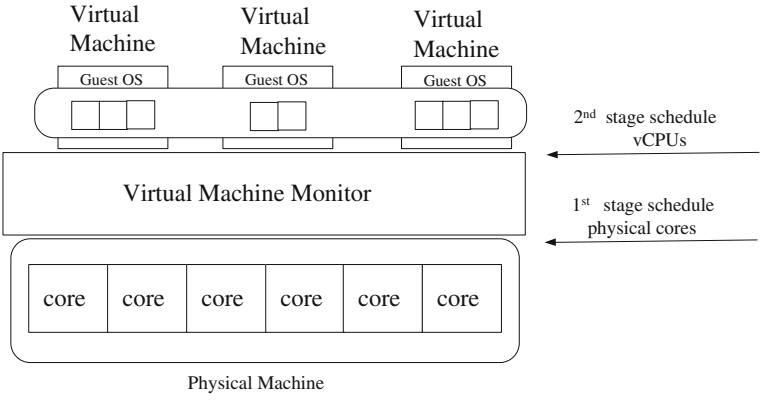


Fig. 94.2 Typical architecture of the scheduling system based on the multi-core processors

This may confront many challenges while scheduling application thread of virtual machine on physical CPU. Operation system has been designed to own entire privilege of physical CPU, but instructions from applications are allocated to some physical CPU cores. The mapping between application threads and physical CPU is difficult to establish in this framework. Due to the semantic gap, it is not easy for VMM to catch the work load of virtual machine. Scheduling of cache sharing, I/O blocking and job distribution is hard to be done in VMM [4].

So the concept of adopting schedule strategies and algorithms of existing operating systems without any modifications can lead to drastic degradation of the system performance and needs to be reconsidered.

94.3 Deploy Virtual Machine in Decomposed Way

94.3.1 The Core Idea

Obviously, the high work load applications should get more computing resources. Current solutions are just to apply more CPUs to these virtual machines. Four slots with six-core CPU are more powerful than two slots with six-core CPU in physical machine while this may not be right in virtualization. Like VMware, a virtual machine has eight vCPUs (virtual CPU) unequal to summarizing eight virtual machines with one vCPU. VMM uses FCFS algorithm and rotation principle to share resources with all virtual machines. A virtual machine that needs 4 vCPUs can be satisfied only when there are 4 available vCPUs. Figure 94.3a presents the linear relationship between virtual machine computation and the numbers of vCPUs.

Figure 94.3a presented that the virtual machine computation will approach a limit with the increase of vCPUs. Current deploying method of virtual machines is not the good choice for virtualization. As presented in Fig. 94.3b, if doing migration of large UNIX business system on X 86 platforms, like SAP [5], the application server should be divided to several systems for seeking more computation from the resource pool. Of course, the application server should support the distribution architecture. The advantage of this deployment is that the availability of getting single vCPU is larger than that of several vCPUs. The entire computation of several servers is larger than that in non-distribution way which means $1 + 1 > 2$.

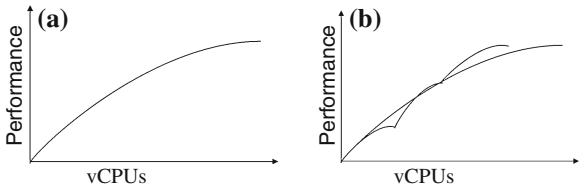


Fig. 94.3 Performance of virtual CPU

94.4 Experiment Results and Analysis

94.4.1 Test Platform

This paper uses three HP DL580 servers and VMware as the test platform. The server has two slots Intel E7440 CPUs and 16 GB memory. Two servers were used for VMware Vcenter and ESX. Another server will be installed with windows server 2003 directly. Virtual machines will be installed on VMware ESX and administrated by Vcenter (Table 94.1).

94.4.2 Evaluation Standard and Tools

Pass Mark is the important measure for CPU benchmark. Stress Prime is a good stress testing tool and can reflect real CPU pressure in our experiment. VMware Vcenter has a figure statistics function and can be used to give us the direct image in test scenario.

94.4.3 Experiment Design

In the following experiments, this paper will try to prove all physical CPUs can be shared by vCPUs and the computation of all virtual machine is larger than that of physical machine. The result will show the deployment of virtualization in a decomposed way is a meaningful consideration.

So the first experiment will start the virtual machine and then get the CPU statistic figure from VMware Vcenter. This will show whether CPUs of physical machine can be shared by vCPUs of virtual machine.

The second experiment will start one virtual machine with half computation of its host machine (physical machine) and use Pass Mark to calculate the detail values. This will show whether the decomposed way is the right choice.

The third experiment will start two virtual machines which have the same CPU and Memory configuration but give different proportions (4:3). This will show whether the VMM need more resources with the increase of virtual machine quantity. Different proportion is used to check the effectiveness of VMware adjustment mechanism.

Table 94.1 Experiment platform

	Type	CPU	MEM	OS	Remark
Host machine	HP DL580G5	2*4 cores, E7440	16 GB	Windows 2003 server R2	
Virtual machines	VMware	4 cores	16 GB	Windows 2003 server R2	On same VMware ESX

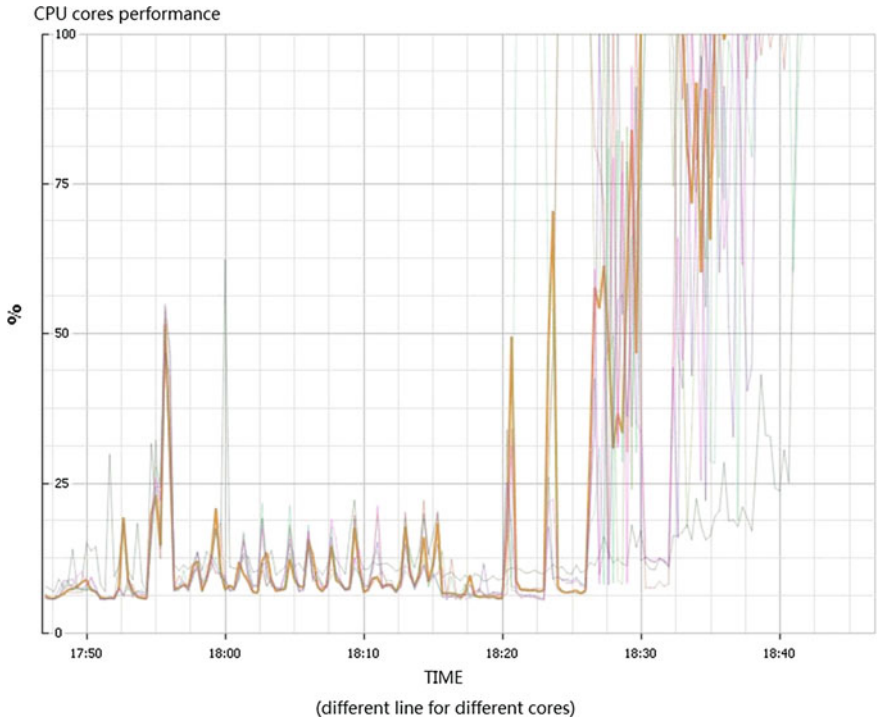


Fig. 94.4 Computation distribution of virtual CPU on different cores

94.4.4 Comparison Experiment and Analysis of Results

Figure 94.4 presented the result of the first experiment. The vertical axis stands for CPU computation and the transverse axis stands for starting different virtual machine in different time. The result in Fig. 94.4 shows that the computation of all virtual machine can be distributed among all physical CPU cores. The result also shows the linear relationship between virtual machine computation and the numbers of vCPUs.

The Pass Mark numbers of the second experiment is showed in Tables 94.2 and 94.3. Table 94.2 presents the score of virtual machine with half computation of its host machine. Table 94.3 presents the score of physical machine. In this case, virtual machine and physical machine get the same memory configuration for the least impact of non-CPU factors.

The result shows that all remarks of virtual machine are more powerful than that of 1/2 of its physical machine except "Encryption" item. The gap can achieve even 30 % of wider! Deploying virtual machine in a decomposed way gets more advantages.

The reason is that current operation system doesn't support multi-core CPU very well. On the other hand, VMM also need computation which makes the

Table 94.2 Result data of virtual machine

Item	SSE/ 3Dnow	Comp.	Encryption	Image rotation	String sorting	CPU mark	Memory mark
Virtual machine	13000	12000	70	3000	8000	4000	600

Table 94.3 Result data of physical machine

Item	SSE/ 3Dnow	Comp.	Encryption	Image rotation	String sorting	CPU mark	Memory mark
Physical machine	20000	23000	200	5000	15000	6000	700

Table 94.4 Result data of two virtual machines

Item	SSE/ 3Dnow	Comp.	Encryption	Image rotation	String sorting	CPU mark	Memory
Virtual machine 1	13700	14000	130	4000	11400	4500	310
Virtual machine 2	10000	11000	100	3000	8500	3400	250

memory score less than 14 % score of its physical machine. More memory is good for virtualization.

As presented in Table 94.4 of the third experiment, the summary of virtual machine one and two is larger than that of the physical machine (presented in Table 94.3) again. The proportion of these scores (nearly 4:3) confirms VMware adjustment mechanism. The memory is still less than 15 % of its physical machine which means VMM needs constant resources to run its scheduling job [6].

94.5 Conclusion and Future Work

Aiming at solving the problem that current method of deploying virtual machine doesn't include the multi-core situation, this paper proposes a decomposed method for virtualization. The optimized solution highlights the fact that computation of vCPU is distributed on all multi-cores. Experiments show that the method can provide a better result and more flexibility comparing to using the traditional deploying method.

Future work includes the following two aspects. First, deploying in a decomposed way can't be boundless. So we need to deal with the decomposed degree level to avoid resource wasting. Second, heterogeneous multi-core CPU also needs to be tested to get more precise results.

References

1. VMware. Virtualizing Business-Critical Applications on vSphere, pp. 31–35 (2012)
2. Hai, J., A-lin, Z., Wu, S.: Virtual machine VCPU scheduling in the multi-core environment: Issues and challenges. J. Comput. Res. Dev. **48**(7), 1216–1224 (2011)
3. Li Y.-d, Hang, L.: Survey of multi-core operating system. Appl. Res. Comput. **28**(9): 3215–3219 (2011)
4. Kim, H., Lim, H., Jeong, J., et al.: Task-aware virtual machine scheduling for I/O performance . Proceeding of VEEp09. ACM, New York, pp. 101–110 (2009)
5. Henter, P. Virtualization of SAP applications with VMware vSphere 5 on IBM puresystems, 1.0, pp. 23–27 (2012)
6. Grund, M., Schaffner, J., Krueger, J., Brunnert, J., Zeier, A.: The effects of virtualization on main memory systems. In Proceedings of Sixth International Workshop on Data Management on New Hardware, New York, pp. 41–46(2010)

Chapter 95
An MDA Based Widget Development Framework

Peng Xiao, Minghui Wu, Bin Peng and Jing Ying

Abstract The paper aims to solve the repeated work problem existing in multi-platform widget development and supply a widget development environment which can support for cross-platform development or cross-standards development. A widget development framework called Model-driven Widget Development Framework (MWDF) is proposed. By using MDA, the MWDF can make developer get a visual programming and model driven development environment, and the generated widget can be deployed on many platforms supported by MWDF. For illustration, the architecture of MWDF and implementation of model driven development module are shown to describe the framework. MWDF abstracts the implement in widget development which saves a lot of time for the developer and increases the expandability of the widget.

Keywords Widget · Model driven architecture (MDA) · Model transformation

95.1 Introduction

Widget is a small application that can be installed and executed within html-based environment by an end user, and it is typically created in DHTML, JavaScript. With the development of mobile internet and mobile terminal device, widget is entering the phase of rapid development. Nowadays widget application should support for accessing to device local resources besides invoking DOM APIs to

P. Xiao · J. Ying
College of Computer Science and Technology, Zhejiang University, Hangzhou, China

M. Wu · B. Peng (✉)
Department of Computer Science and Engineering, Zhejiang University City College, Hangzhou, China
e-mail: pengb@zucc.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_95,

operate HTML Document. But on different terminal devices the way to access local resources is different to each other, and different runtime environment has different development standard. The existing development platform cannot support cross-platform development or cross-standards development. There is too much repeated work in developing a cross-platform deployment widget. A development framework is needed by using which a developer only need to build a platform independence model to develop a cross-platform widget; the frame also should contain a visual programming interface and a model driven development environment.

Model Driven Architecture (MDA) [1] is a software design approach for the development of software systems proposed by Object Management Group (OMG) in 2001. And MDA is based on several of standards proposed by OMG, separating business logic and technology depending on certain platform. The main idea of MDA is building Platform Independent Model (PIM) first, then using model transformation technology to map PIM to Platform Specific Model (PSM) which containing platform information, finally mapping PSM to generate executable code for target platform [2]. The mapping from PIM to PSM and PSM to code is the core technology in MDA.

Using MDA provides a good solution to the problem in widget development: establishing a widget development framework based on MDA. Developers complete business requirement by building PIM of the widget, then model transformation engine provided by the framework map PIM to PSM and code using platform configuration file. The development environment provided by the frame should also support for publishing the developed widget applications in Widget Store by means of widget store service in Cloud Service. Terminal devices have a Widget Engine which is used as an interpretive execution environment. Widgets downloaded from Widget Store can run on the environment.

The rest of this paper is organized as follows: Sect. 95.2 introduces the architecture of MWDF, and then the architecture of development environment implementation technique is presented in Sect. 95.3. Section 95.4 describes the development process of widget by using MWDF. Finally, Sect. 95.5 concludes this paper.

95.2 MWDF Architecture

As Fig. 95.1 shows, MWDF can be divided into three parts. Figure 95.1A presents the functions provided by Development Environment in MWDF, including project management, Model Driven Development module, code editor, emulator and project publish. Figure 95.1B illustrates Cloud Services supported by MWDF. The Cloud Services supplies services for developer and user. In Fig. 95.1C, the architecture of Terminal Device for Widget is shown. Terminal Device mainly contains Widget Engine and Widget Store. Widget Engine is used as a runtime environment for users and provides access capacity to local resources. Widget

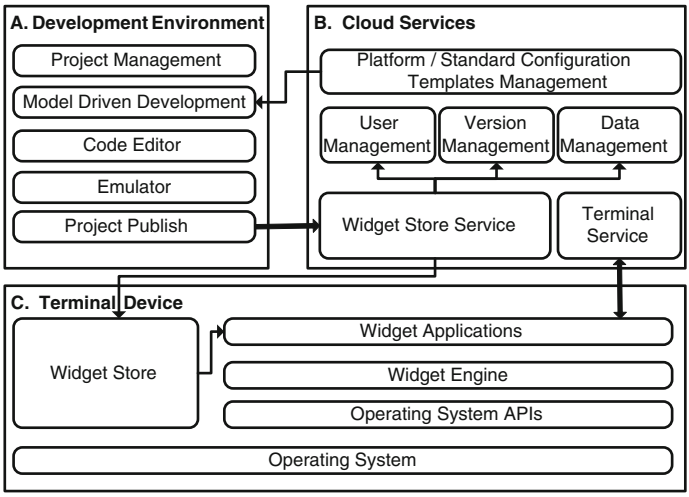


Fig. 95.1 Architecture of MWDF

Store implements its functions by connecting to Widget Store Service in Cloud Service. This section introduces the three modules in detail and communication among them.

95.2.1 Development Environment

The development environment supports for five functions as mentioned above. The developer uses Project Management to create a new widget project including project configuration file, source package, and platform configuration file and so on. After creating a new project, developer could build PIM in the editor embedded in Model Driven Development module, the PIM will be transformed into PSM and the executable code on the basis of platform configuration Template offered by the Cloud Services. The Code Editor helps the developer to modify the generated code by adding functions or business logics that cannot be generated by Model Driven Development module. Emulator is a component on which developer can preview the widget or debug the widget. In the end of the development, Project Publish module publishes the complete widget to Widget Store Service in Cloud Services.

95.2.2 Cloud Services

The Cloud Services are in charge of providing services for widget developers and users. Services offered by Cloud Services include Platform/Standard Configuration

Template Management, User Management, Version Management, Data Management, Widget Store Service and Terminal Service. The Platform/Standard Configuration Template Management is a template library for different platform or standard, which provides template-download service for Model Driven Development module in Development Environment. User Management is responsible for managing developers and terminal users' accounts including login certification, user permissions validation and new user registration. Version Management takes care of the different version of widget project, carries out multi-versioning management. Data Management focuses on collecting data in Cloud Services, and providing them to Widget Store Service with fast access to data and information. Widget Store Service integrates User Management, Version Management and Data Management, provides publish interface for developers and download interface for terminal user. Terminal Service is used as an information transfer station; it pushes information ordered by terminal user and forwards information between terminal users.

95.2.3 Terminal Device

Terminal Device comprises Operating System, OS APIs, Widget Store, Widget Engine and Widget Applications. After being published, the Widget Application can be downloaded from Widget Store via Widget Store Service. Downloaded Widget Application is deployed on Widget Engine, which can interpretively execute the HTML JavaScript in Widget Application. At the same time, Widget Engine enables Widget Application to access local resources by invoking Operating System APIs supported by Operating System. The installed Widget Application also has the ability to interact with Terminal Service in Cloud Services; both of which receiving and sending information is allowed.

Wherever Times is specified, Times Roman or Times New Roman may be used. If neither of them is available on your word processor, please use the font closest in appearance to Times. Avoid using bit-mapped fonts if possible. True-Type 1 or Open Type fonts are preferred. For math etc., please embed symbol fonts as well.

95.3 Implementation of Model Driven Development Module

The Model Driven Development module in Development Environment is the kernel of MWDF. The architecture of Model Driven Development module implementation is shown in Fig. 95.2. The Model Driven Development module is implemented based on the Eclipse Extension Mechanisms, and it integrates

techniques supported by OMG such as Meta Object Facility (MOF), Unified Modeling Language (UML) and Common Warehouse Metamodel (CWM). Model Driven Development module also uses Eclipse Modeling Framework (EMF) and Velocity Template Technique to implement relevant functions. The Model Driven Development can be divided into four parts; all of them are based on Eclipse Platform and Eclipse Extension Mechanisms which is illustrated in Fig. 95.2.

The Infrastructure contains the specific implementation of MDA standard which enables us to access the model and modify it. The capacity is needed in model transformation. The Model Repository Interface handles the several of models accessing and provides the service for storage, management and publish the metadata. By using Façade pattern, Model Repository Interface simplifies operation on Meta model and Meta model interface, which is the basis of model transformation. In general, MDA system Object Constraint Language (OCL) is used to constrain models, but OCL is too complex to satisfy the flexible and simple development requirement in widget development. So here we use JavaScript to describe model behavior. Except that, there are also some other libraries are contained in this layer.

The Interlayer mainly includes some interface encapsulation for Infrastructure such as Model Repository Interface Encapsulation and Model Interface Encapsulation. Template Engine is also a part of this layer; in MWDF we use Velocity template technique to implement Template Engine. In addition, the Model Transformation Language is contained, and Extensible Stylesheet Language Transformations (XSLT) is chosen for this work.

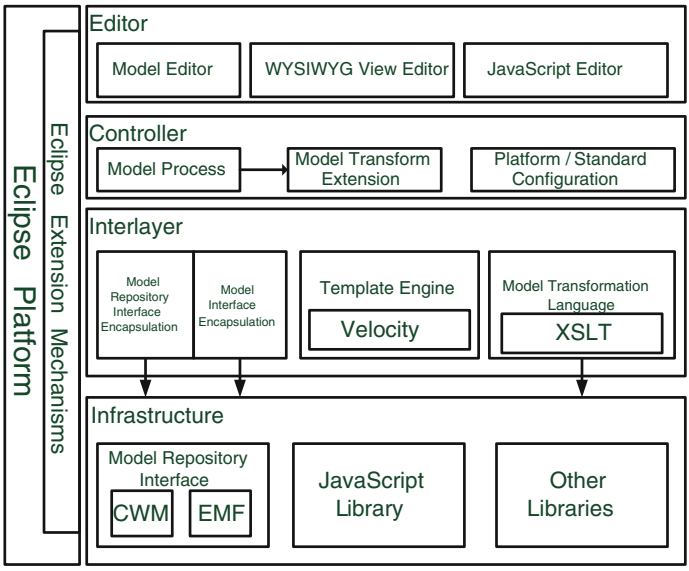


Fig. 95.2 Architecture of model driven development module

Controller is responsible for controlling the development process, such as changing platform configuration, invoking functions in other layer to processing model and so on. The new template downloaded from Cloud Service and created is integrated in this layer. The kernel in this layer is a MDA engine, which is in charge of coordinating different module in system to process the model.

On the top of Fig. 95.2 is the Editor layer. As the user interface for widget developer, Editor Layer plays an important role. There are three kinds of editors: Model Editor, WYSIWYG View Editor and JavaScript Editor. Model Editor is used to build PIM by developer or domain expert, the models are presented in the form of UML, and the Model Editor should provide a data persistence in XML format. WYSIWYG is short for What You See Is What You Get. As its name implies, WYSIWYG View Editor is a visual editor for developing the layout and components in the view. With WYSIWYG View Editor, developers just need to drag and drop the components displayed in graphical way to the correct place to complete the development of view. JavaScript Editor cooperates with WYSIWYG View Editor, using JavaScript to describe the components' behavior.

Different parts in Model Driven Development Module are closely associated. They supports for the model driven development together. In the form of plug-ins of Eclipse, Model Driven Development Module takes full advantage of expansibility supported by Eclipse.

95.4 Development Process Based on MWDF

Development process based on MWDF is a process that builds PIM and then uses model transformation to generate code [3]. Figure 95.3 describes the development process based on MWDF in the view of MDA.

PIM is described in UML, which is established by Model Editor. The Class Diagram is used to present data model in PIM. The State Diagram is used to present business process in PIM. Both of Class Diagram and State Diagram are stored in the form of XML, because of UML model is not suitable for model transformation. In MWDF we use XSLT to transform UML model to XML document [4] which is suitable for model transformation in widget development. The product is called Widget Domain Description File (WDDF). At the meantime, the view description generated by WYSIWYG View Editor and JavaScript Editor is added to WDDF.

WDDF is the PIM description file in MWDF. It is used as the input of PIM to PSM model transformation. According to the chosen Platform Specific Template, PSM is created. And PSM can be mapped into Widget under the restriction of the Code Template. In the development process, both the Platform Specific Template and Code Template are provided by Cloud Services. Developed widget can be published to Cloud Services.

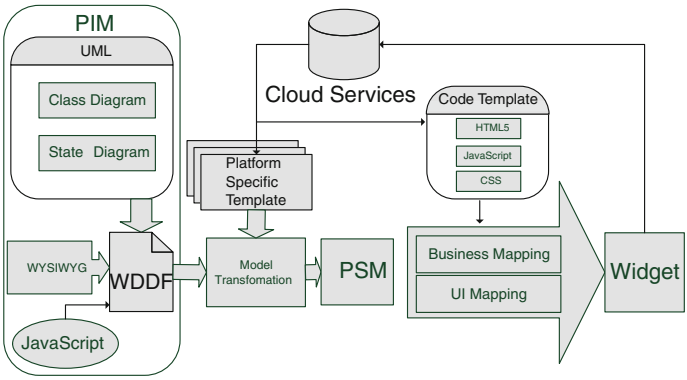


Fig. 95.3 Development process based on MWDF

95.4.1 Building PIM

In MWDF widget development, PIM is used to describe data model, business model and view model, each state in State Diagram is corresponding to a view in widget and each state transaction is corresponding to a view change. Every Widget State Model is formulated as a quad:

$$\text{WSM} = (\text{Id, Transactions, DataModels, View})$$

Id identifies a state, Transactions presents transactions among states, DataModels describes data models in each state and View is stand for the collection of layout and components. Every transaction in Transactions is formulated as a triple:

$$\text{Transaction} = (\text{Transaction TargetId, Arguments, TriggerEvent})$$

TransactionTargetId is the Id of transaction target state. Arguments are arrays of input arguments. TriggerEvent is the event that triggers the transaction.

Every WSM can be illustrated in a WDDF, the structure of a WDDF as Fig. 95.4 shows.

95.4.2 Model Transformation

After building PIM, MWDF gets Platform Specific Template and relevant Code Template from Cloud Services. The Platform Specific Template contains information such as screen type, screen resolution and local resource APIs. Model Transformation Module adds information in Platform Specific Template to WDDF to create PSM. PSM is mapped automatically to executable code which is then packed into a Widget. The template technique is fully used in model

Fig. 95.4 WDDF structure

```
<?xml version="1.0" encoding="utf-8"?>
<WidgetState Id="WidgetStateId">
    <Transactions>
        <Transaction>
            <TransactionTargetId/>
            <Arguments/>
            <TriggerEvent/>
        </Transaction>
        .....
        <Transaction>
            <TransactionTargetId/>
            <Arguments/>
            <TriggerEvent/>
        </Transaction>
    </Transactions>
    <DataModels/>
    <View/>
</WidgetState>
```

transformation, which the core idea that: there is a template to map it into another model or code [5] for each element in source model. The mapping rules listed as following:

1. Each WDDF should be mapped to a page in widget, and the page is identified by the Id of the WDDF.
2. Content in Transaction tag is mapped to a JavaScript function. The function is triggered by TriggerEvent and the values in Arguments tag is used as its parameters. When the TransactionTargetId is not equal to the Id of the current page, the page jump.
3. DataModels are corresponding to data used in the page, and the Meta data in DataModels tag bind it to certain components in View tag.
4. The layout and component tags are in the form of XML standard; it is one–one correspondence with tags in HTML.
5. JavaScript in View tag just need to be copied into script tag in the widget file.

95.5 Conclusion

In this paper the authors propose a widget development framework MWDF based on MDA. It supports model driven widget development for cross-platform or cross-standard. This framework provides a development environment, a Cloud Service for service support and architecture for terminal device to run widget. MWDF abstracts the implement in widget development in which the developer only needs to focus on the specific business logical and user interface. It saves a lot of time for the developer and increases the expandability of the widget.

Acknowledgments This project was supported by the Special Funds for Key Program of the China (No. 2011ZX0302-004-002); the Science Foundation of Zhejiang Province (No. 2010R50009,2011C33015); the Major Projects on Control and Rectification of Water Body Pollution of China (No. 2009ZX07424-001).

References

1. Miller, J., Mukerji, J: MDA guide version 1.0.1 [EB/OL]. http://www.omg.org/mda/specs.htm (2003)
2. Liu, Y., Kang, J., Lu, W: Overview of model-driven architecture. Comput. Sci. (Jisuanji Kexue) **33**(3), 224–228 (2006) (in Chinese)
3. Wu, M., Chen, Z., Jing, Y.: A MDA based approach for multi-platform application development. Inf. Int. Interdiscip. J. **14**, 765–772 (2011)
4. Mattsson, A, Beekveld, M: Simplifying maintenance by using XSLT to unlock UML models in a distributed development environment. Software Maintenance (ICSM), pp. 465–468 (2007)
5. Herrington, J: Code generation in action. Manning Publications co, Greenwich (2003)

Chapter 96
Design of Real-Time Fire Evacuees' State Information Verification System for Fire Rescue

Donghyun Kim and Seoksoo Kim

Abstract The existing fire protection safety management systems are difficult to expect an efficient rescue, because the rescuers cannot confirm the status information of evacuees. They just provide fire and evacuation information. Therefore, this paper suggests a vision based real-time fire evacuees' state information verification system through inputted images from CCTV in the building.

Keywords Human tracking · Fire evacuees' verification · Foreground segmentation · Histogram of oriented gradient

96.1 Introduction

As the number of urban structures is increasing due to domestic industrialization, the number of fire occurrences is also increasing every year. Accordingly, building safety management system is being introduced, which mainly focuses on the performance of fire safety to buildings [1].

However, such a system does not work as designed originally because of the formal inspection of the system inspection and management [2].

In addition, the above system offers the system operator the evacuation information including evacuation routes, fire locations, and fire expansion, but does not provide the system operator the information on the life of person who have not evacuated. Thus, it is difficult to figure out the location of evacuating persons and their status [3–5].

D. Kim (✉) · S. Kim (✉)
Hannam University, Ojeong-dong, Daedeok-gu, Daejeon 306-791, Korea
e-mail: kimdh1986@hun.kr

S. Kim
e-mail: sskim0123@naver.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_96,

Therefore, in this paper, we propose a real-time system for monitoring fire—evacuating persons to rescue their life in case of fire. It is designed to detect a person through images input into CCTV systems in building, generate rescue information, transmit the information to a disaster prevention center.

96.2 Related Works

BuildingEXODUS [6] is the software developed by Fire Safe Engineering Group under Greenwich University. It is designed to simulate evacuating individual behaviors in various spaces, consisting of five sub-models interacting with evacuation subjects, movements, behavior, toxicity, and risk factors. Thus, it enables the user to set details and make a review by implementing a simulation in consideration of factors affecting evacuation.

However, it does not resolve the fire situation itself, so we need a fire simulation. In addition, we need to input the fire results from a fire manually to make analysis on evacuation (Fig. 96.1).

CodeBlue [7] has conducted research on a sensor network used to collect vital signs and locations of survivors in building in case of fire. This study was conducted to see how to obtain evacuating persons' vital signs and location through a sensor attached to their body and the building.

However, the problem with this study is that of the residing persons in building had to all attach a sensor for obtaining vital signs to their body, and carry a mobile terminal with them such as a PDA (Fig. 96.2).

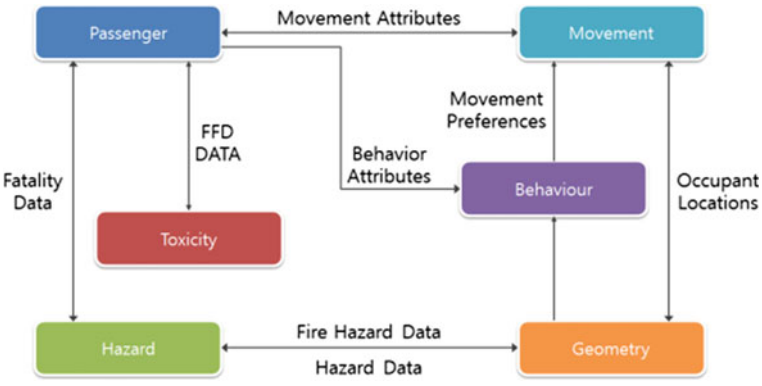


Fig. 96.1 EXODUS sub-model interaction

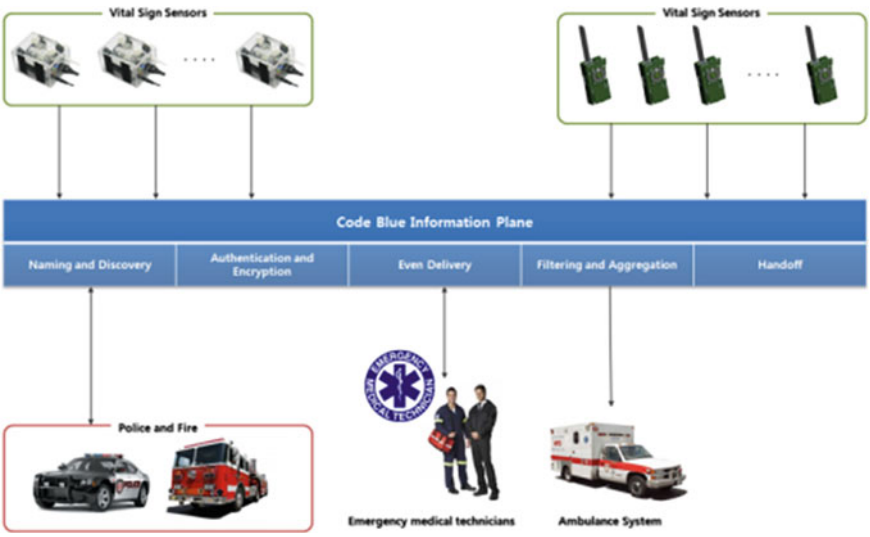


Fig. 96.2 The Code Blue infrastructure

96.3 Real-time Fire Evacuees' State Information Verifications System for Rescue

Figure 96.3 shows a diagram of real-time fire evacuees' state information verification system for fire rescue.

In this paper, we can configured to real time fire evacuees' state information verification system for fire rescue using a total of three modules as shown in Fig. 96.3, and for each module are described follow.

96.3.1 Human Detection Module

96.3.1.1 Dynamic Object Detection

To detect a person, we need to first detect a moving object. In this paper, we detect the moving object using a Gaussian Mixture Model (GMM).

We can obtain Eq. (96.1) in the GMM if a pixel value x measured at random time (t) on successive images has consisted of Gaussian distribution in "M" number; we got Eq. 96.1, accordingly.

$$p\left(\overrightarrow{x}|X_T, BG+FG\right) = \sum_{m=1}^{M} \hat{\pi}_m N\left(\overrightarrow{x}; \overrightarrow{\mu}_m, \overrightarrow{\sigma}^2_m I\right) \tag{96.1}$$

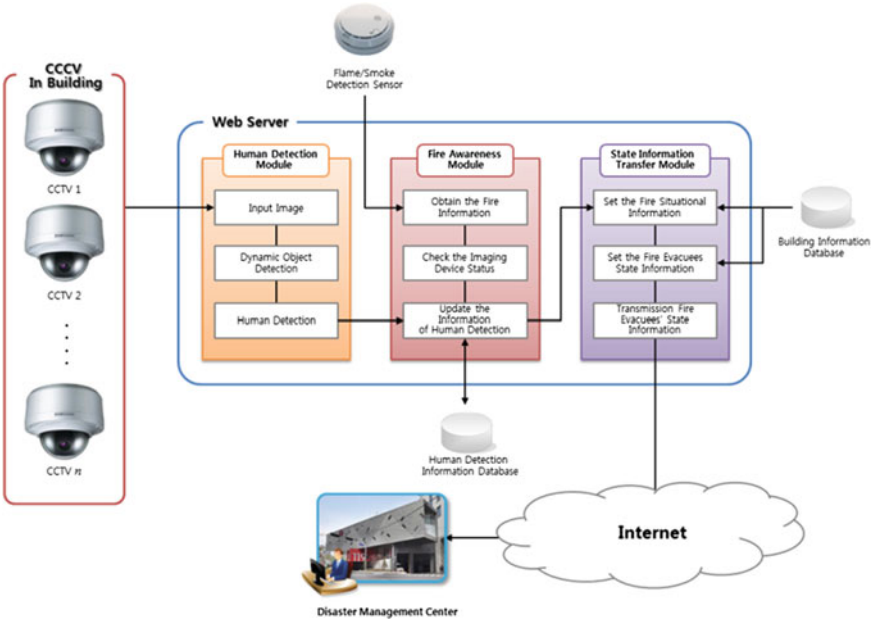


Fig. 96.3 Diagram of real-time fire evacuees' state information verification system for fire rescue

A hobby, $\overrightarrow{\mu}_m$ is the mth average of gaussian distribution and $\hat{\sigma}_m^2$ is mth covariance matrix of gaussian distribution and $\hat{\pi}_m$ is mth weight of Gaussian distribution.

If new pixel value is given at (t + 1) time, it is renewed reflexively by the following equation.

$$\widehat{\pi}_m \leftarrow \hat{\pi}_m + \alpha\left(o_m^{(t)} - \widehat{\pi_m}\right) \tag{96.2}$$

$$\widehat{\pi}_m \leftarrow \hat{\pi}_m + \alpha\left(o_m^{(t)} - \widehat{\pi_m}\right) \tag{96.3}$$

$$\widehat{o_m^2} \leftarrow o_m^2 + o_m^{(t)}(\alpha/\widehat{\pi_m})\left(\overrightarrow{\delta}_m^T \overrightarrow{\delta}_m - \hat{\sigma}_m^2\right) \tag{96.4}$$

In GMM, we sort out the weight of Gaussian distribution in descending order using the difference in the Gaussian distribution on background and foreground. In the sorted weight, we make the approximation against the distribution in "B" number meeting Eq. (96.5). We detect the person candidate region through labeling and set it as Region of Interest (ROI).

$$\mathrm{B} = \mathrm{argmin}_{\mathrm{b}}\left(\sum_{\mathrm{m}=1}^{\mathrm{b}} \hat{\pi}_{\mathrm{m}} > (1 - \mathrm{c_f})\right) \tag{96.5}$$

96.3.1.2 Human Detection

To identify the gradient of the ROI image region, using the following equation, we calculate θ indicating the direction for the dimension (m) of the distribution on pixel changes in axis X and axis Y from the brightness image $\mathrm{I}(x, y)$ of each pixel.

$$\mathrm{m}(\mathrm{x}, \mathrm{y}) = \sqrt{\mathrm{f_x}(\mathrm{x}, \mathrm{y})^2 + \mathrm{f_y}(\mathrm{x}, \mathrm{y})^2} \tag{96.6}$$

$$\theta(\mathrm{x}, \mathrm{y}) = \tan^{-1} \frac{\mathrm{f_y}(\mathrm{x}, \mathrm{y})}{\mathrm{f_x}(\mathrm{x}, \mathrm{y})} \tag{96.7}$$

$$\begin{cases} \mathrm{f_x}(\mathrm{x}, \mathrm{y}) = \mathrm{I}(\mathrm{x}+1, \mathrm{y}) - \mathrm{I}(\mathrm{x}-1, \mathrm{y}) \\ \mathrm{f_y}(\mathrm{x}, \mathrm{y}) = \mathrm{I}(\mathrm{x}, \mathrm{y}+1) - \mathrm{I}(\mathrm{x}, \mathrm{y}-1) \end{cases} \tag{96.8}$$

Using gradient, we designated the 8×8 pixel dimension as one cell. And we drafted out the direction for changes in the brightness in this cell.

We execute the normalization of defining the directional histogram for the brightness created in each cell as one block for the 3×3 cell. The feature amount (9-dimension) of cell (i, j) in Column i, Row j, is expressed using the following equation.

$$\mathrm{F_{i,j}} = [\mathrm{f}_1, \mathrm{f}_2, \ldots, \mathrm{f}_9] \tag{96.9}$$

The feature amount (81 dimensions) of the block in k number is indicated using the following equation.

$$\mathrm{F_{i,j}} = \left[\mathrm{F_{i,j}}, \mathrm{F_{i+1,j}}, \mathrm{F_{i+2,j}}, \mathrm{F_{i,j+1}}, \mathrm{F_{i+1,j+1}}, \mathrm{F_{i+2,j+1}}, \mathrm{F_{i,j+2}}, \mathrm{F_{i+1,j+2}}, \mathrm{F_{i+2,j+2}},\right] \tag{96.10}$$

If we assume the feature vector after normalization as v, we can get the normalization using the next equation.

$$v = \frac{f}{\sqrt{\|B_k\|_2^2 + \epsilon^2}} \quad (\epsilon = 1) \tag{96.11}$$

In this case, the block moving is based on moving one cell to the right and below, respectively. In case the input image is 128×64 in pixel, 6 blocks are generated in the transverse direction and 14 blocks in the longitudinal. Then, the normalization for the total of 84 blocks is complete. The feature vector quantity after normalization by block becomes "84 blocks $\times$ 81 dimensions", so we can get the feature vector of Histogram of Oriented Gradient (HOG) [8] in 6804 dimensions. Using AdaBoost [9], we can detect a person by combining HOG feature vectors.

96.3.2 Fire Awareness Module

Figure 96.4 show the flowchart of fire awareness module.

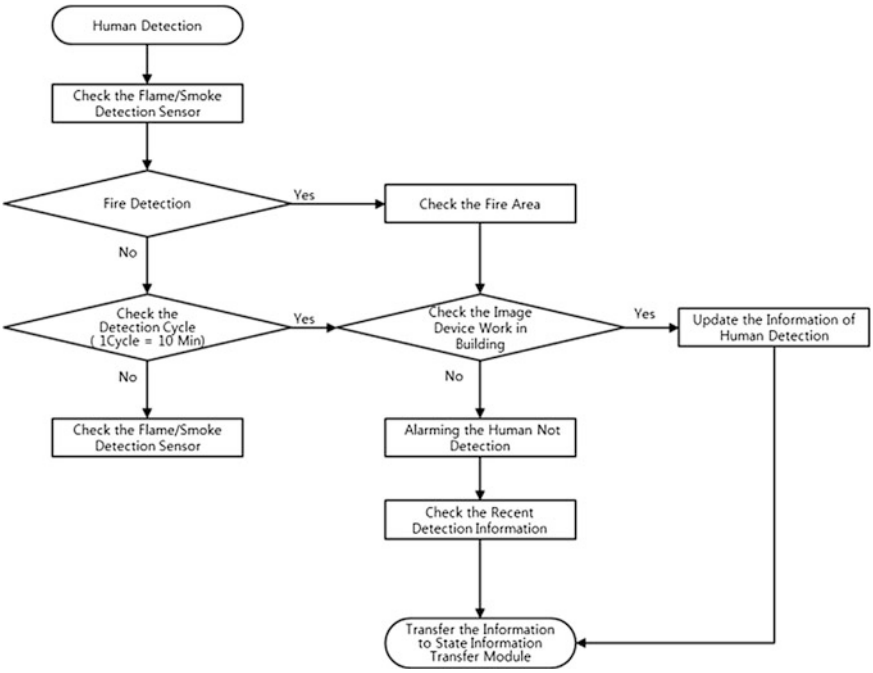


Fig. 96.4 Flowchart of fire awareness module

If the person is detected, using a flame detector, we check whether a fire has occurred using a flame detector and a smoke detector in building. In case a fire, the location of the detector that has detected the fire is detected. Since CCTV can be damaged due to a fire, we must check all the CCTVs in building.

In case we user finds a region where CCTV systems do not operate, we need to check the detected information on the person from the person-detecting information database. For the region where CCTV systems operate normally, the regional information is updated in the person-detecting information database using the detected information. Recently-detected person information or real-time-detected information is transmitted in a status transmission module.

96.3.3 State Information Transfer Module

96.3.3.1 Set the Fire Situational Information

Fire Dynamics Simulator (FDS) [10] is used to perform modeling and interpretation of such phenomenon as heat from a fire, smoke behavior, radiant between vapor and solid phase/convective heat transfer, pyrolysis, flame propagation, fire growth, sprinklers, fire suppression, etc.

In this paper, we calculate fire risks and casualty areas by substituting fire awareness module detected fire area information and building information database's building information for FDS.

96.3.3.2 Set the Fire Evacuees State Information

Then, we estimate the life virtual evacuation information by substituting the detected in-building personal information updated in Fire Awareness Module and building information of Building Information Database for BuildingEXODUS [6].

The calculated life virtual evacuation information, the information on residual persons who have not evacuated, and the data on fire risks and casualty-expected locality calculated through FDS are transmitted to the disaster management center.

96.4 Conclusion

In this paper, we designed a vision-based real-time system for checking fire evacuating personnel states to solve the problem of not figuring out evacuating persons through the existing life state-monitoring system in case of fire.

First, we detected persons using Gaussian mixture model, HOG, and AdaBoost in the input CCTV video. Additionally, we obtained the situational information on evacuating persons through FDS, BuildingEXODUS based on the detected information and building information database and thereby transmitted the information necessary for life-saving to the disaster management center.

In future studies, we will improve the reliability on CCTV's personnel detection by studying how to improve the problem with the inconsistency of the personal detected information generated in Field of View (FOV), which exists in a number of CCTV images of the disaster management center.

Acknowledgments This work (Grants No. C0023833) was supported by Business for Cooperative R&D between Industry, Academy, and Research Institute funded Korea Small and Medium Business Administration in 2012.

References

1. Santos-Reyes, J., Beard, A.N.: A systemic approach to fire safety management. Fire Saf. J. **36**(4), 359–390 (2001)
2. Berry, D., Usmani, A., Torero, J.L., Tate, A., McLaughlin, S., Potter, S., Trew, A., Baxter, R., Bull, M., Atkinson, M.: FireGrid: Integrated emergency response and fire safety engineering for the future built environment.UK e-Science Programme All Hands Meeting (2005)
3. Zeng, Y., Murphy, S.: Building fire emergency detection and response using wireless sensor networks. 9th IT & T conference, 15 (2009)

4. Chen, T.H., Wu, P.H., Chiou, Y.C.: An early fire-detection method based on image processing. International conference on image processing, 2004, vol. 3, pp. 1707–1710 (2004)
5. Xue, H., Ho, J.C., Cheng, Y.M.: Comparison of different combustion models in enclosure fire simulation. Fire Saf. J. **36**(1), 37–56 (2001)
6. Gwynne, S., Galea, E.R., Lawrence, P.J., Filippidis, L.: Modelling occupant interaction with fire conditions using the building XODUS evacuation model. Fire Saf. J. **36**(4), 327–357 (2001)
7. Lorincz, K., Malan, D.J., Fulford-Jones, T.R.F., Nawoj, A., Clavel, A., Shnayder, V,., Mainland, G., Welsh, M., Moulton, S.: Sensor networks for emergency response: challenges and opportunities. IEEE Pervasive Comput. **3**(4), 16–23 (2004)
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. IEEE Computer Society conference on 2005: Computer vision and pattern recognition 2005, vol. 1, pp. 886–893 (2005)
9. Ratsch, G., Onoda, T., Muller, K.R.: Soft margins for AdaBoost. Mach. Learn. **42**(3), 287–320 (2001)
10. McGrattan, K.B., Hostikka, S., Floyd, J., Baum, H.R., Rehm, R.G., Mell, W., McDermott, R.: Fire dynamics simulator (version 5). Technical Reference Guide, vol. 1018, pp. 5–6. NIST special publication (2002)

Chapter 97
Detection and Analysis of Unidirectional Links in Mobile Ad Hoc Network Under Nuclear Power Plants Environment

Kai Ji and Tian-Jian Li

Abstract In order to guarantee the effective communication mode under any circumstances in nuclear power plants, Mobile Ad Hoc network will be introduced to achieve the goal. In this paper, a detecting algorithm is provided to study unidirectional links. With the emulation experiment which results in the occurrence rate of both unidirectional and bidirectional links, analyses of unidirectional links which originates from diversity in transmission power and signal interference is presented, hoping that it will provide assistance in routing protocol of Mobile Ad Hoc Network.

Keywords Mobile Ad Hoc network · Unidirectional link · Nuclear power plants

97.1 Introduction

Mobile Ad Hoc Network (MANET) is a kind of self organizing wireless network [1] which does not have any fixed infrastructure. It is different from common network as independent networking, dynamic topology, unlimited mobility, multi hop [2] and so on. MANET can be used in large quantities of important occasions such as military affairs, environment monitoring and so on.

Effective communications is a chief nuclear safety guarantee in nuclear power plants, telephone system is usually the most significant communication system in nuclear power plants and it needs to provide useful communication mode under all kinds of circumstances and events. Common backup strategies such as multiple power supply and multiple servers can't guarantee that because the backup devices may also be damaged by accidents. In the case, because of its characteristics,

K. Ji (✉) · T.-J. Li
Information Technology Center, China Nuclear Power Technology Research Institute, China
Guangdong Nuclear Power Holding Co. Ltd, Shenzhen, China
e-mail: ji-kai@cgnpc.com.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_97,

Mobile Ad Hoc Network can be imported to nuclear power plants to provide high availability of communications even in the calamity.

In the paper, a summary of MANET is given first; then unidirectional link in MANET under Nuclear Power Plants environment is studied; thirdly, we provide a unidirectional link detecting algorithm; the following part the emulation experiment with unidirectional and bidirectional links is presented; lastly the paper ends with conclusion.

97.2 Summary of MANET

MANET is the combination of mobile communication and computer network. It expands the application of common network. It consists of several mobile nodes which hold wireless transmission/receiving device, acting both as terminal and router, familiar as a kind of wireless no infrastructure network [3].

The routing protocol is the essential factor of MANET, it directly relates with the efficiency of the MANET. The routing protocol can be divided into table driven protocol, on demand protocol and mixed one through routing discovery strategy.

Table driven protocol tries to maintain the routing information by broadcasting request packets periodically, everyone has the whole routing table of the network, so the average End-to-End delay is quite short while the routing overhead and throughput is large. Destination Sequenced Distance-Vector (DSDV) [4], Optimized Link State Routing (OLSR) [5], Improved Destination-Sequenced Distance-Vector (I-DSDV) [6], Efficient Destination-Sequenced Distance-Vector (EFF-DSDV) [7] protocols are all of this kind. On demand protocol only starts the routing discovery just before sending data, though the routing information can be cached for some time, the average End-to-End delay is still high while the routing overhead and throughput is small. Dynamic Source Routing (DSR) [8, 9], Ad Hoc On-demand Distance Vector (AODV) [10], Ad Toc On-demand Distance Vector Link-State-Aware (AODV-LSA) [11] protocols are of this kind. Mixed protocol adopts both the advantages of the former two to achieve acceptable performance evaluation parameters, it asks nodes to maintain routing information which is limit in a specific area proactively to achieve both acceptable average End-to-End delay and routing overhead, Zone Routing Protocol (ZRP) [12], Mobile Agents Routing (MAR) [13] protocols are of this kind.

Because MANET has the characteristics such as independent networking, dynamic topology, unlimited mobility and multi hop, it can be used in many fields which contain military affairs, environment monitoring, emergency communication and even nuclear power plants environment. In the meanwhile, dynamic network topology, limited node energy sources and multi-hop route characteristics indicate that link is the most valuable resource of MANET. It directly influences the efficiency of routing protocol. In order to put forward an efficient routing protocol, one should make the best use of every link. As a result, studies on link

especially on unidirectional link [14–17] which massively exists in actual environment are quite necessary.

97.3 Unidirectional Link Study

97.3.1 Unidirectional Link Modeling

As shown in Fig. 97.1, there are two random nodes S_1 and S_2, R_1 and R_2 stand for their effective transmission radius, d (S_1, S_2) represents Euclidean distance between S_1 and S_2. If $R_2 \geq$ d (S_1, S_2) and $R_1 <$ d (S_1, S_2), the link between S_2 and S_1 is called unidirectional link.

97.3.2 Causes of Unidirectional Link

The causes of the unidirectional link in nuclear power plants come from such aspects.

1. The diversity in wireless transmission capability. In nuclear power plants environment, nodes to be used in MANET are mainly different, such as microwave receiver, base station and laptop, it is called heterogeneous MANET [18]. Different nodes including many mobile nodes play different roles in the network, with different transmission power which gives birth to unidirectional link.
2. Signal interference and barrier. Even all the nodes have the same transmission power and receiving capability, unidirectional links may still exist because of interference and barrier. Nodes may come across different interference, confliction and barrier in nuclear power plant, so the effective transmission radius of them is decreased differently, which will result in unidirectional link.
3. Broadcast. When disaster happens, nuclear power plants will take communication through satellite. Transmission device based on satellite can provide high bandwidth links in huge area for broadcast, while the reverse link is limited because of the high overhead of the uplink transmission.

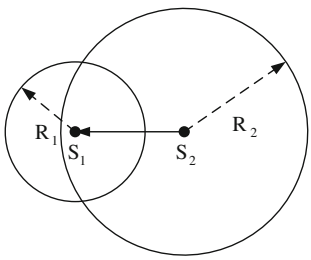


Fig. 97.1 Unidirectional link modeling

4. Transmission limit. Many devices in nuclear power plants are quite sensitive to electromagnetic interference, so only wireless signal receiving is allowed while wireless signal transmission is prohibitive.

97.4 Detecting Algorithm of Unidirectional Link

Because unidirectional link exists massively in nuclear power plants inevitably, and it can influence the routing mechanism in MANET, so we should deal with unidirectional link and bidirectional link on their merits in routing protocol design. The problem stands in the breach is unidirectional link detecting.

Based on the passing study of unidirectional link, a detecting algorithm of unidirectional link is designed, just as Fig. 97.2.

As shown in Fig. 97.2, unidirectional link detecting starts from packet receiving, the detecting procedure can be divided into two stages: reverse power coverage examination and node interference examination, link condition can be judged through the two stages.

97.4.1 Reverse Power Coverage Examination

The stage is directly against the causes of unidirectional links which is the different wireless transmission power.

When node S_i needs to do the routing discovery, it will add its transmission power $P_t(S_i)$ and the smallest power threshold $RX\_Thresh_{S_i}$ it can received in the routing request packet. Then it will broadcast the packet. When the intermediate node S_j receives the packet, it will first measure the received signal power $P_r(S_i)$, and then compare the value with the transmission power $P_t(S_i)$ whose value is in the packet. According to the result, following steps need to be done.

If $P_t(S_j) \geq P_t(S_i)$, it is clear that transmission radius of S_j can cover S_i, the link between them is bidirectional;

If $P_t(S_j) < P_t(S_i)$, goes into Node interference examination directly.

97.4.2 Node Interference Examination

The stage is directly against the causes of unidirectional links which is the signal interference.

1. Computing. We need to compute the channel gain $G_{S_iS_j}$, defined in formula (97.1). Assuming that the channel gain caused by interference in forward and reverse directions is approximately equal, then compute the value $P_r(S_j)$ which

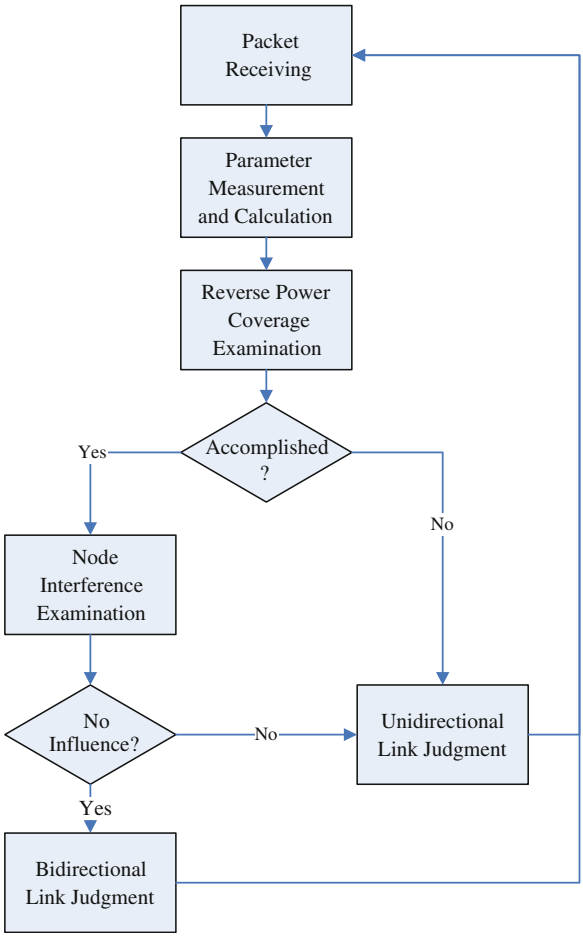


Fig. 97.2 Flow of unidirectional link detecting algorithm

is the received signal power of packet sent by S_j at S_i, it is defined in formula (97.2). Then we will compare the value $P_r(S_j)$ with $RX\_Thresh_{S_i}$ which is the smallest received power threshold of S_i. According to the result, goes into step (b) or (c).

$$G_{S_iS_j} = \frac{P_r(S_i)}{P_t(S_i)} \tag{97.1}$$

$$P_r(S_j) = G_{S_iS_j} \times P_t(S_j) \tag{97.2}$$

2. Comparison. If $P_r(S_j) < RX\_Thresh_{S_i}$, it means the transmission radius of S_j can't cover S_i, the link between them is unidirectional, the detecting procedure

finish. If $P_r(S_j) \geq RX\_Thresh_{S_i}$, it means the transmission radius of S_j can cover S_i, the link between them is bidirectional, the detecting procedure finish.

97.5 Emulation Experiment on Links in Complex Environment

In the section we'll use OPNET software to emulate the actual application environment of MANT. According to the detecting algorithm presented in Sect. 97.4, the experiment directly shows the occurrence condition of both unidirectional and bidirectional links.

1. Link Density (LD): It stands for the percentage of the pair of nodes which have a link between each other in total pair of nodes. Formula (97.3) gives its definition.

$$LD = \frac{n}{N(N-1)} \times 100\,\% \tag{97.3}$$

N stands for the number of all the nodes in MANET, n stands for the number of the pair of nodes which have unidirectional link between each other, what 's more, the pair of nodes (Si, Sj) is not equal to (Sj, Si), they stand for the forward and reverse links between Si and Sj.

2. Bidirectional Link Density (BLD): It stands for the percentage of the pair of nodes which have bidirectional link between each other in total pair of nodes. Formula (97.4) gives its definition.

$$BLD = \frac{n'}{N(N-1)} \times 100\ \% \tag{97.4}$$

N stands for the number of all the nodes in MANET, n′ stands for the number of the pair of nodes which have bidirectional link between each other.

3. Emulation environment: We'll set an emulation environment of $1000 \times 1000\ m^2$ with 40, 50 and 60 nodes in it. Each node's transmission power gets a random value from [0.003w, 0.005w], its smallest received power threshold is −95 dbm. Interference nodes have the transmission power which is 10 % of the regular nodes.
4. Emulation result: As shown in Fig. 97.3, in the actual application environment of MANT in which the wireless transmission capability diversity and signal interference exist, Bidirectional Link Density is always smaller than Link Density. The result shows there are large quantities of unidirectional links, and with the increase of the noise nodes, the number of unidirectional links becomes larger.

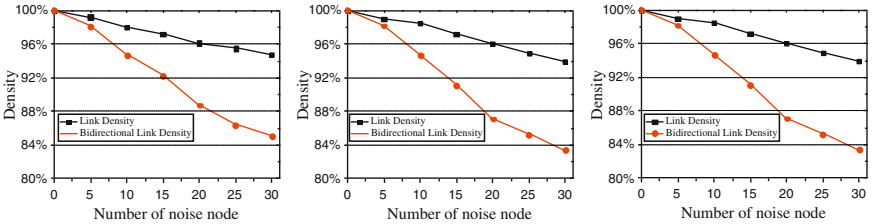


Fig. 97.3 Emulation result (Node Desity = 40/km^2, 50/km^2, 60/km^2)

97.6 Conclusion

The paper studies the deployment of MANET in nuclear power plants, the research focuses on the unidirectional links existing in the actual application environment. A detecting algorithm is presented in the paper and it is taken into practice by emulation experiment in OPNET. The experiment shows that the existence of unidirectional link is quite large which can't be ignored and it becomes larger with the increase of the noise nodes. Research concerning how to improve unidirectional link detection algorithm will be carried out. Then a complete MANET routing protocol will be gradually put forward.

References

1. Sheng, M., Tian, Y., Li, J.D.: A Survey on Wireless Sensor Network and Ad Hoc Network. Xidian University, China (2005)
2. Conti, M., Giordano, S.: Multihop ad hoc networking: the theory. IEEE Commun. Mag. **45**(4), 78–86 (2007)
3. Gopalan, A., Znati, T., Chrysanthis, P.K.: Structuring pervasive services in infrastructureless networks. International conference on pervasive services, pp. 281–290 (2005)
4. Perkins, C.E., Bhagwat, P.: Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers. ACM SIGCOMM Comput. Commun. Rev. **24**(4), 234–244 (1994)
5. Jacquet, P., Muhlethaler, P., Clausen, T.: Optimized link state routing protocol for ad hoc networks. IEEE international multi topic conference, pp. 62–68 (2001)
6. Liu, T., Liu, K.: Improvements on DSDV in mobile ad hoc networks. international conference on wireless communications, networking and mobile computing, pp. 1637–1640 (2007)
7. Khan, K.U.R., Reddy, A.V., Zaman, R.U.: An efficient DSDV routing protocol for wireless mobile ad hoc networks and its performance comparison. Second UKSIM European symposium on computer modeling and simulation, pp. 506–511 (2008)
8. Johnson, D.B., Maltz, D.A., Hu, Y.C.: The dynamic source routing protocol for mobile ad hoc networks (DSR) (2004). http://www.ietf.org/internet-drafts/draft-ietf-manet-dsr-10.txt
9. Patel, R.N.: An analysis on performance evaluation of DSR in various placement environments. Int. J. Comput. Commun. Inf. Syst. **2**(1), 25–30 (2010)
10. Perkins, C.E., Royer, E.B.: Ad hoc on demand distance vector (AODV) routing. IETF MANET Working Group, RFC3561 (2003)

11. Yu, X.H., Yu, O.Y.: A link-state-aware ad hoc on-demand distance vector (AODV) routing protocol for mobile ad hoc networks. International conference on communication technology, pp. 1–4 (2006)
12. Haas, Z.J., Pearlman, M.R., Samar, P.: The Zone Routing Protocol (ZRP) for ad hoc networks. Internet zone routing-protocol-O4.txt (2002)
13. Levy, R., Carlos, P.S., Teittinen, A.: Mobile agents routing-a survivable ad-hoc routing protocol. Military communications conference, pp. 2903–2909 (2005)
14. Ke, X.Z., He, H., Wu, C.L.: A new ant colony-based routing algorithm with unidirectional link in UV mesh communication wireless network. Opto-electronics Letters, pp. 139–142 (2011)
15. Wu, C.X., Wan, C.H., Ye, Y.P.: On demand routing protocol utilizing unidirectional links in mobile ad hoc networks. Comput. Eng. Appl. **33**(5), 1687–1690 (2012) (in chinese)
16. Tang, Y., Li, X., Wang, M.Q.: Improvement of multicast routing supporting mobile ad hoc networks with unidirectional links. International conference on pervasive computing and applications, pp. 502–508 (2011)
17. Shyamala, K., Lokhande, S.K., Kumar, S.: Efficient backup routing scheme in AODV with unidirectional links. Annual IEEE India conference, pp. 1–4 (2011)
18. Zhang, X., Qian, Z.H., Li, T.P.: An efficient routing protocol for heterogeneous wireless ad hoc networks. International conference on multimedia technology, pp. 172–175 (2011)

Chapter 98
Real-time Motion Detection in Dynamic Scenes

Zhihua Li and Zongjian He

Abstract Motion Detection is a fundamental problem in visual surveillance systems, and is especially challenging in dynamic scenes. Temporal difference methods are not robust for dynamic regions, and nonparametric models have preferable results. But nonparametric models have high computational and memory resource requirements. In this paper, a real-time motion detection approach is proposed based on joint space and color model in dynamic regions. Firstly, a simple temporal difference is used in the whole scene. And then nonparametric model is used to reconfirm the motion occurrence in dynamic scene regions with jumping changes. The joint space and color model is built in the variable region of the surveillance scene in order to improve the adaption of dynamic outdoor environment. Experimental results show that the computational complexity and memory requirement of the proposed method are largely reduced, whereas it greatly decreases fault rates in comparison with the temporal difference and achieves the comparable detection accuracy to the original nonparametric model.

Keywords Visual surveillance · Motion detection · Temporal difference · Nonparametric model

Z. Li (✉) · Z. He
College of Information Science and Engineering, Hangzhou Normal University, Hangzhou, China
e-mail: zhihuali_e@163.com

Z. Li
Institute of Advanced Digital Technology and Instrument, Zhejiang University, Hangzhou, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_98,

98.1 Introduction

Motion detection and alarm are important tasks in visual surveillance systems [1]. Automated visual surveillance applications usually use stationary cameras to detect an environment of interest. Temporal difference and background subtraction are two popular and convenient methods for motion detection and alarm. Temporal difference [2–4] includes the subtraction of two consecutive frames followed by parameter thresholding, and is not robust for dynamic events such as ceiling fans, waves on water surfaces, trees swaying in the breeze, and escalators in a dynamical background. Background subtraction [5–12] is based on the subtraction of a background or reference model and the current image followed by a labelling process. It detects moving regions in an image by taking the difference between the current image and the reference background image in a pixel-by-pixel fashion.

Single Gaussian background models [5, 6] are not feasible for the dynamic scenes. Stauffer and Grimson [7–9] have proposed an adaptive on-line parametric model, in which each of the background pixel intensity is modeled as a mixture of Gaussians. The pixel-based mixture model of Gaussians is limited in its capability by the upper bound placed on the number of Gaussian distributions for each pixel. Elgammal et al. [10–12] have proposed nonparametric estimation methods for per-pixel background modeling, in which the background pixel samples are directly used to represent the background color distribution. Background samples are maintained to compute the background distribution using a kernel function (typically a Gaussian function). It can not only detect the motion but also extract the whole motion region, but the model has high computational complexity and large memory requirement. Suppose the scene background sample set uses all the pixel samples from M most recent images as the background samples where M usually ranges from 500 to 1000. This time is much too time-consuming for real-time visual surveillance applications at current CPU speeds, and even more for embedded real-time visual surveillance applications with limited CPU and memory resources.

In this paper a real-time motion detection approach is proposed based on joint space and color model in dynamic scene regions. The proposal is summarized as the following:

1. In the training phase, the surveillance background scene is classified into the stable region and the variable region. The method effectively improves the adaption of the dynamic outdoor environment. And then the joint space and color model is built in the variable region. In the joint model, the pixel space position and nonparametric coefficients of the variable region is computed and recorded for the detection phase.
2. In the detection phase, firstly a simple temporal difference is used in the whole surveillance scene. The simple temporal difference is highly efficient and only needs small memory to store the parameter threshold for each pixel. When the

result pixels that temporal difference detects belong to the variable region, the nonparametric model is used to reconfirm the motion occurrence.
3. In the updating phase, the parameters of the joint space and color model is updated including the pixel space position and nonparametric coefficients of the variable region.

The rest of the paper is organized as follows: the proposed approach is elaborated in Sect. 98.2. Within this section, a description of system framework and the scheme region classification as well as a simple description of the joint space and color model are included. Experimental results are discussed in Sect. 98.3. Finally, some conclusions and possible future work are given in Sect. 98.4.

98.2 The Proposed Motion Detection Approach

In this section the global representation of the proposed motion detection approach is detailed. Suppose an image frame size is $Q \cdot P$, all the pixel samples from M most recent images is used as the background samples in the training phase and H_t represents the image pixel sample value at time t where $0 \leq t \leq M - 1$.

98.2.1 System Framework of the Proposed Approach

The overall system framework diagram of the proposed approach is shown in Fig. 98.1.

The steps of the motion detection system are summarized as follows:

1. Firstly, according to the training video sequence, the scene region is classified into the stable region and the variable region, and a Generalized Clustering Scheme is used to merge the pixels of the variable region to fill the small interspaces in the variable region.
2. The joint space and color model is build in the variable region, in which the pixel space position and nonparametric coefficients of the variable region is computed and recorded, and Pixel difference motion threshold is evaluated by the stable scene region noise learning.
3. Adaptive temporal difference model is used in the whole scene during the detection phase.
4. When the result pixels that temporal difference detects belong to the variable region or its neighborhood, the nonparametric model is used to reconfirm the motion occurrence based on the joint space and color model.
5. Update the joint space and color model parameters including the pixel space position and nonparametric coefficients.

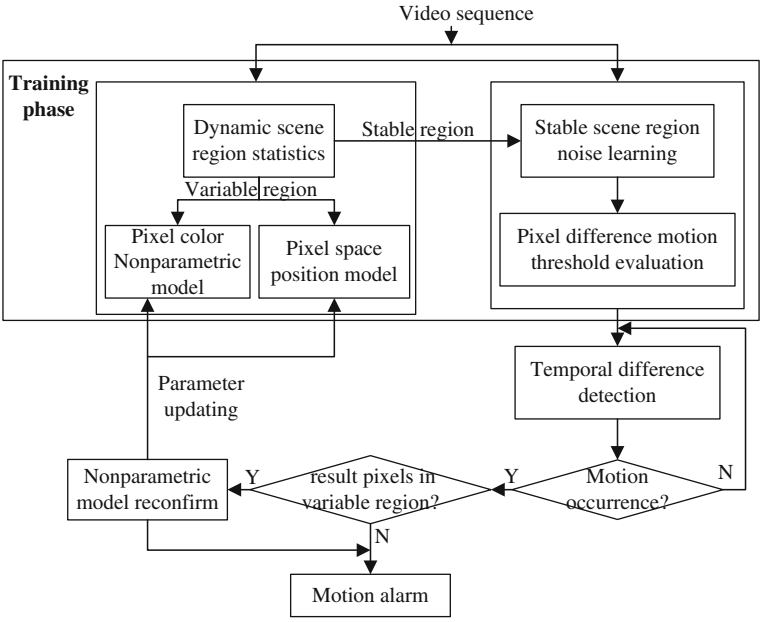


Fig. 98.1 System framework diagram of the proposed approach

98.2.2 Scene Region Classification and Variable Region Joint Model

98.2.2.1 Scene Region Classification

The $|H_{i+1} - H_i|$ of each neighboring pair in the M training background samples is computed in the training phase in order to classify the variable region in the surveillance scene. When any pixels belong to the stable region, all the $|H_{i+1} - H_i|$ should be very small [13]. Figure 98.2b shows the result of classifying a typical surveillance scene (Fig. 98.2a) using (R, G, B) color space. The black area represents the stable region and the white is the variable region. The result shows that the dynamic shadow regions of the fluttering trees and the lightful region in the scene are falsely regarded as the variable region [13]. So the luminance element I and the color element r, g are used to classify the scene region, where $I = R + G + B$, $r = R/(R + G + B)$, $g = G/(R + G + B)$. The criterion of classifying the scene pixels into the stable is as follows:

$$\max_{0 \le i < M-1} (|I_{i+1} - I_i|/I_i) \le \beta, \quad \max_{0 \le i < M-1} |r_{i+1} - r_i| \le \alpha, \quad \max_{0 \le i < M-1} |g_{i+1} - g_i| \le \alpha \tag{98.1}$$

where α, β is the threshold parameter. Figure 98.2c shows the classification result using (r, g, I) color space, where α is set to 0.03 and β is set to 0.2.

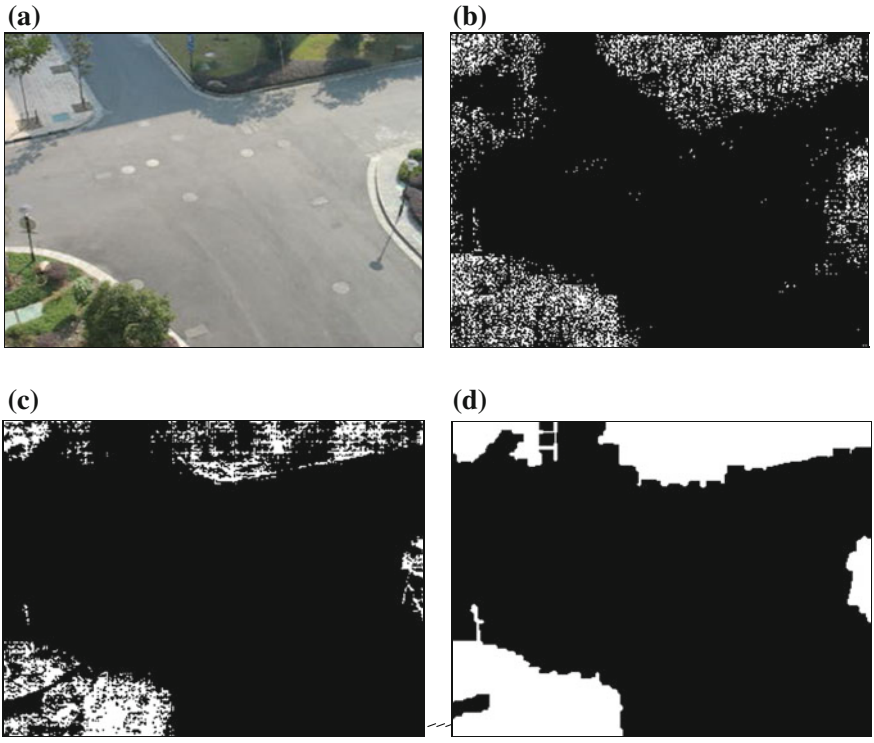


Fig. 98.2 The scene region classification. **a** The original scene frame. **b** Classification result using RGB space. **c** Classification result using rgI space. **d** Combining result of the pixels from **c**

A Generalized Clustering Scheme [14] is used to combine the pixels in the variable region, and then fill the small interspaces in the variable region. The method improves the adaption to the dynamic outdoor environment. The difference measure between two pixel-sets $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ in the algorithm is defined as follows:

$$d_{min}^{ss}(\boldsymbol{v}_i, \boldsymbol{v}_j) = \min_{x \in \boldsymbol{v}_i, y \in \boldsymbol{v}_j} d_2(x, y) \tag{98.2}$$

where $d_2(x, y)$ represents the Euclidean distance between the pixel positions x and y.

The clustering formulation of the two pixel-sets $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ is shown as follows:

$$d_{min}^{ss}(\boldsymbol{v}_i, \boldsymbol{v}_j) = \min_{\boldsymbol{v}_r \in \boldsymbol{R}, \boldsymbol{v}_w \in \boldsymbol{R}} d_{min}^{ss}(\boldsymbol{v}_r, \boldsymbol{v}_w), \quad d_{min}^{ss}(\boldsymbol{v}_i, \boldsymbol{v}_j) \leq d^{T1} \tag{98.3}$$

where $\boldsymbol{R}$ is the set of the clusters and d^{T1} is the threshold parameter. The algorithm is detailed in Fig. 98.3.

In Fig. 98.3, the 8-neighborhood N_8 pixels of a pixel position z are as follows:

$$N_8(z) = \{r | D_8(z, r) = 1\} \tag{98.4}$$

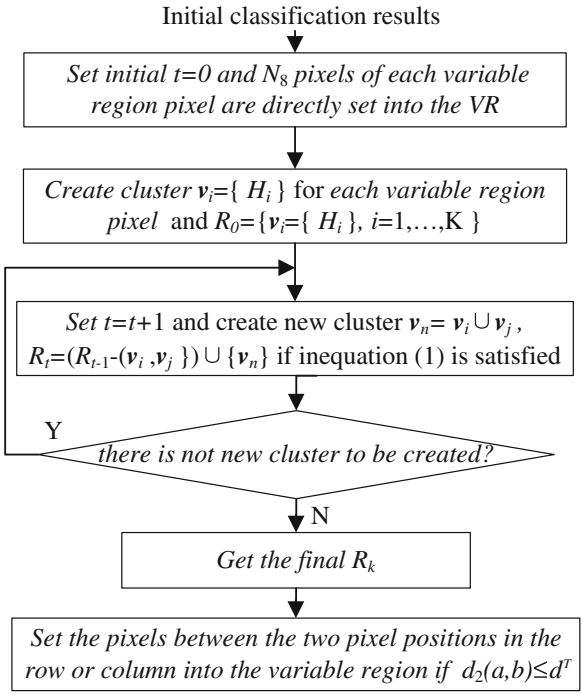


Fig. 98.3 Generalized clustering scheme algorithm flow

where $D_8(S, r)$ represents the l_∞ distance between the pixel positions S and r. And d_2 (a, b) represents the Euclidean distance between the two neighboring pixel locations a and b in a row or column. Figure 98.2d shows combining result of the variable region pixels and filling the interspaces in the variable region from the classification result (Fig. 98.2c), where d^T is set to 10.

98.2.2.2 Joint Space and Color Model in the Variable Region

After the scene region classification, the joint space and color model is built in the variable region in order to improve the adaption of dynamic outdoor environment. The structure of the joint space and color model is shown in Fig. 98.4. In the joint model, the pixel space position and nonparametric coefficients of the variable region is computed and recorded. According to the scene region classification result (Fig. 98.2d), the pixel space position of the variable region (the white area in Fig. 98.2d) is recorded in a 2-D index matrix G[P][Q] based on their locations (i, j) in the image frame where $G[i][j] = \begin{cases} 1, H_{i,j} \in VR \\ 0, H_{i,j} \in SR \end{cases}$, VR is the abbreviation of variable region and SR is the abbreviation of stable region. For each pixel of the variable region, nonparametric model [11] is build and the parameters are recorded for the detection phase.

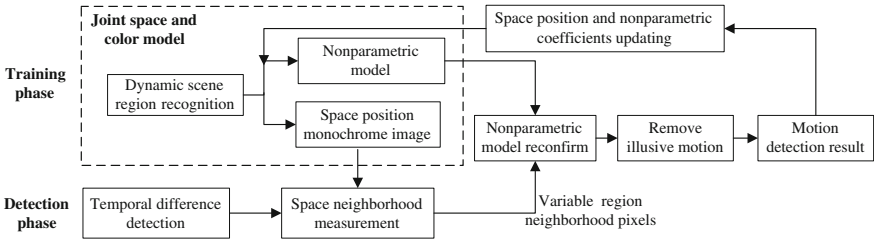


Fig. 98.4 Joint space and color model as well as motion detection flow

98.2.3 Motion Detection and Parameter Updating

In the detection phase, firstly an simple temporal difference is used in the whole surveillance scene. The overall motion detection flow is shown in Fig. 98.4. The criterion of moving pixels is as follows:

$$|H_t - H_{t-1}| \leq T_2 \tag{98.5}$$

where T_2 is the preset threshold parameter. In order to avoid falsely detecting the dynamic shadow regions of the swaying trees, the luminance element I and the color element r, g is used for motion detection where $\mathrm{I} = \mathrm{R} + \mathrm{G} + \mathrm{B}$, $\mathrm{r} = \mathrm{R}/(\mathrm{R} + \mathrm{G} + \mathrm{B})$, $\mathrm{g} = \mathrm{G}/(\mathrm{R} + \mathrm{G} + \mathrm{B})$. The criterion of judging the scene pixels into moving pixels is as follows:

$$|I_t - I_{t-1}|/I_t \leq T_I,\ |r_t - r_{t-1}| \leq T_r,\ |g_t - g_{t-1}| \leq T_g \tag{98.6}$$

where T_I, T_r and T_g are the threshold parameters. The simple temporal difference is highly efficient and only needs small memory to store the parameter threshold for each pixel. When the result pixels that temporal difference detects belong to the variable region or its neighbourhood (24-neighborhood N_{24}), nonparametric model is used to reconfirm the motion occurrence according to the joint space and color model. The 24-neighborhood N_{24} pixels of a pixel z are defined as follows:

$$N_{24}(z) = \{r | D_{24}(z, r) \leq 2\} \tag{98.7}$$

where $D_{24}(z, r)$ denotes the l_∞ distance between the pixel locations z and r.

After the motion detection, the parameters of the joint space and color model are updated including the pixel space position and nonparametric coefficients of the variable region. When nonparametric model reconfirms that the moving pixels is illusive motion because of dynamic environment, the pixel space position 2-D index matrix G[P][Q] of the variable region is updated. At the same time, nonparametric model coefficients are updated according to the literature [11].

98.3 Experimental Results

Two typical video clips are used to compare the proposed motion detection approach with the typical temporal difference and well-known nonparametric model. The first video contains the swaying trees in the breeze and the second video contains the undulating waves on the surface of a lake. The system configuration on which the proposed motion detection approach runs is a Pentium desktop with a 2 GHz CPU and 1 GB memory (RAM). These video frames are 320 × 240 pixels in size.

Figures 98.5b and 98.6b respectively show the classification results of dynamic surveillance scenes (Figs. 98.5a and 98.6a). The simple temporal difference in Figs. 98.5c and 98.6c is very rapidly that averagely takes 15 ms to process a single frame, whereas its detection result in the variable region is so much inaccurate. The threshold T_r, T_g and T_I of the proposed method are respectively set to 0.05, 0.05 and 0.25. The nonparametric model approximately takes 1100 ms to process a single frame, whereas the proposed motion detection method averagely takes 45 ms to process a single frame in video 1 and takes 75 ms to process a single frame in video 2. The proposed motion detection method consumes much less memory than the nonparametric model which is implemented by us for the video 1 and video 2. The current implementation of the proposed method uses a precomputed lookup table in the variable region instead of computing the Gaussian kernel density function directly to improve the run time performance. In addition, the absolute value of the subtraction of each consecutive background sample pair

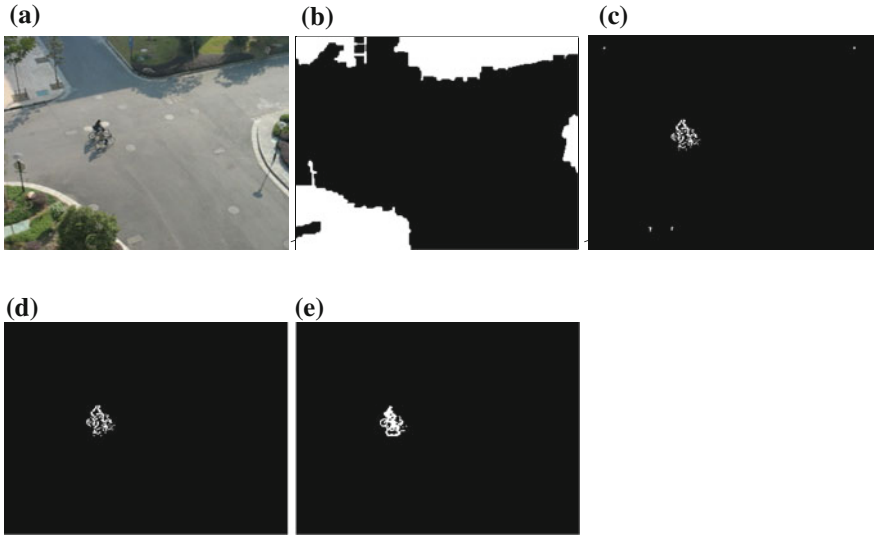


Fig. 98.5 Motion detection in the scene containing the swaying trees. **a** The original scene. **b** Scene classification result. **c** Typical temporal difference. **d** The proposed method. **d** Nonparametric model

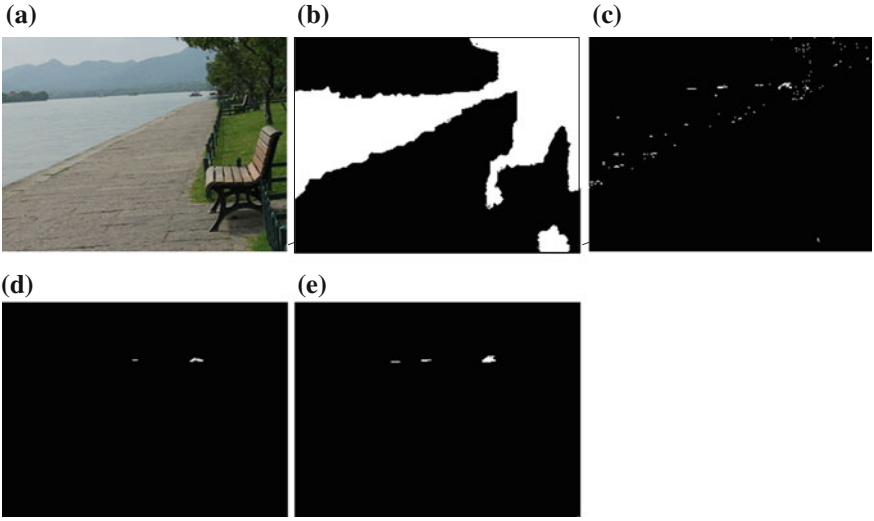


Fig. 98.6 Motion detection in the scene containing the water waves. **a** The original scene. **b** Scene classification result. **c** Typical temporal difference. **d** The proposed method. **d** Nonparametric model

is stored to accelerate the nonparametric model updating in the variable region. The high computational efficiency and low memory requirement of the proposed motion detection method benefits from the scene region classification as well as the simple temporal difference in the stable region.

Figures 98.5 and 98.6 also show the comparisons results of the detection accuracy between the proposed approach and the nonparametric model as well as typical temporal difference. Figure 98.5 shows the comparison result in the dynamic video 1 scene including the trees swaying. Figure 98.6 shows the comparison result in dynamic video 2 scenes containing the water waves. The proposed approach and the nonparametric model have obviously the similar detection accuracy in the variable region of video 1 and video 2. The estimation result shows the simple temporal difference is sufficient to detect the motion in the stable region.

To sum up, the experimental results show that the proposed motion detection approach is more efficient than the nonparametric model, but achieves the similar detection precision.

98.4 Conclusion and Future Work

An efficient motion detection approach is proposed in this paper. Firstly, a simple temporal difference model is used in the whole scene. And then nonparametric model is used to reconfirm the motion occurrence in dynamic scene region with jumping changes. A Generalized clustering Scheme is used to merge the pixels in

the variable region and fill the small interspaces. The joint space and color model is built in the variable region in order to improve the adaption of dynamic environment. The computational complexity and memory requirement of the proposed motion detection method are largely reduced, whereas it greatly decreases fault rates in comparison with the temporal difference and achieves the comparable detection accuracy to the original nonparametric model. The proposed motion detection method is well suited for use in real-time visual surveillance systems such as motion alarm applications. Future work would be to incorporate the moving region extraction algorithms into the scheme.

Acknowledgments This project is supported by the National Natural Science Foundation of China (Grant No. 61001170).

References

1. Hu, W., Tan, T., Wang, L., et al.: A survey on visual surveillance of object motion and behaviors. IEEE Trans. Syst. Man Cybern.-Part C: Appl. Rev. **34**(3), 334–352 (2004)
2. Lipton, A.J., Fujiyoshi, H., Patil, R.S.: Moving target classification and tracking from real-time video. In: Proceedings of IEEE workshop applications of computer vision, 1998, pp. 8–14 (1998)
3. Toyama, K., Krumm, J., Brumitt, B., Meyers, B.: Wallflower. Principles and practice of background maintenance. In: Proceedings of IEEE international conference on computer vision, 1999, pp. 255–261 (1999)
4. Stringa, E., Regazzoni, C.S.: Real-time video-shot detection for scene surveillance application. IEEE Trans. Image Process. **9**(1), 69–79 (2000)
5. Wren, C.R., Azarbayejani, A., Darrell, T., Pentland, A.P.: Pfinder: Real-time tracking of the human body. IEEE Trans. Pattern Anal. Mach. Intell. **19**, 780–785 (1997)
6. Mc, S.J., Jabri, S., Duric, Z., Rosenfeld, A., Wechsler, H.: Tracking groups of people. Comput. Vis. Image Underst. **80**, 42–56 (2000)
7. Bhandarkar, S.M., Luo, X.: Fast and robust background updating for real-time traffic surveillance and monitoring. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Washington, DC, USA (2005)
8. KaewTraKulPong, P., Bowden, R.: An improved adaptive background mixture model for real-time tracking with shadow detection. In: Proceedings of the European workshop on advanced video-based surveillance systems, Kingston, UK (2001)
9. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: Proceedings of IEEE conference on computer vision and pattern recognition, Fort Collins, CO, USA, pp. 246–252 (1999)
10. Elgammal, A., Duraiswami, R., Harwood, D., Davis, L.: Background and foreground modeling using nonparametric kernel density for visual surveillance. Proc. IEEE **90**(7), 1151–1163 (2002)
11. Elgammal, A., Harwood, D., Davis, L.: Non-parametric model for background. In: Proceedings of the 6th European conference on computer vision-Part II, London, UK (2000)
12. Luo, X., Bhandarkar, S.M., Hua, W., Gu, H.: Nonparametric background modeling using the condensation algorithm. In: Proceedings of IEEE conference on video and signal based surveillance, Washington, DC, USA (2006)
13. Li, Z., Zhou, F., Tian, X., Chen, Y.: High efficient moving object extraction and classification in traffic video surveillance. Syst. Eng. Electron. **20**(4), 858–868 (2009)
14. Theodoridis, S., Koutroumbas, K.: Pattern Recognition. Academic, New York (2006)

Chapter 99
Using Kohonen Cluster to Identify Time-of-Day Break Points of Intersection

Yang Jun and Yang Yang

Abstract The historical traffic data is in fact capable of proving abundant of information that can aid in the development of improved current traffic control. Time-of-day (TOD) control system is the most widely used one in the world with the limited funding and system maturity, so it is important to use data mining tools that demonstrate the value of traffic data to enhance the performance of TOD systems. Kohonen cluster approach is very useful for determining TOD break points for better traffic signal control within unsupervised neural network. A case study using an intersection corridor was conducted to identify TOD break points to support the design of signal timing plan by using kohonen cluster. The results of this research indicate that the proposed method can identify TOD break points successfully without deploying multiple signal timing plans on the basis of the subjective judgment.

Keywords Traffic control · Kohonen cluster · Time-of-day · Signal timing plan

99.1 Introduction

Intelligent transportation systems (ITS) have introduced sophisticated real-time sensing and decision support system to surface transportation. It is important to utilize traffic data and to optimize traffic signal control. The data is used to support traffic operations and provides the on-line information for the traffic controllers. A traffic signal system is one of the most effective tools to alleviate traffic congestion, and get improvement in measures of effectiveness such as fuel consumption and delay time. A lot of signal systems currently operate using the

Y. Jun (✉) · Y. Yang
Faculty of Transportation Engineering, Kunming University of Science and Technology, Kunming, China
e-mail: yangjun_km@sina.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_99,

time-of-day (TOD) approach, where a pre-set plan is automatically used for a particular time interval. Therefore, designing a TOD system mainly requires traffic engineers to develop signal-timing plans that are effective for particular time intervals in a day. The current approach used to define the state for TOD plan development is based on single day, hand counted volumes. It may be inadequate. And it is difficult for inexperienced traffic engineers to select different timing plan and an interval break point through identifying significant changes in traffic volume.

In this paper, the kohonen cluster analysis helps classify historical traffic data into different groups or clusters. So the optimal break points for TOD traffic signal operations can be gotten by using the characteristics in different clusters.

99.2 Literature Review

A few methods are recently presented for TOD breakpoints by investigating archived traffic data. One method is based on a genetic algorithm that optimizes TOD breakpoints with explicit consideration of signal timing performance at a representative intersection [1]. Wang [2] uses k-means method to traffic volume data to identify time-of-day (TOD) breakpoints with the assumption that the analyst specifies the number of clusters. Smith et al. [3] applied data mining tool to support the maintenance of traffic signal systems when traffic conditions have changed significantly. Park and Lee [4] present a feature vector of optimal cycle length per time interval instead of traffic volume itself to determine optimal break points. Abbas [5] presents a multi-objective evolutionary algorithm to find optimal selection of timing plans and the optimal TOD schedule.

However, in the above literature, the value of k must be given in k-means cluster analysis to identify time-of-day break points, which could lead to subjective errors. But kohonen cluster analysis can completely avoid within unsupervised neural network. A lot of studies couldn't avoid the noisy allocations of traffic volumes to clusters and lead to the fault TOD breakpoint. In the same, because of few data which is important in cluster analysis, some statistical approaches produce unreliable clusters.

99.3 Methodology

99.3.1 Kohonen Neural Network

Kohonen network is an unsupervised neural network in which its input data with similar features are mapped to form clusters by competitive learning algorithm [6]. The algorithm considers the Euclidean distance between two n-dimensional

vectors which is measured by the similarity between input vectors. The distance of an input vector from each neuron i, D_i is given by,

$$D_i = \left\| W_{ij} - X \right\| = \sqrt{\sum_{j=1}^{n} (x_j - w_{ij})^2} \tag{99.1}$$

where $X = (X_1, X_2, \ldots, X_n)^T$ denotes an input vector;

$W_{ij} = (W_{11}, W_{12}, \ldots, W_{im})^T$ denotes the weight vector of the ith neuron.

The kohonen with a minimum distance is called as the winner. In other words, the weight vector is closest to the input vector.

$$D_w = \min\{D_j\} i \in \{1, 2, \ldots, m\} \tag{99.2}$$

In the kohonen algorithm, the concept of the winner-take-all units are related to the biological concept of grandmother cells because they are responsible for selecting one specific feature, for example, the feature presenting the stereotypical grandmother. However, this representation is not robust because when one unit is removed; all information concerning the corresponding class would be lost. For robust competitive learning, kohonen proposed the self organizing training algorithm. Ideally neighboring neurons classify neighboring features and thus the loss in one neuron will result in a decrease of accuracy but not in a complete loss of information. For each input vector only one such unit will respond, namely the unit characterized by the maximum output, respectively minimum distance, for this input vector x. During the training, the winner adjusts its weights to be closer to the values of data and the neighbours of the winner also adjust their weights to be closer to the same input data vector according to the following relation

$$\begin{aligned} W_i &= W_{ij} + a(W_i - X) \\ i &= \{1, 2, \ldots, m\} \end{aligned} \tag{99.3}$$

The units of the network are thus competing for selection. Only the weights of the winner will be adapted. The adjustment of the neighbouring neuron is instrumental in preserving the order of the input data. Thus, the winning neuron is the closest to the input value. After training, the weight vectors are self organizing and represent prototypes of classes of input vector. The complete algorithm is described as follows:

i. t: = 1; initialize w_{ij} randomly for i = 1,…, m; j = 1,…, n,
   Initialize weight of bias $b_i = e^{[1-\ln(1/k)]}$, where b_i is bias weight neuron-i and K is amount of class. Set learning rate and maximum epoch.
ii. Choose input vector $x \in X$ randomly in the training set.
iii. Determine the neuron i such that its weight vector w is closest to the input vector

$$D_w = \min\{D_j\} i \in \{1, 2, \ldots, m\} \text{ for all i.}$$

iv. Update the weight vector w_i, i = 1,…, m;

$$W_i = W_{ij} + a(W_i - X)$$
$$i = \{1, 2, \ldots, m\}.$$

v. Update bias weight:
$c(i) = (1 - a)e^{(1-\ln(b(i)))} + ay(i)$ for the winner neuron
$b_i = e^{[1-\ln(1/k)]}$ for else.

vi. Increment the time t: = t + 1.

vii. Go to step (ii) until maximum epoch is reached.

99.3.2 Traffic Data and Data Quality

From the inspections of real data sets, many missing data were hidden between the normal data with various values. There were many possible reasons for these mistakes, and there was no way to avoid them. Even though many bad records were deleted from the raw data set after data screening tests, some missing observations were still in the data set. To improve the accuracy of the raw data set, many "bad or missing" data are returned from the detectors. Data quality must be taken into account. The cluster methodologies will be disturbed with these noises.

In this paper, the anomaly function of SPSS Clementine is used to inspect the data set by different time steps and weekdays. The SPSS Clementine anomaly procedure calculates the mean, median, 75th percentile and 25th percentile of the variables, and then uses these values to get the Inter Quartile Range (IQR), which is the distance between the 75th percentile and 25th percentile. The lower far fence is located at 2*IQR below the 25th percentile; the upper far fence is located at 2*IQR above the 75th percentile. These observations belong to outliers if they are located beyond the lower or upper fence. The outliers will be replaced by the mean.

99.3.3 State Definition

In an urban traffic network, the traffic conditions can be represented with the state vectors that contain variables necessary to indicate the performance of the urban network. A state is an abstract representation of the condition of that system at some point in time. The defined state serves as a sufficient statistic for the condition of the system, i.e., it contains all possible information regarding current status, propensity to change and response to external control, as well as the information necessary to evaluate the defined indices of performance for the system [7]. The concept of state-based control is to use a set of established rules or

policies to guide the selection of a control strategy for a system as the system transitions from one state to another.

Data is collected at signalized intersections by single inductive loop detectors in usual. Volume, occupancy, and speed are the three types of traffic data which system detectors collected. Because speed was not a directly measure and occupancy was often bad quality, only traffic volume was selected to identify TOD break points in this research. A three-direction intersection was studied in this paper, so the state is as follows, with each variable number assigned according to its direction number:

$X(t) = (NV, EV, SV)$
where $X(t)$ = system state at time t
NV = north import volume at time t
EV = east import volume at time t
SV = south import volume at time t.

99.4 Research Case Study

The data set in this study was extracted from the database in Kunming Traffic Police Detachment signal system. This system uses inductive loop detectors positioned well upstream of intersections (system detectors) to collect basic demand data. The proposed methodology is implemented at a three intersection of Dianchi Road and Guangfu Road, which is the most important intersection in Dianchi Road. Figure 99.1 shows the intersection used in the study.

The data set used for this case study consists of 24-h observations from Monday to Friday, because weekends have different traffic distributed pattern in evidence. And this data covered 5 days; it was enough to illustrate the method in this paper. Table 99.1 shows arrival traffic volumes data set in the 24 h, and the numbers from

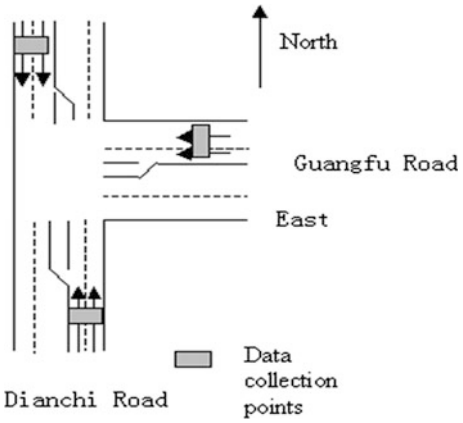


Fig. 99.1 Intersection of Dianchi road and Guangfu road in Kunming

Table 99.1 Arrival traffic volumes data set

Time from	1EV	1SV	1NV	2EV	2SV	2NV	3EV	3SV	3NV	4EV	4SV	4NV	5EV	5SV	5NV
23:00	96	187	301	126	279	355	117	258	353	132	303	410	117	223	344
0:00	78	118	148	74	129	193	89	142	238	69	130	240	80	126	221
1:00	40	73	101	46	86	116	49	88	142	39	84	131	39	94	147
2:00	12	39	67	14	40	77	39	50	66	18	50	78	15	50	71
3:00	6	28	42	11	20	40	13	46	48	20	22	40	18	20	44
4:00	21	23	35	8	29	***98***	13	36	53	12	35	47	19	35	73
5:00	79	234	188	102	229	202	103	214	219	82	220	165	92	210	194
6:00	513	1163	802	449	1085	839	509	1079	947	456	1052	843	432	1075	878
7:00	506	982	1534	390	781	1245	508	939	1614	543	1029	1736	510	1006	1586
8:00	536	1101	1564	642	1138	1643	475	1051	1376	540	1014	1363	621	1096	1536
9:00	524	1095	1247	567	1044	1368	512	925	1347	526	961	1125	627	1095	1332
10:00	527	1011	1186	536	1004	1245	498	892	1153	487	903	1046	538	1075	1242
11:00	438	763	948	478	777	1018	469	773	960	464	793	1071	513	882	1060
12:00	557	906	1028	523	1001	1064	609	1033	1188	458	1019	1037	654	1177	1176
13:00	649	967	1194	656	1067	1256	641	1071	1154	597	1118	1176	631	1062	1239
14:00	690	1028	1156	568	1024	1001	631	1123	1175	548	1167	1218	532	1065	1239
15:00	711	1160	1102	564	1108	930	641	1253	1124	556	1153	1157	542	1182	1242
16:00	709	1363	1398	718	1349	1209	658	***1897***	1281	738	1329	1320	776	1377	1329
17:00	668	947	1352	636	846	1546	661	887	1364	622	914	1386	730	975	1525
18:00	414	716	1092	408	743	1128	445	836	1233	420	764	1250	492	835	1344
19:00	446	1020	1099	448	1269	1024	387	981	1067	431	1032	1074	518	1177	1278
20:00	291	754	1018	390	930	1050	337	849	954	381	884	1041	475	1160	1157
21:00	218	654	729	248	756	859	243	616	748	237	669	776	304	793	992
22:00	150	380	378	147	455	488	121	367	489	136	386	466	199	568	577

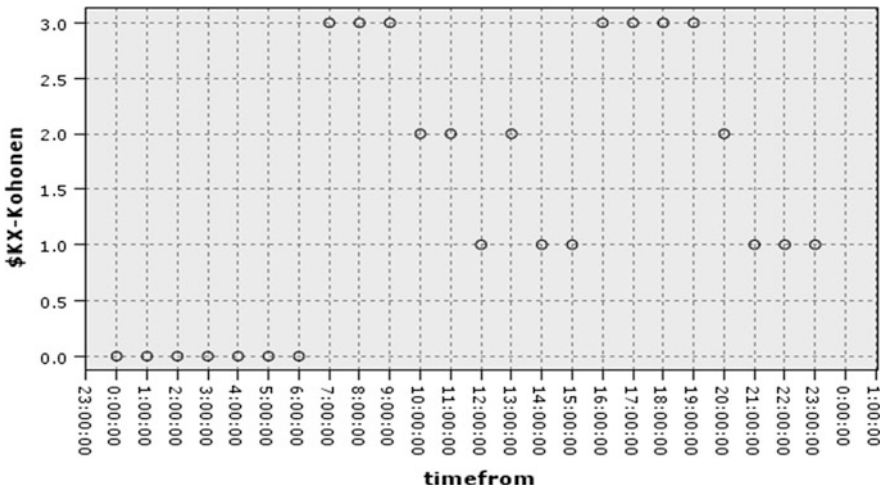


Fig. 99.2 Break points and traffic pattern in 1-h

1 to 5 in the Table 99.1 were used to represent the days of the week: Monday, Tuesday, Wednesday, Thursday and Friday. So, the column of 1EV expresses hourly traffic volume of the east import occurring during the 24-h period of a day on Monday, and the other means the same.

Some outliers had been found through detection by the SPSS. The outliers were shown in the Table 99.1 using Italic, wavy line and bold, and these outliers had been replaced by the mean. Figure 99.2 was obtained that shows break points and traffic pattern in 1-h interval. A number of observations may be gotten when considering Fig. 99.2.

(1) Cluster1, off-peak traffic pattern, 0:00–6:00.
(2) Cluster2, Mid-Day traffic pattern, 10:00–15:00, since changing timing plan frequently may worse traffic conditions, so 10:00–11:00 belonged to Mid-Day traffic pattern.
(3) Cluster3, PM-Day traffic pattern, 20:00–23:00, 20:00 belonged to PM-Day traffic pattern.
(4) Cluster4, PM-peak traffic pattern, 16:00–19:00, in the same.
(5) Cluster5, AM-peak traffic pattern, 6:00–9:00.

99.5 Conclusion

The present common signal timing control strategy for dynamic traffic is time-of-day modes. The control strategies of timing plans include one timing plan for 1 day or different time periods based on predetermined rates or designed traffic demands. Because of unsupervised neural network, it is reliable for kohonen cluster to improve signal-timing plans avoiding subjective errors. This paper

describes the method of using the kohonen cluster methodology to identify time break points with enough traffic volumes. The methodology is shown to be effective.

References

1. Park, B., Santra, P., Yun, I., Lee, D.-H.: Optimization of time-of-day breakpoints for better traffic signal control. Transp. Res. Rec. **1867**, 217–223 (2004)
2. Wang, X., Cottrell W., Mu, S.: Using k-means clustering to identify time-of-day break points for traffic signal timing plans. Intell. Transp. Syst. Proc. IEEE **13**(15), 586–591 (2005)
3. Smith, B.L., Scherer, W.T., Hauser, T.A.: Data mining tools for the support of traffic signal timing plan development. Transp. Res. Rec. **1768**, 141–147 (2001)
4. Park, B., Lee, J.: A procedure for determining time-of-day break points for coordinated actuated traffic signal system. KSCE J. Civil Eng.**12**(1), 37–44 (2008)
5. Abbas, M.: Optimization of time of day plan Scheduling using a multi- objective evolutionary algorithm. Joint international conference on computing and decision making in civil and building engineering, pp. 2427–2436 (2006)
6. Kohonen, T.: The self-organizing map. Proc. IEEE **78**(9), 1464–1480 (1990)
7. Puterman, M.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley& Sons, New York (1994)

Chapter 100
Botnet Emulation: Challenges and Techniques

Bo Lin, Qinfen Hao, Limin Xiao, Li Ruan, Zhenzhong Zhang and Xianchu Cheng

Abstract Botnet Emulation is an emerging method to research on Botnet which is attracting widespread attention. It is referred to build a closed Botnet with virtualization technology to do analysis on Botnet. Although superior to other traditional methods for its flexibility, reproducibility, validity as well as lawfulness, Botnet Emulation is facing challenges from security, transparency, scale and so on. In this paper, we shed light on some of the key challenges in building Botnet Emulation systems. Furthermore, we discuss various techniques used to address or alleviate these problems, along with the pros and cons of each technique. We hope to motivate future research in this area to develop practical solutions to these challenges.

Keywords Botnet emulation · Virtual machine monitro · Virtualization Management platform

100.1 Introduction

Botnet has become one of the greatest threats to computer security around the world. What makes Botnets such a grave threat is its use of command of control (C&C) channels to connect bots to their botmasters. The botmaster can disseminate unified commands to their bot armies, thus it would lead to greater damage than other isolated malware. From some angle, Botnets are complex distributed systems with hundreds to thousands of scale, and in some cases, millions of computers. In order to have a good understanding of Botnets to avoid or mitigate the threat of Botnet, it is necessary to research on Botnet as a whole. Traditional Botnet research methods [1] include Bot Anatomy, Wide-area Measurement Study

B. Lin (✉) · Q. Hao · L. Xiao · L. Ruan · Z. Zhang · X. Cheng
School of Computer Science and Engineering, Beihang University, Beijing, China
e-mail: lin_victor@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_100,

and Botnet Modeling and Simulation. Bot Anatomy is always used to analyze the behavior of a new Botnet on the level of binary execution Bot program. We can get the initial information of Botnet including its C&C architecture and command format, but it cannot give the whole operating mechanism of Botnet. Wide-area Measurement Study is consist of malware collection model, malware analysis model, Botnet tracing model and data analysis model, which is a complex and fragile system. Little changes of the Botnet would lead to the system ineffective. Especially, with the innovation of Botnet, the decrease of Botnet size, the higher concealment and antagonism, Botnet is bringing greater challenges to the traditional research methods illustrated in the following:

1. In order to study the real Botnet, it is usually necessary to create a number of instances into the Botnet. However, too many instances may cause the Botnet operator's notice and indirectly help the operator to optimize his defense strategy. Meanwhile, if the number of instances is too small, we cannot get precise test results leading to our research inconvincible.
2. Wide-area Measurement Study always needs building complex detecting, tracing and protection system. In particular, we must carefully design security management module. But due to Botnet's potential nature threat, it is impossible to guarantee absolute safety. For example, some researchers reported that their Botnet Measurement Platform was attacked by Botnet DDOS attack. Besides, Botnet Measurement experiment usually takes 6 months or longer to collect a relatively reliable and comprehensive information, which is a long life cycle compared to the amazing evolution speed of Botnet.
3. Sometimes, we need to study the effect of the defense strategy on one Botnet to choose the best solution. But the realistic experimental parameters lack certainty and reliability due to the complex Internet Network environment and operator's intervention by hand.
4. Mathematical modeling is often complex and difficult to understand, and it cannot reflect the network behavior which is an important Botnet characteristics.

Finally, it comes to the moral law to add Botnet instances, which in some parts of Europe is a banned behavior, even if it is only used for research.

Due to the above defects of traditional Botnet research methods, researchers are beginning to study the possibility to build a closed Botnet for research, which would make up for all above disadvantages of traditional Botnet research methods.

100.2 Botnet Emulation Research and Related Work

Botnet Emulation research is to build a virtual emulation platform, similar with real network environment, to have a research on the Botnet. Compared to the traditional methods, firstly, with building a closed Botnet environment, Botnet Emulation research is able to provide relatively higher secure and accurate

research environment, which ensure the Botnet cannot break outside of the experiment network. Secondly, unlike other traditional methods, which they are acting as the bot armies in the wild Botnet, Botnet Emulation builds the whole picture of Botnet including bot armies, network environment as well as Bot Admin console, which has a better control of experimental parameters and a shorter cycle to collect precise information. Finally, with an isolation between experimental environment with real Internet environment, researchers are free to do any experiments on their own Botnet regardless of legal issues.

Botnet research based on virtualization technology is actually not new. In order to trap bot executive program, researchers [2–4] use Honeypot technology, which employ virtual machines, to imitate the activities of the real systems and intentionally expose system flaws to Botnet. We can restore a Honeypot quickly after the Honeypot is compromised by malware. Furthermore, many research organizations build Honeynet [5–7] to carry out large-scale Botnet detecting and tracking measurement experiments. And a Honeynet Project is raised to investigate the latest attacks, develop open source security tools to improve Internet security and learn how malicious hackers behave.

As early as the beginning of Botnet research, it was proposed to build a closed environment to study Botnet. Based on the design thinking, some researchers built software-based Botnet simulation environment to study Botnet. Its fatal drawback was poor emulation of the real Internet environment leading to distortion and unreliability.

Recently, researchers are using network testbeds such as PlanetLab, Emulab, and DETER to build Botnet Emulation Platform. For example, in 2007, Paul Barford [8] designed a Botnet Emulation Environment (BEE) on the Emulab [9], aiming to provide a safe and enclosed test platform for Botnet. It startup 5 guest operating systems with VMware Server 1.0.1 software on 100 physical nodes to build the 500-scale Botnet emulation platform. Based on Emulab, BEE made good use of its flexible configuration features to deploy variety of network topology. Jackson [10] use DETER to deploy their System for Live Investigation of Next Generation bots (SLINGbot) which enables researchers to construct benign bots for the purposes of generating and characterizing botnet command and control traffic. However these experimental platforms are not that suitable for high risk security experiments, such as Botnets emulation, owing to the risk of malware "breaking" through logical barriers and escaping into the wild.

In 2009, Sandia National Laboratory [11, 12] successfully achieved one million scale of Botnet with the Thunderbird Super Clusters (4460 nodes) and the Hyperion (1024 nodes). Scientists from Canada and France installed 3000 copies of Windows XP on 98-blade server clusters With VMware Series. Through using xCat cluster management tools, they [13] built a Botnet emulation platform. Moreover, they built Waledac [14] Botnet environment on their Botnet emulation platform and systematically analyzed the defensive effect of Sybil attack on Waledac, fully demonstrating the superiority of Botnet emulation research.

100.3 Challenges and Techniques

Botnet Emulation research has incomparable advantages than other traditional research methods. However, just as described above, Botnet Emulation research is still in the research stage. In this section, we would focus on the major challenges Botnet Emulation research is facing.

100.3.1 High Security

We take it for granted that VMM is highly secured, which many companies also advocate that. However, Gartner analyst Neil MacDonald, pointed out that 60 % of the virtual server is less secure than the physical servers in one his research reports. On the BlackHat immunity built in USA 2009, Director VRT illustrated its discovery that A VMware Guest can escape to Host through using one bug in the svga driver. Especially considering that malware can be potentially highly contagious and is developed for malicious intents, we must ensure our Botnet Emulation environment is highly enclosed and secured such that malware cannot be released into outside networks or destroy Botnet Emulation environment.

In general, the security challenges come from these following areas:

1. VM Escape
   Theoretically, the Guest Operation System within VMM is not capable of knowing if it is actually running in a virtual or real environment, and there is no way to jailbreak or directly interact with the VMM. However, many VMMs dropped this important point for the pursuit of higher performance. For example, Para-Virtualization VMM promoted its performance through establishing the direct communication between Guest and VMM, which leaves more possibilities to escape the VMM. Besides, the bug existed in the VMM is also a way to break the virtual wall between Guest and VMM. Nelson Elhage [15] used the one Qemu bug: CVE-2011-175 existed in KVM, successfully promoted its privilege form VM user to Host administrator.
2. Denial of Service
   Denial of Service attack can occur in a virtualized environment when an attacker takes over a VM and is then able to gain control on the physical resources of the other VMs on the same physical host. Although the attacker cannot get the administrator privilege, but his attack would lead to other VMs's request for hardware resources unresponsive.
3. Communication among VMs or between VM and Host
   The introduction of features such as shared clipboard that allows data to be transferred back and forth among VMs and Host has introduced a security challenge, which may be treated as a gateway for transferring malicious codes among VMs or between VM and Host.

In order to address these above security challenges, we must first consider the security of Virtual Machine Monitor (VMM), which is the core of Botnet Emulation research. VMM is responsible for the maintenance of data processing and network communication on all virtual machines. So the analysis of the VMM security would be an important indicator of the whole virtual machine system, which also is a guideline when we are choosing one type of VMM for our Botnet Emulation research.

VMware, Xen, KVM are the most widespread VMM, which are widely used in various fields especially on the Cloud Computing. On the opinion of security stack, the smaller size of one software, the less interfaces exposed, the faster security patch updated, the software is relatively higher secure. So Atsec information security company make a comparison of these above three VMM, and I add other three high lightweight VMM to this table, which is illustrated in Table 100.1.

On the other hand, based on the implementation level of VMM, it is a commonsense that the more underlying layer VMM implemented on, the higher secure of the VMM. Thus we have an overall security assessment: Native Hypervisor (such as VMware ESX, Xen) ≫(is much higher secure than) Hosted VMM (Hardware assisted virtualization VMM (such as KVM) >(is higher secure than) Full Virtualization VMM (such as Qemu, Bochs and VirtualPC) >(is higher secure than) Para-Virtualization VMM (such as lguest, UML)) ≫(is much higher secure than) Operating System Virtualization VMM (such as OpenVZ, LXC).

Besides the VMM introduced above, researchers are also looking for the possibility to build a trusted VMM. Ge cheng [16] pointed out that the general-purpose VMMs are relatively large as a trusted computing base (TCB), then he addressed the challenge to reduce the overhead of virtualization and established a dynamic chain of trust when using VMM to enhance the security of VMM. However, TVMMs introduced in the research papers are always closed leading to results unverified.

In addition to the VMM security, we should also consider other security factors, such as the security of the Botnet Emulation Platform (xCat, vMatic, Cluster management tool, IaaS and so on), which can be used to manage the VMs. And in order to improve competitiveness in the Virtualization fields, many companies proposed their security strategies on the Virtualization Management Platform. VMware offers security built into the hypervisor with vShield. Other vendors can provide security to VMware virtual environments using vSafe APIs plugging into

Table 100.1 Comparison results considering security concerns

Security stack	KVM	Xen	VMware ESX	Open VZ	LXC	Lguest
Size of software	Medium	Medium to High	Medium to High	Few	Few	Few
Number of interfaces	Medium	High	High	Medium	Medium	Medium
Assurance of Development environment	Medium	Medium	Medium	Medium	Medium	Few

the hypervisor. Check Point's VPN-1 VE, which is designed to run on VMware ESX servers, separates VMs and protects inter-VM traffic. VPN-1 VE firewall would inspect traffic by moving all packets outside the VM and through the firewall instead of allowing packets to move internally form VMM to VMM without Inspection. Fortinet provides the ability to customize and segregate security offerings for different clients. Red Hat also enhanced the security of its KVM by providing the SElinux and sVirt function, which can be supported on the libvirt-based Platform [17].

100.3.2 High Transparency

With the fast innovation of Botnet technology, more and more Botnet have enhanced their self-protection ability. When the bot code found it running in a virtual environment, it will take counter-measures such as uninstall itself or camouflage itself. So if needed, our VMM used for Botnet Emulation research should be added the anti-detection features.

The main principle of detecting a virtual environment is based on the CPU architecture defects. In order to maintain the safety of system, a virtual machine monitor must trap and handle any sensitive instruction. Then it is different when executing a sensitive and non privileged instruction, because in order to trap the sensitive instruction, the VMM must expend additional time handling the instruction.

There have been many methods to detect the existence of VMM, using the methods introduced in [18], we respectively use the time-stamp counter (TSC) based timing, NTP based timing and parallel-threads based timing to analyzed KVM. We take CPUID, which is a sensitive instruction but not a privileged instruction, as the experimental instruction. We calculate the time ratio between running CPID and running ordinary instruction within native environment and KVM environment. Results illustrated in the following figures are proved that simple calculations within the virtual machine can provide accurate prediction of the existence of KVM. And other VMMs have the similar problems (Figs. 100.1, 100.2, 100.3).

Although many methods are also proposed to defend the detection of Malware such as camouflaging itself throughing hiding VM fingerprints or providing fake data, several works have shown limitations on how well the presence of a VM can be hidden. Just described in the above, the difference of running sensitive and non-privileged instruction between within VM and within Native environment cannot simply emulated. And also the discrepancies revealed by physical resources are inherently difficult to eliminate [19–21]. To successfully fake such a genuine test, a VM would need orders of magnitude more computing power than the hardware that it is emulating. Maybe the best optimal resolution is to have an overall understanding on the detection technique owned by Botnet we are willing to run on our Botnet Emulation environment.

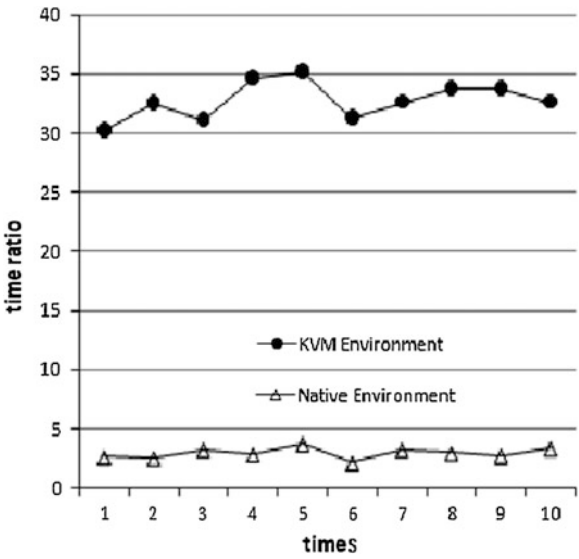


Fig. 100.1 The time ratio comparison within KVM environment (filled circle) and within Native environment (open triangle), using TSC based timing

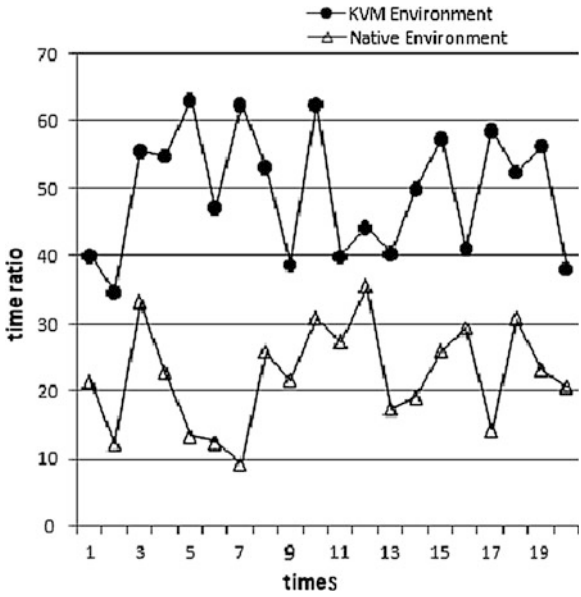


Fig. 100.2 The time ratio comparison within KVM environment (filled circle) and within Native environment (open triangle), using NTP based timing

100.3.3 Scale

One big challenge of Botnet Emulation comes from the scale of Botnet. Botnet can reach hundreds to thousands of scale, the size of some widespread Botnets can even reach to amazing one million [22]. In order to meet the requirement of scale, Sandia Library [11] deployed their Botnet on the Thunderbird Super Clusters with

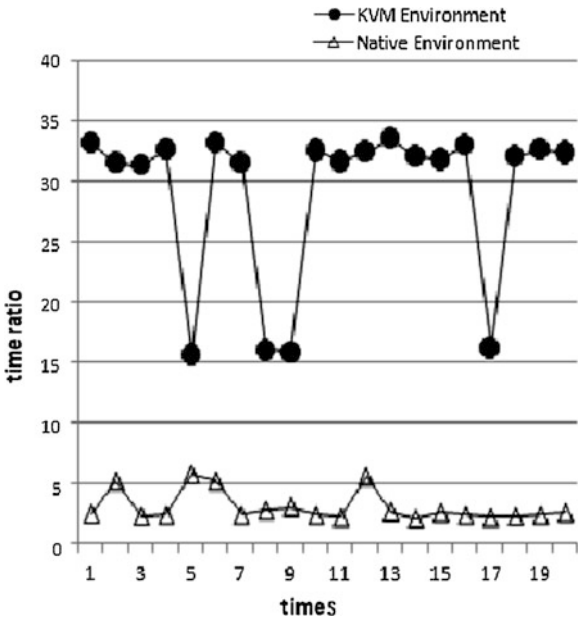


Fig. 100.3 The time ratio comparison within KVM environment (filled circle) and within Native environment (open triangle), using parallel-threads based timing

4460 nodes, and [13] used their cluster with 98 blades, each having a quad-core processor and 8 Gb of RAM, and even BEE [8] make full use of resources of Emulab global network emulation platform [9]. It is beyond the power for an ordinary researcher to own so superior hardware configuration. It is also not easy to apply for hardware resources from global network emulation platform.

In order to be able to build Botnet Emulation environment with ordinary hardware conditions, one solution is to choose lightweight virtual machine manager (VMM). In general, the higher layer VMM implemented on, the more number of virtual machines VMM can support on a physical machine. Besides, referring to a specified VMM, there are also many factors affecting the maximum number of virtual machines supported in a physical machine, some of which include:

1. Physical machine configuration: including memory size, CPU performance, whether supporting hardware-based virtualization technology as well as disk space.
2. Per-machine Metadata and Working Set Size of guest operating system: Per-machine Metadata is the key data of guest, which must be permanently existed within RAM. Per-machine Metadata limits the maximum virtual machine number. Respectively, Working Set Size is the RAM usage needed to support guest running service normally. Considering our Botnet's network I/O intensive behavior features, I/O performance is the key factor affecting maximum of virtual machine number supported by VMM.

In addition, some virtualization technology may use some optimization strategies to expand the maximum number of virtual machines, such as simplifying the

ISA architecture, adjusting the scheduling policy of the virtual machine as well as modifying the guest operating system.

We did some experiments on the maximum of virtual machines one VMM can support. We do the test on one physical machine with DDR2 667 MHz 3*1G RAM and Intel® Xeon® 5160 3.00 GHz dual-core with four threads. We take KVM and lguest as experimental objects.

Within zero-load network environment, the bottleneck is the RAM usage. We first test the effect of virtual memory size configuration of one virtual machine on physical RAM usage, which is illustrated in Fig. 100.4.

We can see that, simply configuring the virtual memory size, we can startup 30 VMs with KVM (3G/100 M) and 60 VMs with lguest (3G/50). Lightweight VMM such as OpenVZ and LXC theoretically can support more VMs. The test result [23] proved that it is possible to run up to 120 virtual environment on a 768 M memory hardware with OpenVZ, and which is a linear extrapolation. In other words, it can support 480 virtual environments on 3G memory hardware.

However, within high load network environment, it is not so optimistic.

From above picture, it is obvious that within high load network environment, when the number of running VMs on one single physical machine reach three, the host CPU usage would be used up. DDOS attack, as one of the characteristic of Botnet, would reproduce high network load, thus the high load problem must be addressed before doing DDOS experiments in the Botnet Emulation environment (Fig. 100.5).

Except for the above three challenges Botnet Emulation faces, many other factors should be considered when building the Botnet Emulation environment. For example, we need Botnet Emulation Platform to support various VMM types and have a control on the life cycle of all VMs in one Botnet. And Realism and feasibility should also be within consideration. Thus Botnet Emulation still needs progressing for a better research experience.

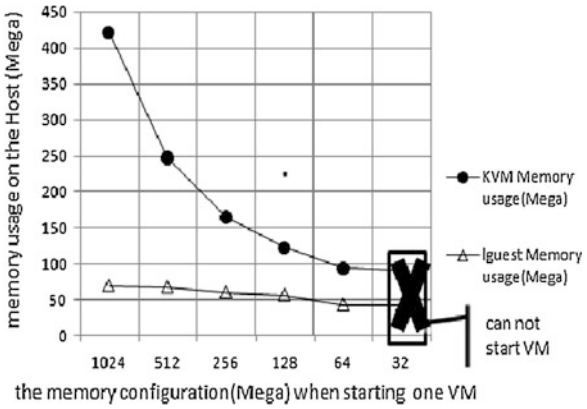


Fig. 100.4 The effect of virtual memory configuration of one VM on the Host RAM usage

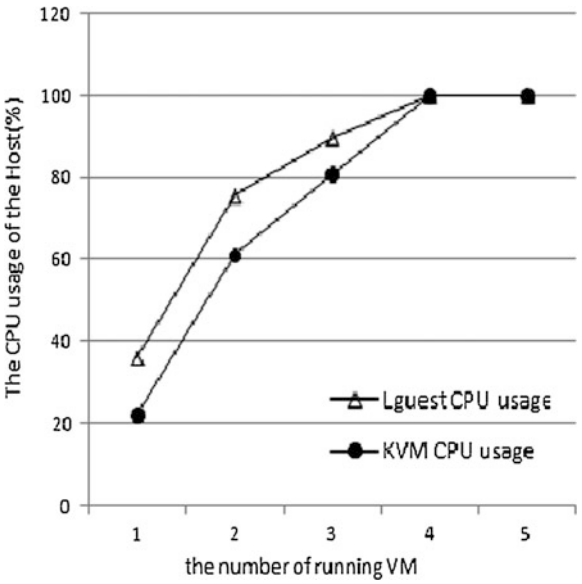


Fig. 100.5 The CPU usage within high load network environment

100.4 Conclusion

In summary, this paper describes an emerging Botnet research method—Botnet Emulation research. The article illustrated the advantages of Botnet Emulation research compared to other traditional Botnet research. And then it systematically researched on the key challenges Botnet Emulation research is facing. We did some experiments to show the difficulties and also proposed some techniques and advice to address these problems.

With the development of Botnet technology, the C&C channel has evolved from IRC to more robust P2P or hybrid architecture. Also some strategies such as fast-flux technology are deployed on Botnet to behave more hidden and hard to track. It becomes more and more difficult to research on Botnet with traditional Botnet Research methods. Thus Botnet Emulation research is a promising method to substitute for traditional methods. However, Botnet Emulation is still in a preliminary stage, which needs a lot of work to do in the future. We hope that our survey of the challenges and technology on Botnet Emulation will help motivate future research in this critical area to solve these practical challenges.

Acknowledgments This study is supported by the Hi-tech Research and Development Program of China (863 Program) under Grant No.2011AA01A205, the National Natural Science Foundation of China under Grant No.61003015, the Doctoral Fund of Ministry of Education of China under Grant No.20101102110018, the National "Core electronic devices high-end general purpose chips and fundamental software" project under Grant No.2010ZX01036-001-001, and the National Natural Science Foundation of China under Grant No. 60973008.

References

1. Zhu, Z., Lu, G., Chen, Y., Fu, Z.J., Roberts, P., Han, K.: Botnet research survey. Northwestern University, Evanston (IEEE) (2008)
2. Seifert, C., Endicott-Popovsky, B., Frincke, D., Komisarczuk, P., Muschevici, R., Welch, I.: Justifying the need for forensically ready protocols: a case study of identifying malicious web servers using client honeypots, vol. 11, no. 1 (2008)
3. Jiang, X., Wang, X.: Out-of-the-Box Monitoring of VM-based High-Interaction Honeypots. Springer, Heidelberg (2007)
4. Alosefer, Y., Rana, O.: Clustering client honeypot data to support malware analysis. Knowledge-Based and Intelligent Information and Engineering Systems. Lecture Notes in Computer Science, vol. 6279, pp. 556–565 (2010)
5. Spitzner, L.: Definition and value of honeypots. http://www.tracking-hackers.com/papers/honeypots.hmtl
6. Balas, E., Viecco, C.: Towards a third generation data capture architecture for honeynets. Indiana University, Bloomington (2005)
7. Levine, J., LaBella, R., Owen, H., Contis, D., Culver, B.: The use of honeynets to detect exploited systems across large enterprise networks. In: IEEE 4th Annual Information Assurance Workshop, West Point, NY, June (2003)
8. Barford, P., Blodgett, M.: Toward botnet mesocosms. University of Wisconsin-Madison, Madison (2007)
9. Benzel, T., Braden, R., Kim, D., Neuman, C., Joseph, A., Sklower, K., Ostrenga, R., Schwab, S.: Experience with deter: a testbed for security research. In: Testbeds and research infrastructures for the development of networks and communities, TRIDENTCOM 2006. 2nd international conference on, 2006, p. 10 (2006)
10. Benzel, T., Braden, R., Kim, D., Neuman, C., Joseph, A., Sklower, K., et al.: Experience with deter: a testbed for security research. In: Testbeds and Research Infrastructures for the Development of Networks and Communities, TRIDENTCOM 2006. 2nd International Conference on, Pub Place: IEEE, Barcelona, pp. 379–388 (2006)
11. Jackson, A.W., Lapsley, D., Jones, C., Zatko, M., Golubitsky, C., Strayer, W.T.: SLINGbot: A system for live investigation of next generation botnets. In: Conference For Homeland Security, CATCH '09. Cybersecurity Applications & Technology, Pub Place: IEEE, Washington, DC, pp. 313–318 (2009)
12. Minnich, R., Rudish, D.: Ten million and one penguins, or, lessons learned from booting millions of virtual machines on HPC systems (2009)
13. Emulating a Million Machines to investigate Botnets. http://www.hpcs2010.org/
14. Calvet, J., Davis, C.R., Fernandez, J.M., Marion, J.Y., St-Onge, P.L., Guizani, W., et al.: The case for in-the-lab botnet experimentation: creating and taking down a 3000-node botnet. In: Proceedings of the 26th Annual Computer Security Applications Conference, ACSAC '10, Pub Place: ACM. New York, pp.141-150 (2010)
15. OpenVZ density. http://zh.wikipedia.org/wiki/OpenVZ#.E5.AF.86.E5.BA.A6
16. Nelson Elhage, Virtunoid. A KVM Guest: Host privilege escalation exploit. Black Hat USA 2011
17. Cheng, G., Zou, D.Q., Li, M., Ji, C.: Trusted lightweight VMM based security architecture. Jisuanji Yingyong Yanjiu **27**(8), 3045–3049 (2010)
18. Waledac Botnet. http://en.wikipedia.org/wiki/Waledac_botnet
19. Libvirt. http://libvirt.org
20. Garfinkel, T., Adams, K., Warfield, A., Franklin, J.: Compatibility is not transparency: VMM detection myths and realities (2007)

21. Kennell, R., Jamieson, L.H.: Establishing the genuinity of remote computer systems. In: Proceedings of the 12th USENIX Security Symposium, Pub Place: USENIX Association, Washington, DC, pp.295–310 (2003)
22. Thompson, C., Huntley, M., Link, C.: Virtualization detection: new strategies and their effectiveness
23. DAMBALLA, Top 10 Botnet Threat Report, 2010

Chapter 101
A New System for Summoning and Scheduling Taxis

Hengjian Liu, Jing Xiong and Yiren Ding

Abstract In order to solve the problems in the process of artificial scheduling for taxis, which include expensive labor cost, limited service time and concurrency, a new self-help taxi service schema is proposed. The schema is composed of mobile phone, intelligent server and taxis equipped with GPS. Based on the schema we designed a new system for summoning and scheduling taxis. In the system, three ways, including searching, GPS, electronic map, are available for passengers to locate themselves using mobile phones. And on the server side, we introduced Geographic Information System (GIS) and adopted the ant colony algorithm based on taxi drivers' driving habits. As a result, passengers can get their locations easily with almost all kinds of mobile phones, and the server can identify passengers' requests by itself and give a consistent route with the one adopted by taxi drivers. Finally, the system has the advantages of wide applicable scope, low labor cost, and large concurrency.

Keywords Mobile phone · Taxi · Self-help · GIS · GPS · Colony algorithm

101.1 Introduction

As an important complement way for buses, taxis provide convenience for public travel. However, under some special conditions such as at night and in remote areas, it's not so easy to get a taxi, because the taxi driver could not know the position of passengers timely, only blindly wandering in the city [1]. To solve this question, the phone-call taxi service gradually arises and develops, and it has gradually affected people's taxi travel habits. In Suzhou, for example, the phone-call taxi business

H. Liu (✉) · J. Xiong · Y. Ding
Faculty of Information Engineering, China University of Geosciences, Wuhan, China
e-mail: liuhengjiangis@foxmail.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_101,

grows from daily 300 items in 2004 to daily 10,000 in 2010 [2], the front and rear to a 50-fold increase. This shows that the traditional way of street hail taxi service is being challenged by the development of phone-call taxi service. But in another way the phone-call taxi service has high human costs and small amount of service concurrent. And these issues will be important obstacles to its further development. How to develop in the next stage? By analyzing the trajectory of the development of the taxi industry and limitations of the phone-call way, we propose a new direction in this paper: *the self-help taxi service* (Fig. 101.1).

The self-help taxi service means that passengers can take taxis by mobile phone instructions, without help from service staff. To attain this goal, there must be an intelligent server to schedule taxis. However, before scheduling taxis, the server must understand two questions well, one is where taxis are and the other is where passengers are. For now, the former question can be solved easily thanks to the essential equipment—GPS in taxis. For the latter one, some scholars suggest phones with GPS module [1], and some others suggest LBS (Location-based service) provided by mobile operators. But all mobile phones have not a GPS module which may be invalid around tall buildings and the LBS is difficult to use because of its dependence on operators and other legal issues. In this paper, three ways, including searching, GPS and electronic map, are provided to locate passengers for the first time, with lower requirements for mobile phones and wider applicable scope. Based on the above considerations, this paper proposes a new taxi summoning and scheduling system, by which passengers could locate themselves precisely and ask for taxis by instructions. Besides, passengers could also obtain a predicted route from current location to destination and some other information.

The system uses Android as a mobile development platform, because the Android platform is very representative, occupying 90.6 % of the smart phone shipments in August 2012 [3]. Electronic map selecting the Baidu mobile map, because it has lots of data and is easy to invoke. GIS platform in the server side to select MapGISK9, because it's a popular and powerful GIS software in China.

101.2 Main Functions of the System

The system uses C/S architecture. The client features include GPS positioning, address search, map operations, place favorites, etc. The main server-side features include POI search, taxis monitoring and scheduling. Figure 101.2 shows the normal processing flow.

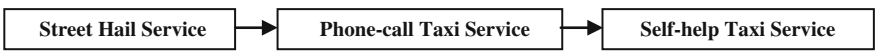


Fig. 101.1 The trend of taxi service

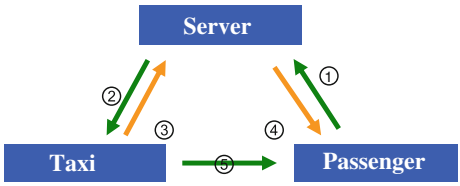


Fig. 101.2 A five-step flow for processing passengers' requests

Process Description:

1. Send request

Passengers can send requests mainly in two ways: mobile phone application or a short message. Mobile phone application conforms to the trend of smart phones, and has the following characteristics: consuming internet traffic; with vivid embedded map, enabling convenient operations on map; enabling automatic positioning with GPS; receiving positions of specific no-load taxis, so that passengers can visually see the positions and movements on map. The short message way applies to the general non-smart phones by sending short messages containing the address name to be located. After receiving the message, the server invokes the published web service of address matching, and then filters out a group of high similarity address records. Afterwards passengers select the addresses, and give the feedback to server which then confirms passengers' geographical location.

2. Ask for confirm from taxi drivers

Firstly, the server receives passengers' locations and other attribute information. Secondly, the server searches no-load taxis within a certain distance. Finally, the server distributes requests to terminals on no-load taxis. While the terminals get requests, they could show locations on the embedded map.

3. Confirm

In this process, the principle of *first-answer-first-get* is adopted for drivers.

4. Receive results

The results include license plates, drivers, distance, arrival time, ID, etc. Passengers could also receive the location information of the taxi according to a certain time interval on their smart phones and update positions on the embedded map.

5. Go for passengers

101.3 Design and Implementation of the System Functions

The server-side applications of the system work on Windows Server 2008, using J2EE framework, the development tool is MyEclipse, and the GIS platform is MapGISK9. The client side is developed based on Android 2.3, using eclipse as the development tool and Baidu Map as the mobile phone map.

101.3.1 Main Functions of the Server Side

The server-side functions are mainly composed of taxi monitoring and scheduling, POI searching, etc.

101.3.1.1 Taxi Monitoring and Scheduling

Taxi monitoring mainly focuses on the real-time location and status of taxis, while scheduling is primarily responsible for receiving and analyzing passengers' requests, and to arrange nearby taxis in response to passengers' demands. Taxi monitoring works on the server, receives the GPS taxi location information and status information, and then marks the position of the vehicle in the electronic map on the monitor screen. To distinguish between load taxis and no-load taxis, icons with different color are used. In the scheduling module, the server opens ports to listen for requests from passengers, and open multi-threads to process tasks, like taxis searching, route planning, requests responding, etc. Specially, for route planning, a model based on ant colony optimization algorithm and taxi driving habits is adopted to get the best route planning results. Taxi driving habits are analyzed by using taxis' GPS data.

101.3.1.2 POI Searching

POI, namely, Point of Interest, is one of the core functions of the geocoding. It is a public service to enable passengers to search for a suitable location, and it's different from the standard address with house number style. Passengers usually use address aliases more often than house number addresses. Given the positioning accuracy requirements and domestic house number inconsistency, interpolation algorithm is denied here, so all searched address results come from the database of addresses. Figure 101.3 shows the normal steps for address matching.

With the development of distributed and loosely coupled GIS applications, it's a better choice to establish services with cross-platform operations, as well as one deployment and multiple calls [4]. This requires a combination of Web Services technology and geocoding technology to encapsulate geocoding into Web Services. In this paper we use Axis engine as Web Service development tool.

101.3.2 The Client

The client-side centers on the acquisition of the position information. There are three ways to get located: mobile phone's GPS module; tapping on the phone map; searching through the POI service.

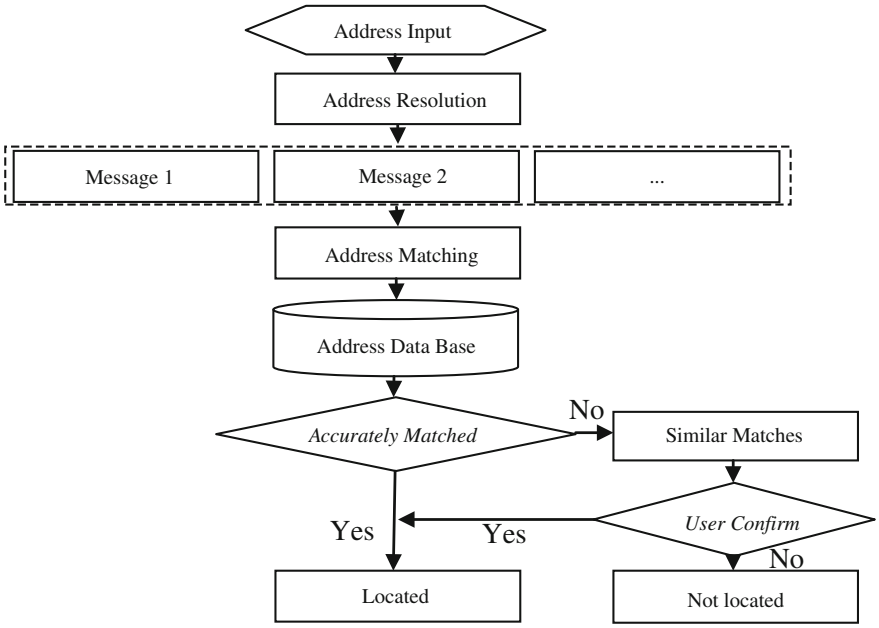


Fig. 101.3 Normal steps of POI service

101.3.2.1 Mobile Phone's GPS Module

Taking Android as an example, there are two steps to attain longitude and latitude information through programming.

1. Get an instance of class *LocationManager*, code as:

LocationManger
lm = (LocationManager)this.getSystemService(Context.LOCATION_SERVICE);

2. Register GPS location listening

LocationListener locationListener = new LocationListener() {…}
lm.requestLocationUpdates(LocationManager.GPS_PROVIDER, 1000, 1, locationListener); In the object *locationListener* you can implement the code to get the latitude, longitude, altitude, etc.

101.3.2.2 Tapping on the Phone Map

In this way, a conversion is necessary to convert screen coordinates into geographic coordinates. The screen coordinate origin is located in the upper left corner, the x-axis positive direction to the right, the y-axis positive direction down. While in

the geographic coordinate system, the x-axis represents the longitude, positive direction to the right, the y-axis represents the latitude, positive direction up.

101.3.2.3 Searching Through the POI Service

To get the location through the POI services, passengers should follow three steps: Firstly, send a string in the form of "'start point address' + ';' + 'endpoint address'", such as "University of Geosciences; Wuchang Railway Station". Then matched addresses would be sent to passengers, like "University of Geosciences (Wuhan Campus) main entrance -1; University of Geosciences (Wuhan Campus) west second entrance-2; University of Geosciences River City Academy -3 @ Wuchang Railway Station West Exit-a; Wuchang Railway Station East Exit-b". In the above example, there is a separator "@" which divided a string into two parts, the former contains similar numbered addresses with the start point address and the latter contains similar numbered addresses with the end point address. Finally, passengers send a message to confirm the present position and destination, such as "1a", in which "1" stands for "University of Geosciences (Wuhan Campus) main entrance" and "a" stands for "Wuchang Railway Station West Exit".

101.4 Algorithm

101.4.1 Address Matching

The key to address matching is to determine the matching strategy and the structure of data in database [5, 6]. The address database used in this system is different from the standard one, because when taking a taxi in daily life passengers often use a commonly used alias rather than the address with house number. In order to reduce the query and comparison time, a better approach is to create indexes for address fields [6, 7]. In this paper, we combine the positive maximum matching rule with TRIE dictionary tree. Firstly, we also create indexes for address items using TRIE dictionary tree. When dealing with the addresses that users enter, we can normalize those addresses and then pick up useful information. After that, we can find all matched address nodes according to the positive maximum matching rule.

101.4.2 Route Planning

Route planning is a classical problem in GIS, and traditional route planning mostly focuses on finding shortest path in terms of time or distance cost. Gradually some new algorithms and models arise and more factors are imported by scholars, like

direction changing constraint and limitation, road level, passing time, road accident, etc. Among these scholars, Luliang Tang etc. think in terms of taxi drivers, that taxi drivers are familiar with every corner of the city and have lots of experience in driving [8]. Along this way, they record passing time and frequency on every road section by collecting float cars information, and then establish the initial road pheromone level network. In this paper, we combine the ant colony optimization algorithm with an initial pheromone network based on taxi traces, and update the pheromone timely as the number of ant increases. Finally, we could get the optimal solution from all ants' traces. The function of the abstract model is as follow:

$$\mathrm{R(t)} = E[L,\ T(t),\ P(t)] \tag{101.1}$$

where L is the path length, T (t) is travel time in time period t, P (t) is the value of pheromone level of the road network. E[L, T(t), P(t)] is a multi-objective path selection decision model based on the above three parameters.

$$L = \sum_{i=1}^{n} L(E_i) \tag{101.2}$$

$$\mathrm{T(t)} = \sum_{i=1}^{n} T(E_i, t) = \sum_{i=1}^{n} \frac{L(E_i)}{\overline{V}(E_i, t)} \tag{101.3}$$

$$\mathrm{P(t)} = \sum_{i=1}^{n} \left(L_{pmax} - P(E_i, t)\right) = nL_{pmax} - \sum_{i=1}^{n} P(E_i,\ t) \tag{101.4}$$

where $\mathrm{L(E_i)}$ is the length of road section $\mathrm{E_i}$, n is the number of edges; $\mathrm{T(E_i,\ t)}$ is travel time through road section $\mathrm{E_i}$ in time period t, $\overline{\mathrm{V}}(\mathrm{E_i,\ t})$ is the average speed on road section $\mathrm{E_i}$ in time period t; $\mathrm{L_{pmax}}$ is the maximum level value of pheromone level of the road network, $\mathrm{P(E_i,\ t)}$ is the pheromone value of road section $\mathrm{E_i}$ in time period t. The function of $\mathrm{P(E_i,\ t)}$ is estimated as follows:

$$\mathrm{P(E_i,\ t)} = \left(\frac{\theta \mathrm{B_l(E_i)}}{\mathrm{B_{lmax}}} + \frac{\mu \mathrm{F_l(E_i,\ t)}}{\mathrm{F_{lmax}}}\right)\mathrm{L_{pmax}} \tag{101.5}$$

where $\mathrm{B_l(E_i)}$ is the construction level of road section $\mathrm{E_i}$, $\mathrm{F_l(E_i,\ t)}$ is the passing frequency level of road section $\mathrm{E_i}$ in period t, $\mathrm{B_{lmax}}$ is the maximum construction level, $\mathrm{F_{lmax}}$ is the maximum passing frequency level, θ is the influence coefficient which describes the influence of construction level on the pheromone level of the road network, μ is the influence coefficient which describes the influence of passing frequency level on the pheromone level.

Based on the abstract model R (t), the process of finding the optimal solution is a process of making L as short as possible, T (t) as little as possible, and P (t) as low as possible. So the decision-making function can be made as follows:

$$R = \min\left(w_1 L' + w_2 T' + w_3 P'\right) \tag{101.6}$$

$$L' = L/L_{min} \tag{101.7}$$

$$T' = T(t)/T_{min} \tag{101.8}$$

$$P' = P(t)/P_{min} \tag{101.9}$$

where L', T', P' are normalized values, L_{min} is the minimum route length from the origin to the destination, T_{min} is the minimum travel time, P_{min} is the minimum pheromone value, w_1, w_2, w_3 are corresponding weight coefficients.

Under the guidance of the decision-making function, it's a good way to find the optimal solution among all solutions through the ant colony algorithm. Firstly, use formula (101.5) to calculate the road network initial pheromone distribution and then set the total number of ants as m, so ant k (k = 1, 2,…, m) moves forward according to the pheromone. In the moving process, the remaining pheromone and heuristic information on the path both affect moving action. The probability for ant k to choose road section E_j in period t is calculated as follows:

$$p_{ij}^k(t) = \begin{cases} \frac{[\tau_j(t)]^\alpha \cdot [\delta_j(t)]^\beta}{\sum_{n \in allowed_k} [\tau_n(t)]^\alpha \cdot [\delta_n(t)]^\beta}, & if \quad j \in allowed_k \\ 0, \text{ if } \quad j \notin \text{allowed}_k \end{cases} \tag{101.10}$$

where $\tau_j(t)$ is the pheromone of road section E_j in period t, $\delta_j(t)$ is the value of the heuristic function of road section E_j in period t, $\delta_j(t) = 1/D_j$, D_j is the vertical distance between road section E_j and destination, $allowed_k$ is a set of possible road sections, which are linked with E_j, except for the used road sections, α, β are corresponding exponential coefficients.

To avoid the case that too much remaining pheromone floods and covers the heuristic information, the pheromone must be updated after the ants complete the path search. The updated pheromone for E_j could be calculated as follows:

$$\tau_{jnew}(t) = (1 - \rho)\tau_j(t) + \Delta\tau_j(t), \ \rho \in (0, \ 1) \tag{101.11}$$

$$\Delta\tau_j(t) = \sum_{k=1}^{m} \Delta\tau_j^k(t) \tag{101.12}$$

$$\Delta\tau_j^k(t) = \begin{cases} \frac{Q}{R_k}, \ if \ ant \ k \ passes \ E_j \\ 0, \text{ if ant k does notpass } E_j \end{cases} \tag{101.13}$$

where ρ is the pheromone volatile coefficient, $\tau_j(t)$ is the origin pheromone on section E_j, $\Delta\tau_j(t)$ is the new pheromone left by ants, $\Delta\tau_j^k(t)$ is the new pheromone left by ant k on section E_j in period t, Q is the strength of the pheromone, R_k is the value of the objective decision-making function for ant k.

Table 101.1 Similarities between planned route and history routes from Ruan Jian Yuan to Wuchang railway station

Matched segments	Route number	Segment match ratio (%)	Average segment match ratio (%)	Full match ratio (%)
18	20	100	97.5	83.3
16	2	88.9		
15	1	83.3		
14	1	77.8		

To evaluate the algorithm, we select some origins and destinations, and then find routes from database of taxis' history GPS trace. For example, from Ruan Jian Yuan to Wuchang Railway Station there are 18 segments in planned route derived from the algorithm, and among the 24 routes derived from the database, 20 routes are the same with the planned one and four are similar (Table 101.1), so the full match ratio is 83.3 % and the average segment match ratio is 97.5 %. Besides, there are 30 pairs of origins and destinations as verifying samples, the minimum full match ratio is 78.0 %, the maximum is 94.4 % and the average is 84.4 %.

101.5 Conclusion

This paper introduces a new taxi summoning and scheduling system, which can not only provide passengers with a self-help taxi service, but also reduce the car management's labor costs and improve service efficiency. The system takes mobile phone as terminal and provides three ways to locate, solving the problem of locating for almost all kinds of mobile phones. In the process of route planning, the system takes taxi drivers-urban traffic intelligent individuals as ants, and uses the ant colony algorithm, which turns out to be effective and consistent with the actual route. Furthermore, there is much work to do to make the system perfect, especially in aspects like address database supplement and standardization, taxis' GPS data mining, etc.

Acknowledgments The project was supported by the Fundamental Research Founds for National University-China University of Geosciences (Wuhan), under the grant 1210491B08.

References

1. Zhou, X., Zhao, H., et al.: Taxi calling and scheduling system based on GPS. Comput. Eng. Des. **30**(21), 4995–4997 (2009)
2. Suzhou Evening News. Changes of public transportation in Suzhou in the decade from 2002 to 2012. http://sz.xinhuanet.com/2012-08/14/c_112717193_3.htm
3. China Academy of Telecommunication Research of MIIT. The operational status of the domestic (China) mobile phone industry in August 2012. http://www.catr.cn/tegd/shchgch/201209/t20120911_847434.html

4. Li, X., Chen, Y., et al.: The application and web publishing techniques of address matching based on web services. Remote Sensing Information, vol. 5, pp. 56–59. Beijing (2006)
5. Guo, H., Song, G., et al.: Design and implementation of address geocoding system. Comput. Eng. **35**(1), 250–252 (2009)
6. Sun, C., Zhou, S., et al.: Chinese geo-coding based on classification database of geographical names. J. Comput. Appl. **30**(7), 1953–1955 (2010)
7. Zhou, S., Tang, B., et al.: The study of address tree coding based on the maximum matching algorithm in courier business. Commun. Comput. Inf. Sci. **113**, 38–45 (2010)
8. Tang, L., Chang, X., Li, Q., et al.: Public travel route optimization based on ant colony optimization algorithm and taxi GPS data. China J. Highw. Transp. **24**(2), 89–95 (2011)

Chapter 102
A Social Interest Indicator Based Mobility Model for Ad Hoc Network

Kaikai Yue, Demin Li and Peng Li

Abstract Due to the deficiency of the social relationship in many synthetic models such as RPGM and RWP in ad hoc network simulations, a new mobility model based on the interest indicator is proposed in this paper. By analyzing the moving gathering relationship among individuals, the authors propose the concept of interest indicator as well as the interest distance and then further propose a social interest indicator based mobility model for ad hoc network. Simulation result shows that the proposed model is more realistic than RPGM and RWP and the authors also analyze the influence of the model on DSDV and AODV. This proposed model can be used in circumstances such as exhibition, party and battle.

Keywords Mobility model · Social relationship · Ad hoc network

102.1 Introduction

Ad hoc network is a new technology which is often used for emergency relieves. Often the performance of an ad hoc network depends on what protocols it is using. Presently, the validation of a new protocol for ad hoc network relies on simulations, which, in turn, relies on how realistic the mobility models are. However, most synthetic models available now are very simplistic and put emphasizes on the ease of implementation other than soundness. It is so important to establish a model that can better describe the mobility of real individuals. Mobility is strongly affected by humans to socialize and cooperate with each other in one form or

K. Yue · D. Li (✉) · P. Li
College of Information and Science Technology, Donghua University, Shanghai, China
e-mail: deminli@dhu.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_102,

another. Fortunately, humans are known to socialize in some particular ways that can be mathematically modeled and studied in social science in the years to come.

Recent years, lots of synthetic models have been put forward. A comprehensive review of the present mobile models can be found in [1]. The simplest one is the Random Walk Mobility Model [2] (similar to Brown Motion), in which mobility of the entities in this system is purely based on randomness. An improvement is the Random Way-Point Mobility Model [3], in which there are pauses between the changes in direction and speed. However, the two most-used models mostly focus on the mobility of single individuals and do not consider the relationship of individuals to groups. In order to depict the relationship of individuals to groups, there are also many group models that have been put forward. The mostly used model for group is the Reference Point Group Mobility Model [4], in which a reference point is selected, the other entities are moving with the vector of superposition with the reference point. All entities are associated to the reference point and mobility of the reference point determines the mobility of the whole group.

In reality, most mobile devices are carried by humans and the mobility of the devices is based on the socialization and cooperation of human. Thus, it is necessary to define group mobility model that heavily depend on the structure of the relationships among people carrying mobile devices. In order to construct a model which reflects the mobility of real individuals, we should take the social relationship into consideration.

Fortunately, some models on this subject have been proposed. One model is Community Based Mobility Model [5]. In this model collections of hosts are grouped together in a way that is based on the social relationship of an individual to a group. Individuals are grouped into several sub groups according to the attraction of a group to an individual. Another one is Dynamical Interesting Induced Mobility Model [6]. In which an interest vector is proposed to represent the state of an individual at some time, but it mainly focuses on the mobility of individuals, not involve the concept of groups.

In this paper we propose a new model based on interest factor, in which the mobility of an individual is determined by the relationship of itself to the whole group. And we also consider the relationship between individuals, and use the relationship as a parameter to determine the mobility of individuals.

102.2 Design of the Model

The mobility of an individual is affected by the others. In this section, we will discuss the establishment of our model in detail according to the gathering relationship of an individual to the others around it.

102.2.1 Indicator Calculation

Recent researches in social network have obtained great achievements, which can enlighten us to design more realistic social models. When we consider a big party or a convention, we will figure out that there is a gathering relationship of an individual to a group. That is when an individual has the similar interests to the whole group it will linger in the group for a long time to communicate with the members; otherwise, he will ignore the group and walk away. But as time elapses, his interest will change, when he loses interest in the group, he will leave this group and walk along to the others.

We can use a time-varying vector to represent the strength of a person's interest in different things. We assign each individual a vector called interest vector, which represent the interest level in something.

We use $(e_1\ e_2\ e_3 \ldots\ldots e_N)$ to represent N interests, and $I_k^{e_i}(t)$ (ranging from 0 to 1) represents the interest level of individual k has in interest e_i at t. So the vector $\left(I_k^{e_1}(t)\ I_k^{e_2}(t)\ I_k^{e_3}(t) \ldots\ldots I_k^{e_N}(t)\right)$ which is called interest vector represents the interest state of individual k at t.

When an individual comes to an interest group, it will first exchange its interest vector with each member of the group to determine whether this group conforms to its interest or not. If this group conforms to its interest, the individual will linger in this group for a while to communicate with the members inside; otherwise, it will ignore the group and walk along.

We use formula (102.1) to calculate the interest distance of individual k and j:

$$dis(k,\ j) = \frac{\left(\sum\limits_{\substack{i=1 \\ k \neq j}}^{N} \left(I_k^{e_i}(t) - I_j^{e_i}(t)\right)^2\right)^{1/2}}{\sqrt{N}} \tag{102.1}$$

$dis(k,\ j)$ is called interest distance and represents the difference level of the two individuals. It is obvious that smaller interest distance means the two individuals have similar interest and bigger chance that they will mobile in the same scenario.

Assuming that there is a group containing $z$members, we use an interest distance vector $(dis(k,\ 1)\ dis(k,\ 2)\ dis(k,\ 3) \ldots\ldots dis(k,\ z))$ to represent the interest distance of an individual to all of the members in the group.

We also need an indicator to measure the relationship of individual k to the whole group. Here we define formula (102.2):

$$DF_k = \frac{\sum\limits_{\substack{j=1 \\ j \neq k,}}^{z-1} dis(k,j)}{z-1} \tag{102.2}$$

In which, DF_k is the difference factor which describes the interest difference between single individual k and the whole group. z is the total number of people characterized in the system. It is obvious that smaller DF_k mans smaller difference between individual k and the whole group, which in turn, means that the whole group has bigger attraction on k. In this scenario, there is bigger possibility that individual k will linger in this group for some time to communicate with the members inside.

102.2.2 Model Establishment

In this section, we will use the difference factor to establish our mobility model in detail. When an individual comes to a group, it will first exchange its interest vector with all of the members inside the group and calculate its difference factor DF_k to determine whether to linger in the group or not. If DF_k is bigger than the threshold value we set before, the individual will ignore the group and walk along to another one; otherwise the individual will come into the group and select a member which is mostly close to it to communicate with. Since the interest of some individual is not stable and will change as time elapses, so it will repeatedly calculate its DF_k at each interval δt and once its DF_k is bigger than the threshold value, it will leave the group and walk along to another one (Fig. 102.1).

As we know that difference factor can be a factor which affects the mobility vector between an individual and a group. When individual k starts to move inside the group, it will select an individual j which is mostly closest to it as a reference point, the mobility vector of individual k is the sum of the mobility vector of j and a random vector. It can be measured with the following formula (102.3):

$$\begin{cases} V_k(t) = dis(k,\ j)V_r + (1 - dis(k,\ j))V_j(t-1) \\ \theta_k(t) = dis(k,\ j)\theta_r + (1 - dis(k,\ j))\theta_j(t-1) \end{cases} \tag{102.3}$$

where $\{V_k(t),\ \theta_k(t)\}$ and $\{V_j(t),\ \theta_j(t)\}$ respectively represent the mobility vector of individual k and j, $\{V_r, \theta_r\}$ represents the random mobility vector of individual k which conforms to uniform distribution.

Figure 102.2 shows us the mobility of individual k from $t-1$ to t, when the individual calculates its DF_k and determines to move inside the group, it will select

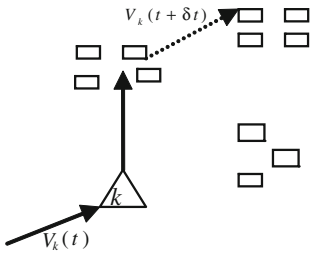


Fig. 102.1 Mobility of an individual to a group

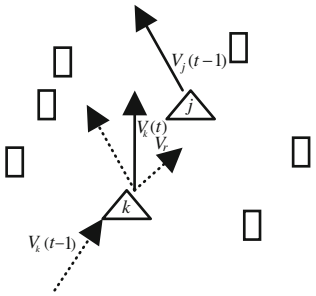


Fig. 102.2 Mobility of two individuals

the nearest individual j as a reference point. At t, the mobile vector of individual k is the sum of the random vector V_r and $V_j(t-1)$.

102.3 Simulation and Evaluation

In the Sect. 102.2.2, we introduced a general overview of the proposed model. In this section, we will describe how interest factor influence the evolution and the dynamics of the mobile scenario by simulation. Here we compare our model with the RPGM and RWP to show that our model is more realistic and can better reflect the characteristics of the mobility of humans.

We divide 50 nodes into five groups in RPGM and the same way in the proposed model. Then set the interest level of each interest from 0 to 1 in random, the threshold value of the DF_k is set to 0.3. We will evaluate our model by those three following indexes:

(a) Average degree of spatial dependence: an index to reflect the relationship of an individual to the neighbors.
(b) Routing overheads: an index to reflect the performance of a routing protocol.
(c) Average link duration: an index to reflect the duration to interact with the other nodes.

Figure 102.3 shows the average degree of spatial dependence. We can figure out that value is high when the individual moves at low speed. That is because when an individual moves at low speed, there is great chance that is communicating with the others. RPGM has the highest value because all individuals are related to the reference individual in the group; there are the most similarities among the individuals' moving vectors. Since individuals in RW are all moving randomly, there are no relationships among individuals, so the value is very low.

Figure 102.4 shows the average link duration. We can see that with the increasing of the speed, the value is decreasing for each model. RPGM has the highest value because group members are related to the reference individual and they have to communicate with each more often. The value of IIBM is lower

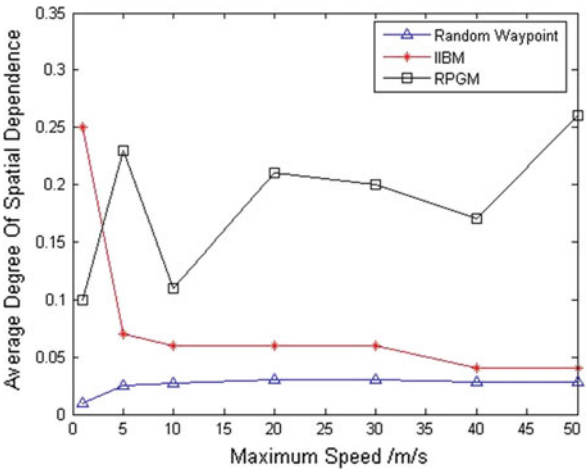


Fig. 102.3 Average degree of spatial dependence

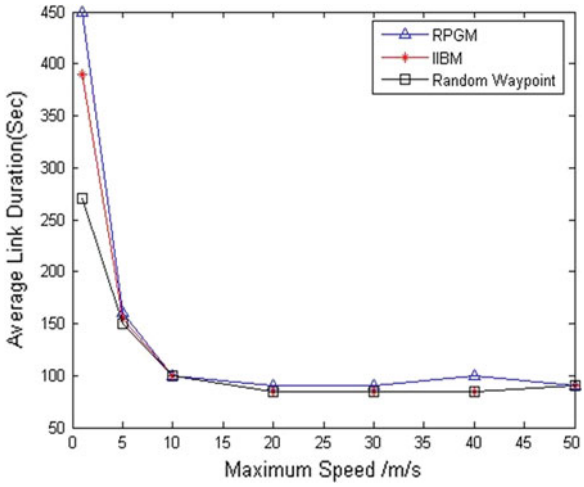


Fig. 102.4 Average link duration

because it reflects the gathering relationship of the individuals. The relationship is not as strong as RPGM but stronger than RWP. Since individuals in RW are all moving randomly it has the lowest value.

Figures 102.5 and 102.6 shows the routing overheads of DSDV and AODV. We can see that routing overheads is relatively lower using DSDV than AODV. That is because when individuals are moving at low speed, routing topology does not chance quickly. So it is better to use DSDV when nodes are moving in a scenario like IBMM when considering routing overheads.

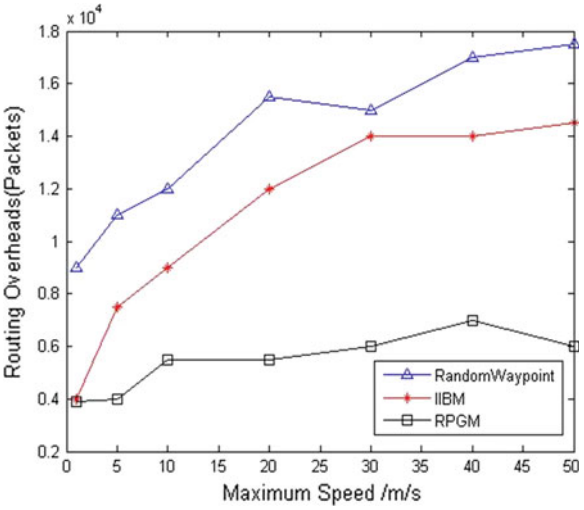


Fig. 102.5 Routing overheads of DSDV

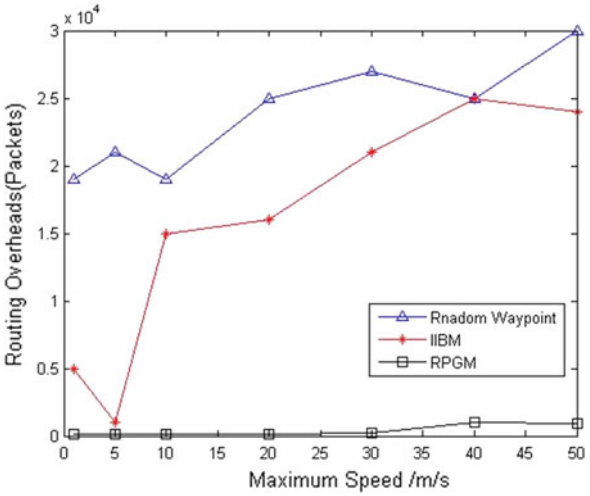


Fig. 102.6 Routing overheads of AODV

102.4 Conclusions and Future Work

In order to test the performance of the routing protocols in ad hoc network, researchers usually use simulation. The performance of routing protocols depends on the mobility of the entities in the scenario. Most present mobility models are simplistic and put emphasis on ease of implementation other than soundness. In this paper, the authors proposed a new mobility model based on relationship which can be used for conventions, parties, etc. and interpret how relationship affects the mobility of two individuals. They also use simulation to prove that the proposed

model is more realistic than those proposed before. The authors also proved that it is better to use DSDV other than AODV when considering routing overheads in a scenario like IIBM.

Though it is still a heavy task to implement communication based on social network, but the authors believe that it is possible to design mechanisms based on the evaluations of social networks that connects the individuals carrying devices, in order to build more efficient and, at the same time, more reliable systems.

References

1. Camp, T., Boleng, J., Davies, V.: A survey of mobility models for ad hoc network research. Wirel. Commun. Mob. Comput. **2**(5), 483–502 (2002) (Special issue on Mobile Ad Hoc Networking: Research, Trends and Applications)
2. Einstein, A.: Investigation on the Theory of the Brownian Mobility. Dover Publication, New York (1956)
3. Johnson, D.B., Maltz, D.A.: Dynamic sources routing in ad hoc wireless network. In: Imielinski, T., Korth, H. (eds.) Mobile Computing, pp.153–181. Kluwer Academic Publishers, NewYork (1996)
4. Hong, X., Gerla, M., Pei, G., Chiang, C.C.: A group mobility model for ad hoc wireless networks. In: MSWiM 99 Proceedings of the 2nd ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems, pp. 53–60. ACM, New York (1999)
5. Musolesi, M., Mascol, C.: A community based mobility model for ad hoc network research. REALMAN'06 (2006)
6. Li, X., Hu, T., Wang, D., Zhang, D.-F.: Dynamical interesting induced mobility model for simulation of mobile ad hoc networks. J. Syst. Simul. **21**(16), 4926–4930 (2009)

Chapter 103
Topological Map and Probability Model of the Multiple Plane and Multiple Stage Packet Switching Fabric

Xiangjie Ma, Xiaozhong Li, Xinglong Fan and Lingling Huo

Abstract The Multiple-Plane and Multiple-Stage packet switching fabric is one of novel scalable switching technologies in Internet and new generation of information network. As lack of theoretical basis, many important MPMS technical problems were still not resolved currently. In this paper, researchers studied the universal topology and the logical mapping graph. Based on the directed graph theory, the directed graph model (DGM) and probability model of the MPMS packet switching fabric are provided. The primary theoretical basis is built for MPMS fabric. Simulation experiments show that the probabilities of the proposed probability model are quite stable with a deviation less than 2 %, which is much better than present results.

Keywords Multiple plane and multiple stage · Universal model · Logical mapping · Directed graph · Probability model

103.1 Introduction

The scalability of switching fabric in switching equipments is faced with severe challenges with fast evolution of public information network (Internet) towards high-speed transmission [1, 2]. The MPMS packet switching technology, which converges a lot of small-scale switching units into a large-scale switching network, can break up the capability limitation of a single switching unit and greatly improve the scalability of switching equipment [3]. Therefore, the MPMS packet switching technologies are attracting increasing attention from academic and industry areas [4, 5]. The MPMS packet switch prototype was firstly put forward

X. Ma (✉) · X. Li · X. Fan · L. Huo
Air Force Engineering University, Beijing, China
e-mail: maxiangjie100@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_103,

by HJ Chao and JS Park from Polytechnic University, and they studied an MPMS packet switching system called TrueWay based on large number of FPGA devices [6]. In 2007, HJ Chao and JS Park further studied the problem of "flow control" in MPMS system based on level to level flow control, end to end flow control and credit based flow control [7]. In 2009, a novel scalable MPMS switching fabric was provided and the graphic model were studied [8, 9]. A novel bandwidth guaranteed scheduling algorithm called BG-CRRD was provided, delivering 100 % throughput under uniform traffic, and allocating output-link bandwidth fairly for the reserved flows in the overloaded case [10].

As lack of theoretical basis and mathematical model, many important MPMS technical problems, however, have not been resolved up to now [11]. For example, which kind of topology properties should be adopted to realize MPMS scalability [12]? How does the internal topology affect the MPMS switching and scheduling performance [13]? Based on former work, this paper proposed an MPMS general topology and logical mapping graph (LMG), and studied the MPMS probability model with the directed graph method based on above research results. The rest of this paper is organized as follows: the general topological graph and logical mapping graph is provided in Sect. 103.2; the MPMS graph model and probability model are set up in Sect. 103.3; we analyze the MPMS experiment and simulation results in Sect. 103.4, and conclude this paper in Sect. 103.5.

103.2 General Topological Graph and Logical Mapping Graph

Symbols used in this paper are listed as follows:

N	The port count of *IP*/*EP* in MPMS packet switching fabric.
W	The count of cell demultiplexers in MPMS packet switching fabric.
P	The count of switching planes in MPMS packet switching fabric.
Q	The count of cell multiplexers in MPMS packet switching fabric.
Z	The zenith set in MPMS DGM model.
Z_i $(i = 1, 2,\ldots,5)$	The zenith sub-sets in MPMS DGM model.
z_{ik} $(i = 1, 2,\ldots,5;$ $k = 1, 2,\ldots,N(W, Q, P))$	Zeniths in Z_i.
B	The directed border set in MPMS DGM model.
B_j $(j = 1, 2, 3, 4)$	The directed border sub-sets in MPMS.
$Bi(k, l)(i = 1, 2, 3, 4;$ $k,l = 1, 2,\ldots,N(W, Q, P))$	Directed borders in DGM model.

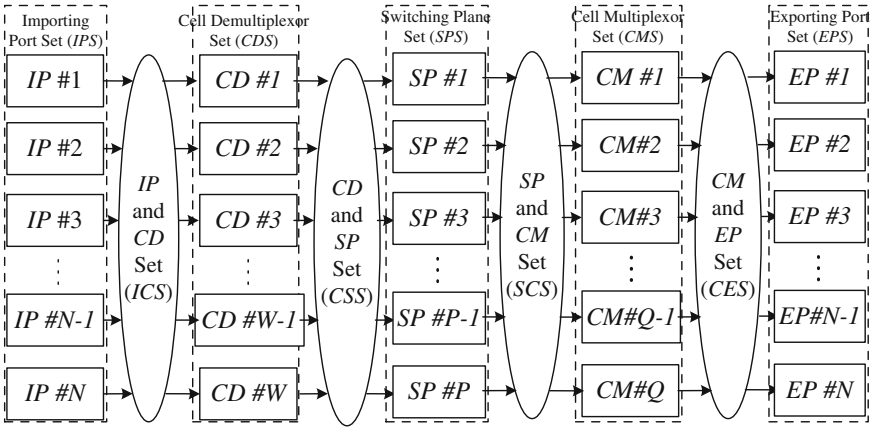


Fig. 103.1 General topological graph of MPMS packet switching fabric

According to the research result in [8], the general topological graph of the MPMS packet switching fabric can be abstracted as shown in Fig. 103.1. We can see that the general topological graph includes five categories of function unit sets and four categories of connecting link sets. And the five categories of function unit sets are composed of switching Importing Port Set (*IPS*), Cell Demultiplexor Set (*CDS*), Switching Plane Set (*SPS*), Cell Multiplexor Set (*CMS*) and switching Exporting Port Set (*EPS*), while the four categories of connecting link sets include the *IP* and *CD* Set (*ICS*), the *CD* and *SP* Set (*CSS*), the *SP* and *CM* Set (*SCS*) and the *CM* and *EP* Set (*CES*).

However, the composition methods of functional unit sets are quite different: IPS set is composed of N importing ports (*IP*), namely $IPS = \{IP_1, IP_2,\ldots, IP_N\}$, corresponding to external line interfaces, which complete cell slicing and buffering. *CDS* sets are composed of W cell grouping units (called *CD*), namely $CDS = \{CD_1, CD_2,\ldots, CD_W\}$, and they realize cell distribution and allocation among parallel packet switch planes (*SP*). *SPS* is composed of P *SP*s, namely $SPS = \{SP_1, SP_2,\ldots, SP_P\}$, and each *SP* is constituted by multi-stage switching networks, whose internal properties were studied in our previous work [9]. *CMS* is composed of Q *CM*s, namely $CMS = \{CM_1, CM_2,\ldots, CM_Q\}$, mainly realizes cell aggregation and combination. *EPS* is composed of N *EP*s, namely $EPS = \{EP_1, EP_2,\ldots, EP_N\}$, connecting the output links of line interface cards.

The interconnect structure of link sets are also different: *ICS* is composed of $(N \times W)$ *IC*s, namely $ICS = \{IC(k,l)|k = 1,2,\ldots,N; l = 1,2,\ldots,W\}$. *CSS* is composed of $(W \times P)$ *CS*s, namely $CSS = \{CS(k,l) \mid k = 1,2,\ldots, W; l = 1,2,\ldots, P\}$. *SCS* is composed of $(P \times Q)$ *SC*s, namely $SCS = \{SC(k,l)|k = 1,2,\ldots,P; l = 1,2,\ldots,Q\}$. *CES* is composed of $(Q \times N)$ *CE*s, namely $CES = \{CE(k,l)|k = 1,2,\ldots,Q; l = 1,2,\ldots,N\}$.

The MPMS functional unit sets are abstracted as domain, and the connecting link sets as mappings, and we get logical mapping graph (LMG) as in Fig. 103.2. Thus *IPS* can be abstracted as: $IPS \leftrightarrow IPD$, and *CDS* can be abstracted as:

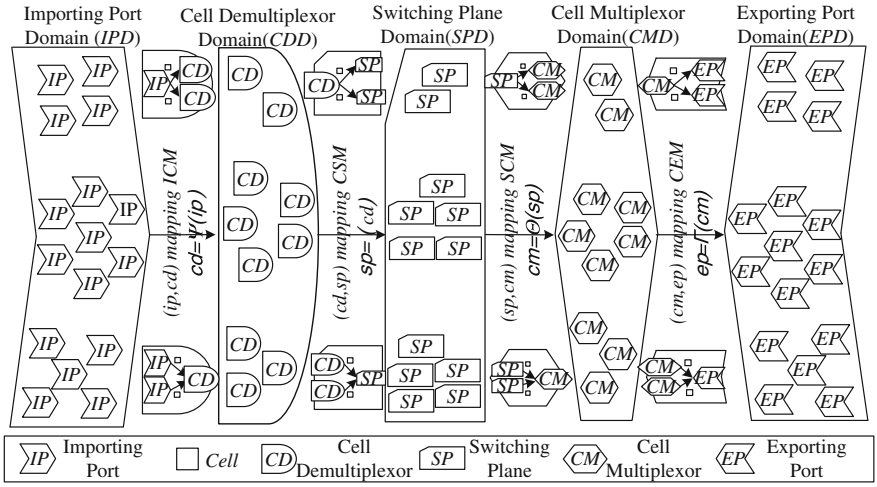


Fig. 103.2 Logical mapping graph of MPMS packet switching fabric

CDS↔*CDD*, and *SPS* can be abstracted as: *SPS*↔*SPD*, and *CMS* can be abstracted as: *CMS*↔*CMD*, and *EPS* can be abstracted as: *EPS*↔*EPD*. In the same way, *ICS* can be abstracted as: $cd = \Psi(ip)$, and *CSS* can be abstracted as: $sp = \Omega(cd)$, and *SCS* can be abstracted as: $cm = \Theta(sp)$, and *CES* can be abstracted as: $ep = \Gamma(cm)$.

103.3 Graph Model and Probability Model of MPMS

As is shown in Figs. 103.1 and 103.2, cells are sequentially switched from input ports IP, passing all corresponding functional units, and eventually reaching output ports EP, which means the MPMS switching fabric has the characteristics of directions. According to directed graph theory [8, 14], the functional domains in LMG logical mapping graph can be abstracted as the zenith sets and the LMG mappings in MPMS logical mapping graph can be abstracted as the directed border in the theory of directed graph, we get the DGM model of the MPMS fabric shown in Fig. 103.3.

Thus, according to the directed graph model and logical mapping relationship in MPMS fabric, the mathematical model of DGM directed graph model can be described as Eqs. (103.1) and (103.2) as follows:

$$Z \triangleq \bigcup_{i=1}^{5} Z^i = Z^1 \cup Z^2 \cup Z^3 \cup Z^4 \cup Z^5,\ \bar{B} \triangleq \bigcup_{j=1}^{4} \bar{B}^j = \bar{B}^1 \cup \bar{B}^2 \cup \bar{B}^3 \cup \bar{B}^4 \quad (103.1)$$

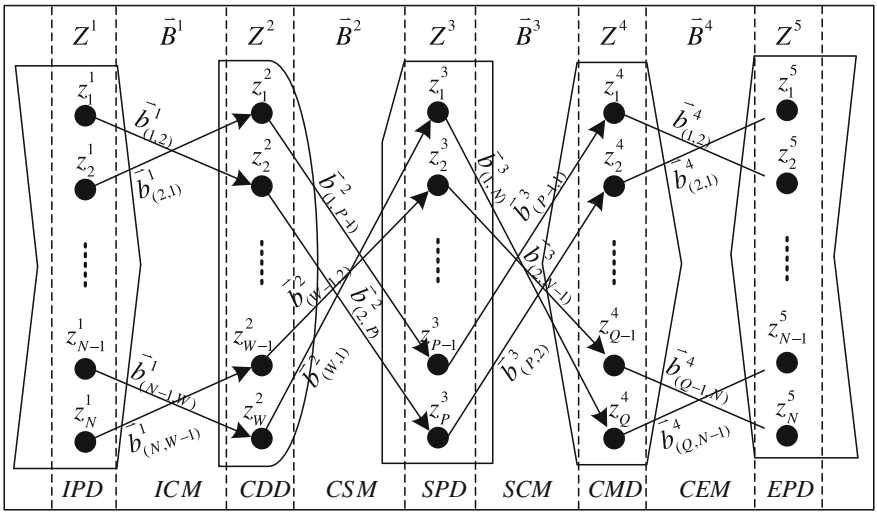


Fig. 103.3 Directed graph model of MPMS packet switching fabric

$$Z^i \triangleq \begin{cases} \{z_1^i, z_2^i, \ldots, z_{N-1}^i, z_N^i\}, i = 1,\ 5 \\ \{z_1^i, z_2^i, \ldots, z_{W-1}^i, z_W^i\}, i = 2 \\ \{z_1^i, v_2^i, \ldots, z_{P-1}^i, z_P^i\}, i = 3 \\ \{z_1^i, z_2^i, \ldots, z_{W-1}^i, z_Q^i\}, i = 4 \end{cases},$$

$$\vec{B}^i \triangleq \begin{cases} \{\vec{b}_{(k,l)}^i = \vec{x}\ y | x = z_k^i, y = z_l^{i+1};\ k = 1, \ldots, N;\ l = 1, \ldots, W\},\ i = 1 \\ \{\vec{b}_{(k,l)}^i = \vec{x}\ y | x = z_k^i, y = z_l^{i+1};\ k = 1, \ldots, W;\ l = 1, \ldots, P\},\ i = 2 \\ \{\vec{b}_{(k,l)}^i = \vec{x}\ y | x = z_k^i, y = z_l^{i+1};\ k = 1, \ldots, P;\ l = 1, \ldots, Q\},\ i = 3 \\ \{\vec{b}_{(k,l)}^i = \vec{x}\ y | x = z_k^i, y = z_l^{i+1};\ k = 1, \ldots, Q;\ l = 1, \ldots, N\},\ i = 4 \end{cases} \tag{103.2}$$

According to the definition Eqs. (103.1) and (103.2) in the directed graph model in MPMS fabric, we can get the logical mapping definition equations of the MPMS packet switching fabric as equations expressed in Eq. (103.3):

$$Z^{i+1} \triangleq \begin{cases} \Psi(Z^i), i = 1 \\ \Omega(Z^i), i = 2 \\ \Theta(Z^i), i = 3 \\ \Gamma(Z^i), i = 4 \end{cases}, \begin{bmatrix} Z^2 \\ Z^3 \\ Z^4 \\ Z^5 \end{bmatrix} \triangleq \begin{bmatrix} \Psi(Z^1) \\ \Omega(Z^2) \\ \Theta(Z^3) \\ \Gamma(Z^4) \end{bmatrix} \triangleq \begin{bmatrix} \Psi(\cdot) \\ \Omega(\cdot) \\ \Theta(\cdot) \\ \Gamma(\cdot) \end{bmatrix} \times \begin{bmatrix} Z^1 \\ Z^2 \\ Z^3 \\ Z^4 \end{bmatrix}, \tag{103.3}$$

$\forall Z^i \in Z, i = 1, 2, \ldots, 5$

Equation (103.3) defines the logical mapping relationships of the adjacent zeniths in MPMS packet switching fabric, while the mapping and connecting relationships among adjacent zeniths is always random for the actual packet switching process in MPMS fabric. That is the probability model describing randomness, called the MPMS probability model. According to definition Eqs. (103.1) and (103.2), we get the MPMS probability model illustrated in Eqs. (103.4) and (103.6) as follows:

$$\left[\left[P^{(i,\,i+1)}\right]\right] \triangleq \begin{cases} \left[\left[P^{(1,\,2)}\right]\right] = (p^{(1,2)}_{(k,l)})_{N\times W};\ k=1,2,\ldots,N;\ l=1,2,\ldots,W;\ i=1 \\ \left[\left[P^{(2,\,3)}\right]\right] = (p^{(2,3)}_{(k,l)})_{W\times P};\ k=1,2,\ldots,W;\ l=1,2,\ldots,P;\ i=2 \\ \left[\left[P^{(3,\,4)}\right]\right] = (p^{(3,4)}_{(k,l)})_{P\times Q};\ k=1,2,\ldots,P;\ l=1,2,\ldots,Q;\ i=3 \\ \left[\left[P^{(4,\,5)}\right]\right] = (p^{(4,\,5)}_{(k,\,l)})_{Q\times N};\ k=1,2,\ldots,Q;\ l=1,2,\ldots,N;\ i=4 \end{cases} \tag{103.4}$$

$$\left[\left[P^{(1,2)}\right]\right] \triangleq \begin{bmatrix} p^{(1,2)}_{(1,1)} & p^{(1,2)}_{(1,2)} & \cdots & p^{(1,2)}_{(1,W)} \\ p^{(1,2)}_{(2,1)} & p^{(1,2)}_{(2,2)} & \cdots & p^{(1,2)}_{(2,W)} \\ \vdots & \vdots & \vdots & \vdots \\ p^{(1,2)}_{(N,1)} & p^{(1,2)}_{(N,2)} & \cdots & p^{(1,2)}_{(N,W)} \end{bmatrix} \hbar \left[\left[P^{(2,3)}\right]\right] \triangleq \begin{bmatrix} p^{(2,3)}_{(1,1)} & p^{(2,3)}_{(1,2)} & \cdots & p^{(2,3)}_{(1,P)} \\ p^{(2,3)}_{(2,1)} & p^{(2,3)}_{(2,2)} & \cdots & p^{(2,3)}_{(2,P)} \\ \vdots & \vdots & \vdots & \vdots \\ p^{(2,3)}_{(W,1)} & p^{(2,3)}_{(W,2)} & \cdots & p^{(2,3)}_{(W,P)} \end{bmatrix} \tag{103.5}$$

$$\left[\left[P^{(3,4)}\right]\right] \triangleq \begin{bmatrix} p^{(3,4)}_{(1,1)} & p^{(3,4)}_{(1,2)} & \cdots & p^{(3,4)}_{(1,Q)} \\ p^{(3,4)}_{(2,1)} & p^{(3,4)}_{(2,2)} & \cdots & p^{(3,4)}_{(2,Q)} \\ \vdots & \vdots & \vdots & \vdots \\ p^{(3,4)}_{(P,1)} & p^{(3,4)}_{(P,2)} & \cdots & p^{(3,4)}_{(P,Q)} \end{bmatrix}, \left[\left[P^{(4,5)}\right]\right] \triangleq \begin{bmatrix} p^{(4,5)}_{(1,1)} & p^{(4,5)}_{(1,2)} & \cdots & p^{(4,5)}_{(1,N)} \\ p^{(4,5)}_{(2,1)} & p^{(4,5)}_{(2,2)} & \cdots & p^{(4,5)}_{(2,N)} \\ \vdots & \vdots & \vdots & \vdots \\ p^{(4,5)}_{(Q,1)} & p^{(4,5)}_{(Q,2)} & \cdots & p^{(4,5)}_{(Q,N)} \end{bmatrix} \tag{103.6}$$

According to above results, we can see the corresponding relationship between functional units and connection links as follows: domain functional unit↔functional domain↔zenith set↔probability matrix zenith set, and connecting link↔mapping functions↔directed border sets↔ probability vectors.

103.4 Experiments and Simulation Analysis

The probability calculation methods are illustrated as follows:

1 To calculate matching probability $p^{(1,\,2)}_{(k,\,l)}$ between any zenith $z^1_k(k=1,\ldots,N)$ in zenith set of Z^1 and any zenith $z^2_l(l=1,\ldots,W)$ in adjacent zenith set of Z^2, there exist two cases: (1) When the topology parameters of the MPMS packet

switching system satisfies $W \geq N$, for any zenith of $z_k^1(k = 1, \ldots, N)$, there always exists a certain zenith $z_l^2(l = 1, \ldots, W)$ matching with the zenith $z_k^1(k = 1, \ldots, N)$, and therefore $p_{(k,\,l)}^{(1,\,2)}$ is equal to the reciprocal $(P(W,\ N))^{-1}$of the arrangement $P(W,\ N)$. (2) When the topology parameters of the MPMS packet switching system satisfies $W < N$, for any zenith $z_k^1(k = 1, \ldots, N)$, there are at most W zeniths matching successfully in each matching, and therefore $p_{(k,l)}^{(1,2)}$ is equal to $W/(NW!)$, which is the reciprocal of the product of the factorial of topological parameters of W with the magnification of N/W. To sum up, the calculating method of matching probability for any zenith of $z_k^1(k = 1, \ldots, N)$ in zenith set of $Z^1$and any zenith of $z_l^2(l = 1, \ldots, W)$ in adjacent zenith set of Z^2 is illustrated in Eq. (103.7) as follows:

$$p_{(k,\,l)}^{(1,\,2)} = \begin{cases} \frac{(W-N)!}{W!} = (P(W,\ N))^{-1}, & W \geq N \\ \frac{1}{W(W-1)(W-2)\ldots 2\times 1} \times \frac{W}{N} = \frac{W}{NW!}, & W < N \end{cases} \tag{103.7}$$

In similar way, we get $p_{(k,\,l)}^{(4,\,5)}$ between any zenith $z_k^4(k = 1, \ldots, Q)$ in zenith set Z^4 and any zenith $z_l^5(l = 1, \ldots, N)$ in adjacent zenith set of Z^5 in Eq. (103.8) as follows:

$$p_{(k,\,l)}^{(4,\,5)} = \begin{cases} \frac{(N-Q)!}{N!} = (P(N,Q))^{-1}, & N \geq Q \\ \frac{1}{N(N-1)(N-2)\ldots 2\times 1} \times \frac{N}{Q} = \frac{N}{QN!}, & N < Q \end{cases} \tag{103.8}$$

2 To calculate matching probability $p_{(k,\,l)}^{(2,\,3)}$ between any zenith $z_k^2(k = 1, \ldots, W)$ in zenith set Z^2 and any zenith $z_l^3(l = 1, \ldots, P)$ in adjacent zenith set Z^3, because for any zenith of $z_k^2(k = 1, \ldots, W)$ there always exists a certain zenith of $z_l^3(l = 1, \ldots, P)$ that is mathed with the zenith of $z_k^2(k = 1, \ldots, W)$, therefore $p_{(k,\,l)}^{(2,\,3)}$ is equal to the reciprocal of P^{-1} topological parameter of P, as illustrated in Eq. (103.9) as follows:

$$p_{(k,\,l)}^{(2,\,3)} = \frac{1}{P} = P^{-1} \tag{103.9}$$

In similar way, we get $p_{(k,\,l)}^{(3,\,4)}$ between any zenith $z_k^3(k = 1, \ldots, P)$ in zenith set of Z^3 and any zenith $z_l^4(l = 1, \ldots, Q)$ in adjacent zenith set of Z^4, as illustrated in Eq. (103.10) as follows:

$$p_{(k,\,l)}^{(3,\,4)} = \frac{1}{Q} = Q^{-1} \tag{103.10}$$

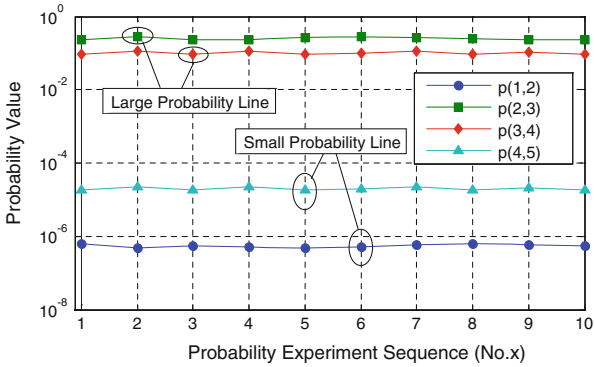


Fig. 103.4 Probability experiments under *MPMS* ($N = 8$, $W = 10$, $P = 4$, $Q = 10$)

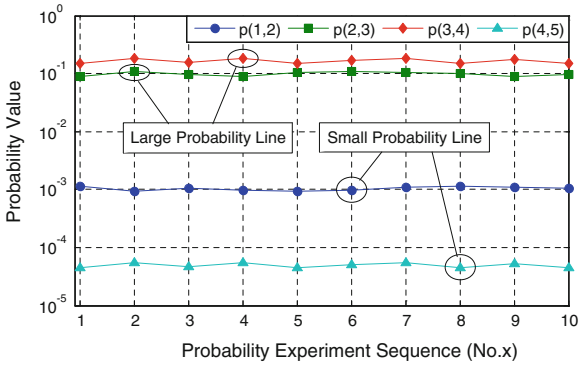


Fig. 103.5 Probability experiments under *MPMS* ($N = 8$, $W = 6$, $P = 10$, $Q = 6$)

In our experiment, the topological parameters are chosen in two methods: one method is to choose the number of switching planes larger than that of demultiplexor and multiplexor, and the other is to choose less switching planes. The experiment for the first method is conducted with MPMS parameter of *MPMS* ($N = 8$, $W = 10$, $P = 4, Q = 10$), and the probability results are shown in Fig. 103.4. The experiment of the second method is conducted with MPMS parameter of *MPMS* ($N = 8, W = 6$, $P = 10$, $Q = 6$), and the probability results are shown in Fig. 103.5. From the results shown in Figs. 103.4 and 103.5, the probabilities of $p(1, 2)$, $p(2, 3)$, $p(3, 4)$, $p(4, 5)$ are quite stable with an deviation less than 2 %, much better than present results [5, 8]. The results of our experiment are quite consistent with our probability model of the MPMS packet switching fabric, which proves the correctness of our probability model of the MPMS packet switching fabric.

103.5 Conclusion

The multiple-plane and multiple-stage packet switching fabric (MPMS) is a novel and promising large-capacity scalable switching technology on the Internet and future information network. In this paper, regarding to the current researching stage

of lack of theoretical basis and mathematical model, researchers proposed a general switching topological map and logical topology map (LMG) in MPMS packet switching system, which laid the theoretical basis and mathematical model for the research of topology and logical map of MPMS switching network. In addition, this paper presents directed graph model (DGM) and probability model for the MPMS packet switching fabric, which laid a preliminary theoretical basis for the further study of probability analysis of the MPMS packet switching network. In simulation experiments, the probabilities are quite stable in the MPMS probability model with a deviation less than 2 %, which is much better than present results.

Acknowledgments This work was supported in part by the grants from Natural Science Foundation of China with No. 61003252. We thank the anonymous reviewers for their constructive comments and suggestions.

References

1. Ye, X.H., Paul, M.J., Yin, Y.W., Roberto, P., Yoo, S.J.B., Akella, V.: DOS: A scalable optical switch for datacenters. In: ACM/IEEE Symposium on Architectures for Networking and Communications Systems, pp. 1–12. La Jolla (2010)
2. Jin, C.Y., Kojima, O., Inoue, T., Kita, T., Wada, O., Hopkinson, M., Akahane, K.: Detailed design and characterization of all-optical switches based on InAs/GaAs quantum dots in a vertical cavity. IEEE J. Quantum Electron. **46**(11), 1582–1589 (2010)
3. Wang, F., Zhu, W., Hamdi, M.: The central-stage buffered clos-network to emulate an OQ switch. In: Proceedings of IEEE Globe Telecommunications, pp. 1–5. Francisco (2006)
4. Wang, F., Hamdi. M.: Analysis on the central-stage buffered clos-network for packet switching. In: IEEE ICC Proceeding, pp.322–335. Korea (2005)
5. Wang, F., Hamdi, M.: Scalable central-stage buffered clos network packet switches with QoS. IEEE HPSR Workshop, pp. 455–468. Poland (2006)
6. Chao, H.J., Park, J.S., Artan, S., Jiang, S., Zhang, G.: Trueway: a highly scalable multi-plane multi-stage buffered packet switch. In: Proceeding of IEEE Workshop on High Performance Switching and Routing, pp. 246–253 Hong Kong (2005)
7. Chao H.J., Jinsoo, P.: Flow control in a multi-plane multi-stage buffered packet switch. IEEE International Conference on HPSR, pp. 1–6. June (2007)
8. Ma Xiang-jie, Li Xiu-Qin, Lan Ju-long, Graphic model and performance analysis of a novel scalable multiple-plane and multiple-stage packet switching fabric, Journal of Electronics & Information Technology(China), 31(5):1026-1030(2009)
9. Ma, X., He, L., Lan, J., Zhang, B.: Study on a novel scheduling algorithm of the multiple-plane and multiple-stage switching fabric. In: The 10th IEEE International Conference on High Performance Computing and Communications, pp. 412–417 (2008)
10. Ma, X., Gu, X., Lan, J., Zhang, B.,. Performance study on the MPMS fabric: a novel parallel and distributed switching system architecture. In: The 10th IEEE International Conference on High Performance Computing and Communications, Dalian, pp. 69–76. China, Sept. (2008)
11. Rojas-Cessa, R., Oki, E., Chao, H.J.: Concurrent fault detection for a multiple-plane packet switch. IEEE/ACM Trans. Networking **11**(4), 616–627 (2003)
12. Khotimsky, D.A., Krishnan, S.: Stability analysis of a parallel packet switch with buffer less input demultiplexors. IEEE ICC, pp. 850–862. Helsinki (2001)
13. Aslam, A., Christensen, K.: Parallel packet switching using multiplexors with virtual input queues. In: Proceedings of IEEE LCN, pp .270–277 (2002)
14. Wilson, R.J.: Introduction to Graph Theory. Academic Press, NewYork (1972)

Chapter 104
Virtual Reality Enhanced Rehabilitation Training Robot for Early Spinal Cord Injury

Yanzhao Chen, Yiqi Zhou, Xiangli Cheng and Zheng Wang

Abstract In order to compensate for the scarcity of currently available early rehabilitation training means of spinal cord injury (SCI) patients, a method of early spinal cord injury rehabilitation training based on virtual reality and robot is pointed out through analysis of nervous system plasticity of patients. A rehabilitation robot system with eight degrees of freedom (DOF) is established, which is based on a six DOF parallel platform. A virtual reality training scene is built. The hardware and software environment of rehabilitation training is studied, meanwhile the match among the virtual reality scene, robot and muscle training in patients is completed. And then, the relevant training mode is formulated. Finally, the implementation of rehabilitation training program is designed. As a result, a new rehabilitation training method for early patients with SCI is formed.

Keywords Robot · Virtual reality · Spinal cord injury · Rehabilitation training

104.1 Introduction

Spinal cord injury has a high incidence and most patients lose the ability of daily life (ADL). The improvement of the SCI patient's ADL by early rehabilitation treatment is meaningful to reduce the burden of family and society. At present, the rehabilitation training is considered to be a safe treatment to promote the neural

Y. Chen · Y. Zhou (✉) · X. Cheng
School of Mechanical Engineering, Shandong University, Shandong Jinan, China
e-mail: yqzhou@sdu.edu.cn

Y. Chen
e-mail: chyzh1986@163.com

Z. Wang
Institute of Automation, Chinese Academy of Sciences, Beijing, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_104,

plasticity after SCI, and is commonly used [1]. Early rehabilitation training in promoting the regeneration and repair of the spinal cord circuitry is particularly important [2].

The existing training methods by the rehabilitator also have some shortages. Firstly, they depend on rehabilitator, that the training time, strength and accuracy cannot be guaranteed. Secondly, the training process is always repeated, which is usually boring. Therefore, there is an urgent need for new means of training to promote early rehabilitation of patients with SCI. Robot-supported rehabilitation training [3, 4] can provide precise and long-lasting controls and provide timely feedback. Meanwhile, VR supported rehabilitation training can stimulate patients' motive, providing immediate feedback, which is important on rehabilitation of the nervous system [5], thus, becomes widely used in rehabilitation training [6].

There are several researches on VR enhanced rehabilitation training robot at home and abroad such as Nam Gyun Kim.etc. researched and developed a VR robot rehabilitation system of trunk balance [7], M. GIRONE.etc. developed an ankle rehabilitation robot with VR system based on Stewart platform called "Rutgers Ankle" [3]. Jingyuan Huang of Tsinghua University studied and developed a lower extremity movement rehabilitation system by virtual bicycle riding [8]. These researches have made some progress in theory and practice, however, the DOF of the existed robot is limited for patient training, and there are few reports of VR based robot study for SCI patients' early rehabilitation.

In order to repair the damaged neural pathways of SCI patients after spinal shock and rebuild their motor function by muscle training, the paper studied and implemented rehabilitation training robot hardware and software environment by six-DOF parallel mechanism and VR technology, then, built communication and coordination mechanism among muscles of patients, robot and VR scene.

104.2 System Framework

104.2.1 Nervous System Plasticity After SCI

In cases of spinal trauma, the nerve fiber bundles are especially devastating, it results in pronounced and persistent sensorimotor dysfunctions for all body parts below the lesion site. In case of spinal cord lesions, especially in incomplete case, the nervous system will be spontaneous adaptation, and has potential to spontaneous functional recovery, which depends on the retain substance of nerve. Self-healing ability of the central nervous system, known as plasticity [1] has great significance to recovery of motor function. Long-term repeated muscle training of patients will stimulate muscle activity. The sensory afferent feedback transmitted by muscle continuously stimulates the spinal cord and forms afferent reflex [9], then, restructuring of the brain and spinal cord circuits, which contributes to recovery of walking after training.

104.2.2 VR and Robot Based Rehabilitation Training System

Figure 104.1 shows the system framework we designed for rehabilitation training. It contains two key parts, which are robot system and PC system.

Robot system consists of the robot body and its control system. The robot control system is responsible for translating and transmitting the control information from PC system to robot body for execution. The robot body provides physical support to patients, as the implementing mechanism, to realize the movement.

PC system consists of several servers and PCs. The servers are responsible for scheduling and coordination of the devices throughout the environment, and act as the carrier of a variety of services. PC for physician as a terminal equipment provides interfaces between physician and the rehabilitation training system, in which the rehabilitation software runs. Physician control the rehabilitation training of patient by the rehabilitation software. PC for patient as another terminal equipment provides the interface between patient and the entire rehabilitation training system, on which the VR scene deployed, providing rehabilitation training support for patients.

In the process of implementation of the rehabilitation training follows certain rehabilitation training mode, physician monitors and controls the whole process through rehabilitation training software running on PC for physician. The control instructions of the physician is sent by PC to the server, on the one hand, the server transmits control instruction to the robot system, after data processing by robot control system the robot execute the movement; on the other hand, the server communicates with the VR scene. The patient does training exercise with the robot body driven and interacted with the visual scene by PC for patient.

104.2.3 Rehabilitation Training Method

Spinal nerve concludes cervical (C), thoracic (T), lumbar (L) and sacral segment (S), different injury segments will affect different regions of the body. Our method is to stimulate the central nervous system to repair of pathway by muscle training

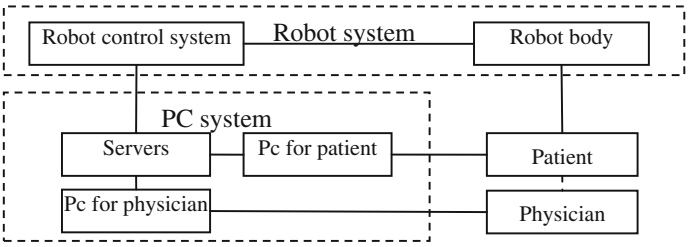


Fig. 104.1 Framework of rehabilitation training system

with the robot in VR environment according to the injury site. We established the context with robot and VR technology and built the communication and coordination mechanism among muscle of patients, robot and VR scene. Early SCI patients are mostly unable to stand. We choose trunk muscles such as rectus abdominis and erector spinae, as well as lower limb muscle such as rectus femoris and tibialis anterior as the target we will train.

Authors have designed two rehabilitation training modes. Firstly, set scheduled trajectory and intensity for training without the detailed control by patients. Its aim is to stimulate targeted muscle in single point and multi-points repeatedly. Single point of stimulation by a single DOF of the robot platform movement can be fixed to train some muscles, and other muscles may be trained less, while the multi-point stimulation through the combined movement of robot may play a better effect. Therefore, it is an aspect in this mode.

In another mode, the patient manipulates the visual reality scene to drive the robot platform, which increases the randomness of movement, and overcomes the monotonicity that may result in "learned disuse" [10] in first mode, meanwhile, enhances the participation of patients.

104.3 Robot System

104.3.1 Robot Body

The robot body of rehabilitation training robot as Fig. 104.2 described can carry out eight DOF of movement, which consists of two platform, one is trunk support platform, and the other is lower limb motor-assisted platform, by which eight DOF of movement can be achieved to meet the needs of multi-modal rehabilitation training. The former is designed based on the six degrees of freedom parallel platform with the features of compact, non-polluting, and high accuracy. It can do six single DOF of movement, including yaw, pitch, roll, level, horizontal, vertical, as well as multiple degrees of freedom of movement, and a seat is mounted on the moving platform, providing support to the patient's body weight. The latter consists of two additional degrees of freedom series connecting to the main platform, which can achieve two degrees of freedom movements, which are knee and ankle rotation.

104.3.2 Robot Hardware Control System

The entire rehabilitation robot hardware control system (shown in Fig. 104.3) is responsible for the communication between upper computer and lower computer, which consists of control system hardware. It includes industrial computer system,

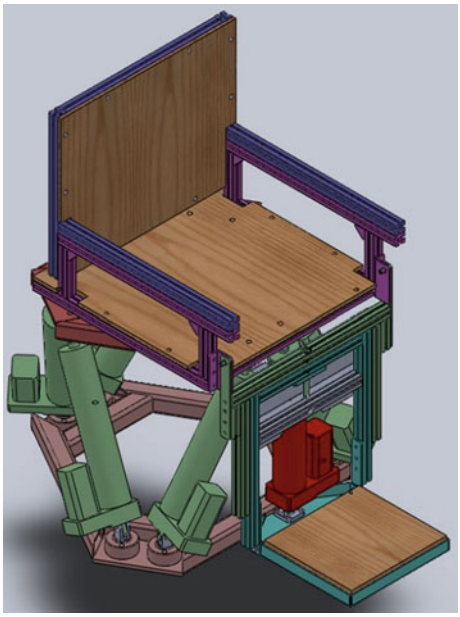


Fig. 104.2 Rehabilitation training robot body

servo driver, servo motor, some control buttons, status display interface and control software.

When this system is running, the control information of upper computer is transmitted to servo drivers by industrial computer system to drive the servo motors. Meanwhile, the motion state data of electric cylinders is collected by sensors attached in it, then, transmitted to industrial computer system, and eventually back to the control software in lower computer or software in the upper PC systems after process.

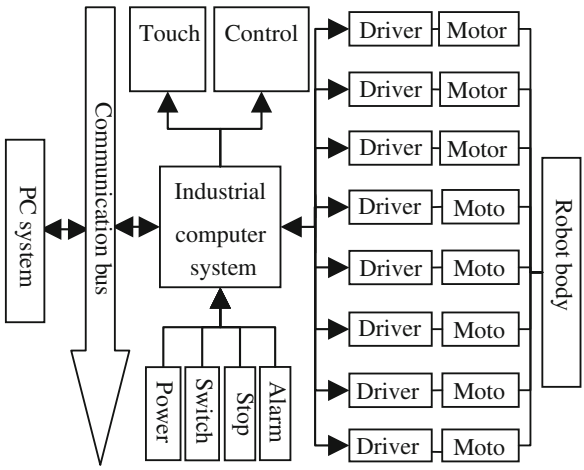


Fig. 104.3 Robot hardware control system

104.3.3 Robot Movement with Muscles Training

The robot designed above can achieve eight degrees of freedom movement, and we establish the mapping from platform movement to muscle training according to the relationship between muscle and movement as follows:

- Pitching motion: mainly training rectus abdominis;
- Roll motion: mainly training erector spinae;
- Rotation of the knee: mainly training rectus femoris;
- Rotation of the ankle: mainly training tibialis anterior.

The mapping above provides support for the rehabilitation training. Besides these degrees of motion mentioned above, some motions may be limited to simulate certain muscle, such as yaw, however, they are also essential in maintaining the continuity and integrity of action in the VR based rehabilitation training environment.

104.4 PC System

PC system in this whole system works as upper computer system and is responsible for the interaction between physician or patient and the robot system. In this paper, we mainly discuss the software running on it.

104.4.1 Rehabilitation Training Software

The rehabilitation training software's function is to control the rehabilitation training process. It contains modules of the patient case database, VR scene, training set and training control. VR scene is a key module of rehabilitation training software, so we introduce it in this section.

This VR scene is the secondary development of an open source game based on OpenGL. We chose a game about aircraft to construct VR training environment, which can do multiple actions, such as forward flight, pitch, turning around, accelerating, and decelerating. In addition, it can achieve a variety of effects, such as lighting, environmental smoke and sound effects. Under certain modes of training, after creating motion mapping between the aircraft and the robot platform in the virtual scene, patient place himself in the virtual scene to gain the immersive feeling, in order to enhance the effect of rehabilitation training.

Two interfaces are designed to support the interaction between visual reality scene and robot system under rehabilitation training mode. One is to receive data by visual reality scene for the first mode, and the other is to send data for the second mode.

104.4.2 Visual Reality Scene for Rehabilitation Training

The rehabilitation software communicates with the robot system through TCP/IP protocol and Socket interface. Under the first rehabilitation training mode, the course and trajectory of the movement is pre-defined in the software, and the rehabilitation software sends the control information to the visual reality scene by the first interface, meanwhile sends it to the robot system, and then realizes the cooperative motion. Under the second rehabilitation training mode, the visual reality scene sends the position and orientation data to the rehabilitation software by interface two. After the data being processed, motion control of the robot is completed.

We designed actions in visual reality and robot environment to training specific muscles, some of them are shown as follows:

- Pitching motion of the robot that corresponding to dive and climb of the aircraft in the virtual scene, which mainly trains rectus abdominis;
- Roll movement of the robot that corresponding to turning around of the aircraft in the virtual scene, which mainly trains erector spinae;
- Knee joint rotation of the robot that corresponding to the flying height of the aircraft in the virtual scene, which mainly trains rectus femoris;
- Ankle joint rotation of the robot which corresponds to the flying speed of the aircraft in the virtual scene mainly trains tibialis anterior.

104.5 Process of Rehabilitation Training

During implementation of the rehabilitation training in the rehabilitation training environment of SCI designed above, patients are sitting on the robot seat and facing to the screen, in which the VR game is displaying. Physician initializes related equipment, connects the hardware and software, selects and starts the VR scene, and then selects the rehabilitation model based on the patient's condition. The patient interacts with the virtual scene by PC for patient, and the physician monitors the training process in real-time, and makes timely adjustments when need by PC for physician. After completion of a training cycle, training information of the patient is recorded into the case database for analysis and sharing. Physician evaluates the training according to the results, if it meets certain requirements, then the training is finish. Otherwise, the physician adjusts the training program, and continues the training of the next cycle. The process usually repeats several times until it meets the rehabilitation requirements.

104.6 Conclusion

The incidence rate of SCI is high, patients' normal life is seriously affected, and the treatment is expensive. Human and animal experiments show that patients with SCI, especially in incomplete case, have multiple levels of plasticity in central nervous system that rehabilitation training can contribute to this plasticity. In this paper, after analysis of the rehabilitation training principle, researchers designed a SCI rehabilitation robot hardware and software environment based on a six DOF parallel platform and VR technology. They also formulated human muscle rehabilitation training mode, which provides a powerful tool and means of rehabilitation training for early incomplete SCI patients. The establishment of the evaluation mechanism about the rehabilitation effect will be the further research work.

Acknowledgments Authors will acknowledge the reviewers to the paper for improvement suggestions. At the same time, they will be thankful to National Science Foundation of China (No. 61103153/F020503), 863 plan (No. SS2013AA010903) and the Key Science and Technology Program of Shandong Province (No. 2010G0020233) in carrying out this research for financial support.

References

1. Fouad, K., Tetzlaff, W.: Rehabilitative training and plasticity following spinal cord injury. Exp. Neurol. **235**(1), 91–99 (2012)
2. Shields, R.K., Dudley-Javoroski, S., et al.: Low-frequency H-reflex depression in trained human soleus after spinal cord injury. Neurosci. Lett. **499**(2), 88–92 (2011)
3. Girone, M., Burdea, G., et al.: A Stewart platform-based system for ankle telerehabilitation. Auton. Robots **10**(2), 203–212 (2001)
4. Van de Meent, H., Baken, B.C.M., et al.: Critical illness VR rehabilitation device (X-VR-D): Evaluation of the potential use for early clinical rehabilitation. J. Electromyogr. Kinesiol. **18**(3), 480–486 (2008)
5. Riener, R., Wellner, M., et al.: A View on VR-Enhanced Rehabilitation Robotics. In: IWVR (2006)
6. Holden, M.K.: Virtual environments for motor rehabilitation: review. Cyber Psychol. Behav. **8**(3), 187–211 (2005)
7. Kim, N.G., Yoo, C. K., et al.: A new rehabilitation training system for postural balance control using virtual reality technology. IEEE Trans. Rehabil Eng. **7**(4), 482–485 (1999)
8. Huang, J., Liu, H., et al.: The Prospects of Research on VR Rehabilitation Engineering. J. Biomed. Eng. **16**(2), 203–208 (1999)
9. Knikou, M.: Neural control of locomotion and training-induced plasticity after spinal and cerebral lesions. Clin. Neurophysiol. **121**(10), 1655–1668 (2010)
10. Raineteau, O., Schwab, M.E.: Plasticity of motor systems after incomplete spinal cord injury. Nat. Rev. Neurosci. **2**(4), 263–273 (2001)

Chapter 105
Syntactic Rules of Spatial Relations in Natural Language

Shaonan Zhu and Xueying Zhang

Abstract Spatial relations are the main part of geographical information in natural language. Their extraction and semantic interpretation play a significant role in bridging the gap between geographical information system and natural language. Normally spatial relations are described with certain spatial terms and syntactic rules in natural language. To overcome the disadvantage of manual induction of syntactic rules, this paper proposes a new machine learning approach based on a sequence alignment algorithm. Firstly, the description instances of spatial relations in a large-scale annotated corpus are extracted and analyzed, and the sequence alignment algorithm is used to calculate the pattern similarity between instances of spatial relations. Then, the instances with high similarity are generalized aspopularly used syntactic rules. Finally, these rules are used for extraction spatial relations in a test data to evaluate their validation. The experimental results indicate that the generalized rules can achieve better performance than those rules induced according to occurrence frequency in the corpus.

Keywords GIS · Spatial relation · Sequence alignment · Syntactic pattern

105.1 Introduction

Spatial relations described with natural language are much more easily understood than geographical information system, because human are used to representing geographic space using natural language [1, 2]. Therefore, to understand and analyze how people express spatial relations is a meaningful issue in geographical information science [3]. National Geographic Information and Analysis Center

S. Zhu (✉) · X. Zhang
Key Laboratory of Virtual Geography Environment Ministry of Education,
Nanjing Normal University, Nanjing, China
e-mail: zhushaonanr@gmail.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_105,

(NCGIA) has implemented a study plan on spatial relations, since 1988. From the perspective of linguistics, the description of spatial relations in natural language is mainly dependent on spatial terms and syntactic structures [4, 5]. Spatial relations terms reflect the state of geographical entities and spatial relations. Syntactic structures delimit the semantic of spatial relations described in the specific pragmatic. Although the spatial relation terms, syntax and semantics have strong fuzziness and uncertainty, there still exist some typical syntactic structures, i.e. Syntactic patterns. Therefore, the identification of syntactic patterns of spatial relations is helpful for the understanding of spatial relations in natural language and their usage in GIS query, spatial reasoning and geographic information retrieval.

The studies aiming at syntactic patterns mainly focus on the field of scene reconstruction from natural language. The Words eye system using part of speech tagging and syntactic analysis, on the basis of the initial annotation and analysis, transforms syntactic parsing tree into a dependency structure, and expresses spatial interdependent relationship between the spatial objects [6]. Reinberger extract the spatial relations with the syntactic structure similar to "subject-verb-object" and special field knowledge [7]. Le used syntactic patterns to extract place names and spatial relations from text [8]. Liu expressed road path by NLRP syntactic patterns [9]. Zhang investigated the spatial relation query with natural language, and proposed analytical method [10]. Zhang used GATE (General Architecture for Text Engineering) as the tool and summarized some syntactic rules to extract spatial relations from the text [11]. In general, rule-based extraction of spatial relations is a significant and reliable method, and the syntactic patterns about spatial relations are manually summarized [12].

Corpus is the resource of linguistics research and information extraction. Therefore, a large-scale annotated corpus will provide a new way to automatically identify of syntactic patterns of spatial relations in text. A rule-based method to extract spatial relations in text will be introduced. Firstly, this paper analyses the instances from the corpus, uses the sequence alignment algorithm to calculate the similarity between instances of spatial relations, builds the similarity matrix, group instances of high similarity, generalized to generate the syntactic patterns of spatial relations; and handles the text to feature sequences; at last, extract spatial relations based on these syntactic patterns.

105.2 Extraction of Spatial Relations Using Syntactic Patterns

In general, a description about the spatial relation is in a sentence or can be divided into sentences, so the context of the description about spatial relation is discussed in a sentence. The part of speech is the most basic language feature in the instance of spatial relations. On the other hand, the vocabulary of spatial relations has a strong indication about spatial relations. GNE (Geographic named entity), POS

(part of speech), and the signal word which is in the vocabulary of spatial relations can express natural language characteristics of the instance of spatial relations. The first step to extract spatial relations is to transform the text to feature sequence which contains three features (GNE, POS, the signal word).Conditional random fields (CRFs) is a method which has be proved as an efficient way to get GNEs from text in the field of natural language processing. Institute of Computing Technology, Chinese Lexical Analysis System (ICTCLAS) is a useful tool to generate POS in Chinese text, and correct rate is more than 97.58 %. So, this paper uses CRFs and ICTCLAS. The signal words are from Corpus which has a large number of spatial relations instances.

A spatial relation can be abstracted as a binary relation between the two geographical entities, has apparent target and reference. In structure, an instance must contain two geographical entities with the order. This paper defines <left, GNE1, middle, GNE2, right> as the structure of the instance. GNE1 and GNE2 mean the target and reference in the instance of spatial relations (GNE: geographic named entity). Left, middle, and right mean the context. Figure 105.1 expresses the structural result of the instances of spatial relations. For example, "ALI Mountain is in the east of Chiayi City" is an instance. After the annotation features are extracted and structured, the final feature sequence is "<GNE><V><SIGNAL><GNE>". If "<GNE><V><SIGNAL><GNE>" is a syntactic pattern, a spatial relation can be found: GNE1 is "ALI Mountain", GNE2 is "JIAYI City", and the relation is "in the east of".

105.3 Syntactic Pattern Recognition

105.3.1 The Similarity Matrix for Instances of Spatial Relations

In order to find the rules from spatial relations, the structure of the sentence which describes spatial relations must be researched. The words in the instances can be regarded as symbols. This paper uses sequence alignment algorithm which is widely applied in the field of biology to compute the similarity. A major theme of genomics is to compare DNA sequences and to find out the public part of the two sequences [13]. Two ways are frequently used. The first one is local sequence alignment, which gets the similarity from local information. Smith-Waterman Algorithm is the classical algorithm. The other one is global sequence alignment, and Needleman-Wunsch Algorithm is the typical algorithm (Fig. 105.2).

Fig. 105.1 The instance of spatial relations

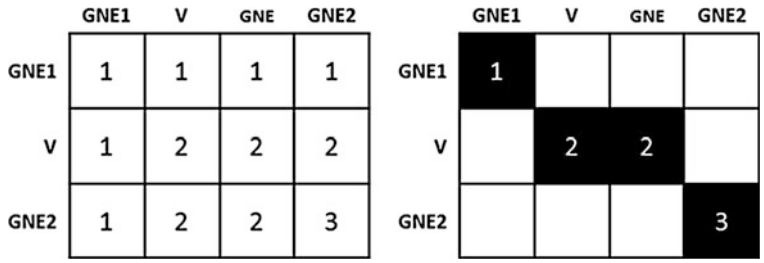


Fig. 105.2 The process of sequence alignment

This paper extends Needleman-Wunsch Algorithm to handle language unit in the sequence. For example, there are two sequences: [GNE1] [V] [GNE] [GNE2] and [GNE1] [V] [GNE2]. Firstly, two-dimensional table must be generated. Secondly, two sequences with the same length can be built (Fig. 105.3).

Thirdly, the similarity between the sequences is quantified (see 105.1).

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{105.1}$$

If at the same position the language unit in the target sequence is the same with the reference sequence, the value of xi is 1, and called one match. The similarity of two sequences means the sum of xi divided by the length of the sequence. The score of the example which is expressed in (Fig. 105.3) is 0.75.

Since the similarity should be computed between two instances, the matrix can be built. The similarity matrix for instances of spatial relations show the similarity of the whole sequence list. It is prepare for the generalization of syntactic patterns.

105.3.2 Generalization of Syntactic Patterns

The sequence can be considered as a syntactic pattern, but it has not a representative. In fact, the instances of spatial relations can be considered as a sample set, then syntactic patterns are generalized according to the value of the similarity matrix. The procedure is defined as follows:

- traverse the instances, and pick up the instance and the list which corresponds the instance in the similarity matrix;
- get the most similar instances, the number of the instances is variable (1,2,…n);

Fig. 105.3 The result of sequence alignment

【GNE1】 【V】 【GNE】 【GNE2】

【GNE1】 【V】 【GNE2】

- generalize the instances from step two. If the result contains language feature, go back to step two. Otherwise, go back to step one;
- loop until all the instances are traversed in step one.

The key point is how to define the generalization templates. When the two sequences are compared, the same kind of feature will be retained. For example, the target sequence is "<GNE1><a/n/n/w/t/v/v/w/t/p><GNE2><n>", and Table 105.1 shows the list which corresponds the instance in the similarity matrix.

The result of the above-mentioned is "<GNE><n/w><GNE><n>", "<GNE><GNE><n>"。

105.4 Experimental Evaluation

This paper used the annotated corpus called GeoCorpus. GeoCorpus got data from encyclopedia of china which contains a full description of the geographic elements of the administrative unit, mountains, rivers, hills, plateaus, plains, basins, etc. This experiment selected 2,355 instances from GeoCorpus. After the instances were structured and generalized, 5,295 syntactic patterns were obtained, and someone repeated several times. Through statistical analysis, there were 996 syntactic patterns except the patterns were repeated. In 5,295 syntactic patterns, there were 920 syntactic patterns which the number of repetitions is less than 10 as Table 105.2 shows, while 50 syntactic patterns had the highest frequency accounted for 70 % of the total.

Fifty syntactic patterns were used to extract spatial relations. GeoCorpus was the test corpus as the data source. The experimental results showed that in Chinese text, spatial relations usually appeared in a sentence; there was almost no qualifier before the former GNE in the syntactic patterns. Some syntactic patterns were the part of other syntactic patterns, and the shorter syntactic patterns were more useful. As Table 105.2 shows, step one used the tradition syntactic patterns by manually summarized, and step two was the result of this paper. The result was that the accuracy was 0.758 and recall rate was 0.433. The accuracy and recalling both increased. These Linguistic phenomena and rules helped to further understand human cognition on spatial relations. Rule-based meant higher accuracy and relatively low recall rate. The right syntactic patterns and rich signal words were the pledge of recognition results (Table 105.3).

Table 105.1 The sample of generalized syntactic patterns

Reference Template	Similarity
<GNE1><m/n/c/w><GNE2><n>	0.71
<GNE1><v/a/n/f/w/c/SIGNAL/w/v/s/w/f/v/w/GNE/w/w><GNE2><n>	0.70
<GNE><SIGNAL/m/n/c/w/w/w><GNE><n>	0.70
<GNE><GNE><n>	0.67
……	

Table 105.2 Syntactic patterns after generalization

Syntactic patterns	Frequency
[GNE1][SIGNAL][GNE2]	622
[GNE1][GNE2][SIGNAL]	403
[GNE1][W][GNE2]	274
[GNE1][GNE][GNE2]	197
[GNE1][V][GNE2]	156
[GNE1][GNE2][N]	135
[GNE1][GNE2][GNE]	120
[GNE1][GNE2][W]	109
[GNE1][W][W][GNE2]	88
[GNE1][SIGNAL][GNE][GNE2]	86

Table 105.3 The comparison of Rule-based extraction

	Precision (%)	Recalling (%)
Step one	75.8	43.3
Step two	78	45.1

105.5 Conclusion

Based on an annotation corpus, this paper proposes a machine learning approach to get the syntactic patterns with the help of the similarity matrix by sequence alignment algorithm. This method overcomes the shortcomings of manual induction, and finds out hidden rules about spatial relations. The future research should focus on how to extract the spatial relations in Chinese text by syntactic patterns, and extend their usage in GIS natural language query.

Acknowledgments This work was supported by National Natural Science Foundation of China (No. 40971231); the Graduates' scientific research and innovation plan of Jiangsu Province (No. CXZZ12_0394).

References

1. Jun, Chen, Renliang, Zhao: Spatial relations in GIS: a survey on its key issues and research progress. Acta Geodaetica et Cartographica Sinica **28**(2), 95–100 (1999)
2. Shihong, Du, Qiming, Qin, Qiao, Wang: The Spatial relations in GIS and their applications. Earth Sci. Front. **13**(3), 069–080 (2006)
3. Du, S., Wang, Q., Li, Z.: Definitions of natural language spatial relations in GIS. Geomatics Inf. Sci. Wuhan Univ. **30**(6), 533–538 (2005)
4. Mark, M.D., Comas, D., Egenhofer, M.J., et al. Evaluating and refining computational models of spatial relations through cross-linguistic human-subjects testing. In: Frank, A., Kuhn, W. Spatial Information Theory: A theoretical Basis for GIS, International Conference COSIT, Semmering, Austria, Lecture Notes in Computer Science, vol. 988, pp. 553–568. Springer-Verlag, Berlin (1995)

5. Egenhofer, M.J.: Locational SQL: syntax extensions. Surveying Engineering Program, University of Maine (1987)
6. Coyne, B., Sproat, R.: Wordseye: an automatic text-to-scene conversion system. In: Proceedings of the 28th Annual Conference on Computer graphics and Interactive Techniques, pp. 487-496. ACM, Los Angeles 12–17 Aug 2001
7. Reinberger, M.-L.: Automatic extraction of spatial relations. In: Proceedings of the TEMA workshop, EPIA 2005, Portugal (2005)
8. Le, X., Yang C., Yu W.: Spatial concept extraction based on spatial semantic role in natural language. Editorial Board Geomatics Inf. Sci. Wuhan Univ. **30**(12), 1100–1103 (2005)
9. Liu, Y., Gao, Y., Lin, B.: Research on GIS path reconstruction based on constrained Chinese language. J. Remote Sens. **8**(4), 323–330 (2004)
10. Zhang, X., Lv, G.: Natural-language spatial relations and their applications in GIS.Geo-information science, **9**(6), 77–81 (2007)
11. Zhang, C., Zhang X., Jiang W.: Rule-based extraction of spatial relations in natural language Text. In: Proceedings of the 2009 International Conference on Computational Intelligence and software Engineering (2009)
12. Chen, X., Hu Y., Lu R.: Extraction of entity relation templates from text collections. Comput. Eng. **33**(22), 199–201; **9**(6), 77–81 (2007)
13. Xu Z.: Bioinformatics. Tsinghua Unversity Press, Beijing (2008)

Chapter 106
An Improved Plagiarism Detection Method: Model and Sample

Jing Fang and Yuanyuan Zhang

Abstract Cosine similarity measure is an efficient plagiarism detection algorithm for documents. However, it may be misled if the document is not properly preprocessed. Furthermore, the weight for the words in the document should depend on its occurrence frequency in the whole digital library. Otherwise, cosine similarity measure may not accurate enough. This paper aims to enhance the accuracy of similarity measure. A preprocessing method and a model to adjust word's weight according to occurrence frequency are proposed in this paper. The paper also develops a sample to illustrate how to preprocess documents, adjust the weight for the words and calculate the similarity. The sample shows that it gets better result after applying the model in this paper.

Keywords Plagiarism detection · Feature vector · Cosine

106.1 Introduction

With the booming development of information technology, most scientific papers and other documents are preserved as digital copies besides hard copies. Thus it becomes much easier to access these documents, and plagiarism happens more frequently because it is easier to copy an existing document than before.

Many plagiarism detection algorithms and tools are developed to find out documents which plagiarize other people's work. A lot of tools compare the target

J. Fang (✉)
Modern Educational Technology Center, North China Institute of Science and Technology, Hebei, China
e-mail: fangj@ncist.edu.cn

Y. Zhang
Library, North China Institute of Science and Technology, Hebei, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_106,

suspicious document with documents in a library which contains millions of documents. If the similarity of the suspicious document and a library document exceeds a preset threshold, this document is assumed to plagiarize other document in a big probability. Another type of algorithms needs no reference library and detects plagiarism inside a single document. Linguistic models are used to find out presentation style difference inside the document and this suggests plagiarism [1]. Some studies also abstract a document to feature vector space and calculate the heterogeneity [2].

For the first kind of method mentioned previously, the key point is how to measure the similarity. The original algorithms pick out plagiarism by string match by comparing the text of two documents [3]. Some tools are based on such kind of algorithm [4]. The limitation is that it can only find out texts which are exactly the same and cannot handle text changes. An improved method is to calculate the word frequency in the document and represent the document with a feature vector, and then calculate the similarity by cosine similarity measure [5]. This method can detect plagiarism when a plagiarizer just messes up the order of the words or sentence, or just deletes or adds some unimportant words. These types of methods can find out plagiarism more accurately. Some practical tools are constructed on this method [6]. Another study topic named versioned document detection is derived from this method to detect versioned document [7].

This paper improves cosine similarity measure by two aspects. The first one is document preprocessing. Preprocessing is needed because a document may be copied by just replacing words of similar meaning or modifying some auxiliary words. After preprocessing, words with similar meaning will be grouped, and auxiliary will be removed from the feature vector. The second improvement is to adjust the weight of words. If word A occurs frequently in the same type of documents of the library, then it does not imply plagiarism even if the word occurs many times in both of two documents. Otherwise, if word B occurs with low frequency in the library and it occurs frequently in both of the two documents, then the probability of plagiarism is much higher. Therefore, in the feature vector, the weight of word A should be decreased and the weight of word B should be increased.

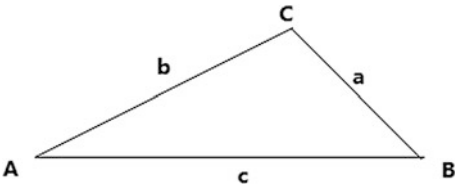
The left parts of this paper are organized as following. Section 106.2 will introduce cosine similarity measure. Section 106.3 will discuss document preprocessing. Section 106.4 will discuss how to adjust word weight. Section 106.5 will give a sample to illustrate the usage. Section 106.6 concludes the work of this paper and future work.

Although the method is universal to all types of documents, this paper mainly focuses on research papers on science and technology.

106.2 Cosine Similarity Measure

For the following triangle, cosine of A can be calculated as formula 106.1. (Fig. 106.1)

Fig. 106.1 Triangle cosine



$$\cos(A) = (b^2 + C^2 - a^2)/(2 * b * c) \tag{106.1}$$

If b and c are understood as vectors, then the previous formula is equal to formula 106.2.

$$\cos(A) = <b, c> /(|b| * |c|) \tag{106.2}$$

In this formula, |b| and |c| is the length of vector b and c, < b, c > is the inner product of vector b and c.

For n-dimensional vectors $< b_1, b_2 \ldots b_n >$ and $< c_1, c_2 \ldots c_n >$, formula 106.2 can be generalized to formula 106.3.

$$\cos(A) = (b_1 * c_1 + b_2 * c_2 + \ldots + b_n * c_n)/(\sqrt{b_1^2 + b_2^2 + \ldots + b_n^2} * \sqrt{c_1^2 + c_2^2 + \ldots + c_n^2}) \tag{106.3}$$

As discussed in Sect. 106.1, a document will be abstracted as a vector. The element of this vector is the occurrence number of all words in this document. For document X and Y, they are expressed by two vectors, $< x_1, x_2 \ldots x_n >$ and $< y_1, y_2 \ldots y_n >$. Then the intersection angle of them is expressed as formula 106.4.

$$\cos(A) = (x_1 * y_1 + x_2 * y_2 + \ldots + x_n * y_n)/(\sqrt{x_1^2 + x_2^2 + \ldots + x_n^2} * \sqrt{y_1^2 + y_2^2 + \ldots + y_n^2}) \tag{106.4}$$

When the cosine value is equals to 1, the intersection angle is zero, and the two documents are totally the same. When the value is zero, the intersection angle is 90°, and the two documents are absolutely different.

106.3 Document Preprocessing

Before calculating the word number of a document to set up the feature vector, the document should be preprocessed to decrease some noise introduced intentionally by the document author.

106.3.1 Remove Words

In a document, especially research paper, noun, verb and adjective are meaningful and not easy to replace. In some documents some other words such as qualifier or numeral are also meaningful. It should be configurable to decide which kind of words should take into account. In this section and the sample, noun, verb and adjective are included. For other types of words, they can easily be added or removed to add noise to avoid plagiarism detection. Therefore, these words are not meaningful in plagiarism detection and should not be included in the feature vector. So these words should be removed before counting word number.

106.3.2 Group Words

A sentence may be restructured to evade plagiarism detection. In this case, the words are disordered and the tense of verbs will change. So, in plagiarism detection, a verb with different tense should be grouped. Therefore, before counting word number, all tense of verb should be replaced by the original tense. For similar reason, plural noun should be replaced with singular form.

After preprocessing, it is ready to count the word number and get the feature vector. For example, if there are two words and their occurrence number are 3 and 7, then the vector should be $< 3, 7 >$.

106.4 Adjust Weight

As mentioned in Sect. 106.1, if two words occur with different frequency in the library, they should have different weight when calculating cosine similarity.

If a word occurs N times in the whole library, then the weight should be 1/N for this word. If it occurs M time in a document, then the number in the vector should be M/N. Therefore, if the words occurrence numbers in a document are M_1, M_2, …, M_k, and their occurrence numbers in the library are N_1, N_2, …, N_k, then the feature vector should be $< M_1/N_1, M_2/N_2, \ldots, M_k/N_k >$.

106.5 Sample

This section illustrates the previous method by a sample. The following paragraph A comes from a paper. Paragraph B is made from paragraph A to simulate plagiarism. Paragraph C expresses the same meaning as A, but it is written from scratch and does not copy anything.

Paragraph A:

"B is a formal method for specifying, refining and implementing software. The main idea of B is to start with a very abstract model of the system under development and gradually add details by building a sequence of more concrete models. A development process creates a number of proof obligations, which guarantee the correctness. Proof obligations can then be proven by automatic or interactive tool. The goal of B is to obtain a proved model."

Paragraph B:

"B is a formal method to specify, refine and implement software. The goal of B is to obtain a proved model. The idea of B is starting with an abstract model of a system under development and gradually adding details by building a sequence of more concrete models with obligations. This sentence is meaningless and just adds noise. A lot of proof obligations, which can be proven by interactive or automatic tool and can guarantee the correctness, compose the development process."

Paragraph C:

"B is a method to develop correct software. Software is initially an abstract specification of B, and then refined step by step to executable program. The refinement steps are correct ensured by theorems and rules. If the initial specification is correct, and the refinement steps are also correct, then the final program is also correct."

106.5.1 Preprocessing

Using the rules of Sect. 106.3, the three paragraphs are processed and the following are the results.

Paragraph A:

"B formal method specify refine implement software main idea B start abstract model system development add detail build sequence concrete model development process create number proof obligation guarantee correctness Proof obligation prove automatic interactive tool goal B obtain proved model."

Paragraph B:

"B formal method specify refine implement software goal B obtain proved model idea B start abstract model system development add detail build sequence concrete model obligation sentence meaningless add noise proof obligation prove interactive automatic tool guarantee correctness compose development process."

Paragraph C:

"B method develop correct software Software abstract specification B refine step step executable program refinement step correct ensure theorem rule initial specification correct refinement step correct final program correct."

106.5.2 Feature Vector

Based on the processed paragraphs, the words are count and the feature vectors are calculated out. To make the three vectors have the same dimension, the corresponding element will set as zero if a word does not exist in one paragraph. The following vectors are of paragraph A, B and C.

Vector A:

<tool:1, detail:1, refinement:0, obligation:2, software:1, correct:0, initial:0, compose:0, proved:1, final:0, prove:1, meaningless:0, theorem:0, automatic:1, number:1, create:1, goal:1, concrete:1, obtain:1, system:1, formal:1, start:1, program:0, specify:1, model:3, idea:1, development:2, B:3, executable:0, proof:2, sentence:0, add:1, build:1, refine:1, interactive:1, rule:0, implement:1, develop:0, correctness:1, ensure:0, abstract:1, main:1, noise:0, process:1, sequence:1, specification:0, guarantee:1, method:1, step:0>

Vector B:

<tool:1, detail:1, refinement:0, obligation:2, software:1, correct:0, initial:0, compose:1, proved:1, final:0, prove:1, meaningless:1, theorem:0, automatic:1, number:0, create:0, goal:1, concrete:1, obtain:1, formal:1, system:1, start:1, program:0, specify:1, idea:1, model:3, development:2, B:3, executable:0, proof:1, sentence:1, add:2, build:1, refine:1, interactive:1, rule:0, implement:1, develop:0, correctness:1, ensure:0, abstract:1, noise:1, main:0, process:1, sequence:1, specification:0, guarantee:1, method:1, step:0>

Vector C:

<tool:0, detail:0, refinement:2, obligation:0, software:2, correct:5, initial:1, compose:0, proved:0, final:1, prove:0, meaningless:0, theorem:1, automatic:0, number:0, create:0, goal:0, concrete:0, obtain:0, system:0, formal:0, start:0, program:2, specify:0, model:0, idea:0, development:0, B:2, executable:1, proof:0, sentence:0, add:0, build:0, refine:1, interactive:0, rule:1, implement:0, develop:1, correctness:0, ensure:1, abstract:1, main:0, noise:0, process:0, sequence:0, specification:2, guarantee:0, method:1, step:4>

106.5.3 Cosine Value

By formula 106.4, cosine between vector A and B is 0.92 and between A and C is 0.17.

106.5.4 Adjust Weight

The occurrence frequency of these words should be calculated for the whole digital library. However, the author does not have such a library. Therefore, in this

sample, some assumptions are set to demonstrate how frequency takes effect. Let's assume "B" occurs frequently in the library because it is a widely studied topic, and assume "obligation" is less frequently present. If "B" occurs three times frequently and "obligation" occurs 1/3 time compared with other words, the weight of "B" should be 1/3, and the weight of "obligation" should be 3. Then the vectors should be as following after adjusting the weight.

Vector A:

<tool:1, detail:1, refinement:0, obligation:6, software:1, correct:0, initial:0, compose:0, proved:1, final:0, prove:1, meaningless:0, theorem:0, automatic:1, number:1, create:1, goal:1, concrete:1, obtain:1, system:1, formal:1, start:1, program:0, specify:1, model:3, idea:1, development:2, B:1, executable:0, proof:2, sentence:0, add:1, build:1, refine:1, interactive:1, rule:0, implement:1, develop:0, correctness:1, ensure:0, abstract:1, main:1, noise:0, process:1, sequence:1, specification:0, guarantee:1, method:1, step:0>

Vector B:

<tool:1, detail:1, refinement:0, obligation:6, software:1, correct:0, initial:0, compose:1, proved:1, final:0, prove:1, meaningless:1, theorem:0, automatic:1, number:0, create:0, goal:1, concrete:1, obtain:1, formal:1, system:1, start:1, program:0, specify:1, idea:1, model:3, development:2, B:1, executable:0, proof:1, sentence:1, add:2, build:1, refine:1, interactive:1, rule:0, implement:1, develop:0, correctness:1, ensure:0, abstract:1, noise:1, main:0, process:1, sequence:1, specification:0, guarantee:1, method:1, step:0>

Vector C:

<tool:0, detail:0, refinement:2, obligation:0, software:2, correct:5, initial:1, compose:0, proved:0, final:1, prove:0, meaningless:0, theorem:1, automatic:0, number:0, create:0, goal:0, concrete:0, obtain:0, system:0, formal:0, start:0, program:2, specify:0, model:0, idea:0, development:0, B:0.67, executable:1, proof:0, sentence:0, add:0, build:0, refine:1, interactive:0, rule:1, implement:0, develop:1, correctness:0, ensure:1, abstract:1, main:0, noise:0, process:0, sequence:0, specification:2, guarantee:0, method:1, step:4>

106.5.5 Cosine Value

With the adjusted vector, cosine between vector A and B is 0.95 and between A and C is 0.08. Compared with the original value 0.92 and 0.17, cosine between vector A and B becomes larger and between A and C becomes smaller. This is more accurate because B is copied from A, and C not.

106.6 Conclusion

This paper improves cosine similarity measure algorithm in plagiarism detection by two ways. One is to preprocess document before calculation. The second is to adjust the weight of the words dependent on the occurrence frequency of the words in the library. The sample proves this improvement is effective.

The future work includes how to preprocess documents more accurately in specific fields. Another more practical work is to build the library and calculate the occurrence frequency so that it is more accurate to adjust the word weight. The threshold that warns plagiarism is also an important topic. The algorithm in this paper only calculates the cosine value while cannot decide the occurrence of plagiarism. It needs more practical training to determine such a threshold.

References

1. Sven, M.E., Benno, S.: Intrinsic plagiarism detection. In: Advances in Information retrieval 28th European Conference on IR Research, ECIR 2006, London, UK, Automatic Conceptual Analysis for plagiarism detection April 10–12, 2006 Proceedings. Lecture Notes in Computer Science, vol. 3936, pp. 565–569. Springer (2006)
2. Zechner, M., Muhr, M., Kern, R., Granitzer, M.: External and intrinsic plagiarism detection using vector space models. PAN's, pp. 47–55 (2009)
3. Kang, N., Han, S.Y.: Document copy detection system based on plagiarism patterns. In: CICLing'06 Proceedings of the 7th international conference on computational linguistics and intelligent text processing, pp. 571–574 (2006)
4. Si, A., Leong, H.V., Lau, R.W.H.: CHECK: A document plagiarism detection system. Proc. ACM Symp. Applied Comput., 70–77 (1997)
5. Dreher, H.: Automatic conceptual analysis for plagiarism detection. J. Issues Informing Sci. Inf. Technol. 601–614 (2007)
6. Kang, N., Gelbukh, A., Han, S.: PPChecker: Plagiarism pattern checker in document copy detection. Proc. TSD, 661–667 (2006)
7. Timothy, H., Justin, Z.: Methods for Identifying versioned and plagiarized documents. J. Am. Soc. Inform. Sci. Technol. **54**(3), 203–215 (2003)

Chapter 107
The Application of I/O Virtualization Framework in TaiShan Nuclear Power Plant

Kongtao Li, Yao Yu and Yi Luo

Abstract With the acceleration in the construction of TaiShan Nuclear Power Plant, there is also a dramatic increase in related application systems, server and storage amount. The traditional I/O framework can no longer meet the needs and it frequently encounters such problems as the severe shortage of servers integrated after virtualization, restrictions in Virtual Manufacturing (VM) performance, etc., which are the disadvantages of the stable operation of the production system. Therefore, the I/O virtualization framework is based on cloud delivery concept emerges, with simple and explicit topology, convenient deploy, bandwidth on demand, high utilization rate and QoS guarantee. It can also boost the transmission speed of data drastically and guarantee the stability in the operation of the nuclear power production system.

Keywords I/O Virtualization framework · Storage server · Transmission bottleneck · Stability

107.1 Introduction

TaiShan Nuclear Power Plant (TSNPP) is the third generation ERP pressurized water reactor (PWR) nuclear power plant. Since server and storage are indispensible components in the nuclear power production system. Fundamental platform for server virtualization has already been deployed by TSNPP, which has improved the utilization rate of resources. But with the deepening of virtualization application, the I/O bottleneck problems become much more evident day by day, and I/O virtualization turns to be the demanding improvement approach. The

K. Li (✉) · Y. Yu · Y. Luo
Information Technology Center, China Nuclear Power Technology Research Institute,
China Guangdong Nuclear Power Holding Co. Ltd,
Shenzhen, China
e-mail: likongtao@cgnpc.com.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_107,

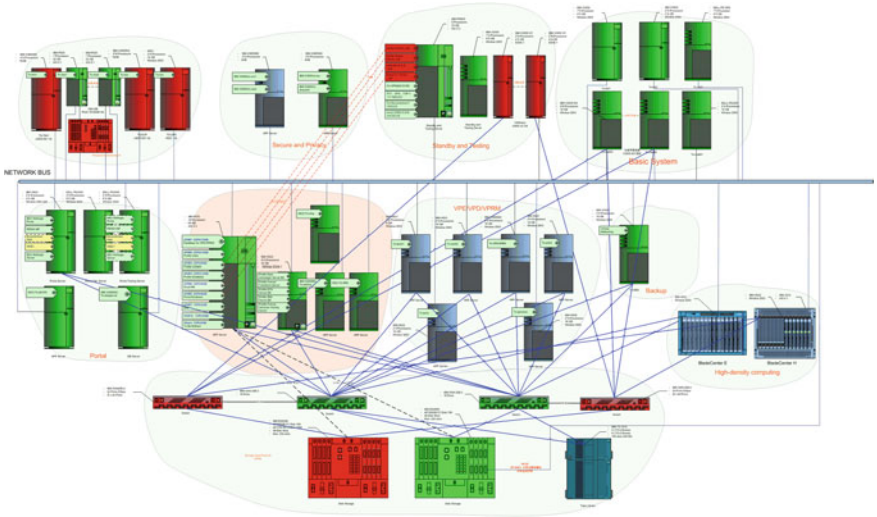


Fig. 107.1 The network structure of servers

topological graph of both the server and storage physical structure in TSNPP is shown in the following figure.

As shown in the Fig. 107.1, there are 46 PC servers, 2 sets of storage, a set of tape library, 2 sets of blade servers, four fiber switches, covering such application systems as safety, test and reserve, Portal, database, VPE/VPD/VPRM, etc. And only parts of the links have been represented.

107.2 Current Situation

The server and storage in the data centre of TSNPP still connect to the website in an old model about a decade ago in which the IT managers should employ substantial cables to connect the network card on the server and fibre card with network switch and fibre switch. Therefore, there is a long-existing and complex problem in management. The I/O framework lack of flexibility not only increases the system cost and lowers the system deploying speed and virtualized bandwidth stability, but also fails to respond to the ever-changing business demands. At present, the virtualized basic cloud computing framework deployed by TSNPP with blade server as well as the bandwidth of its key application system also come to obvious bottleneck.

Since the VMware diagnostic tool is immature, there is a lack of popularity in the current I/O of the server, and as a result it is difficult to locate and settle the performance issues. Even though the managers have confirmed the root of the bottleneck, it may be solved by purchasing more network cards or by employing

load balancing in VM. In face of bottleneck, the I/O connections of the server will be increased by the IT managers to satisfy the demands for virtualization. The root of the unsatisfactory in this method lies in that the virtualized users should deploy 6–16 I/O connections for each server, which not only increases the equipments and consumables, but also needs larger server to insert more I/O card, thus there is a drastic increase in the construction cost of the data canter, occupying space of the equipment and energy consumption.

107.3 Xsigo Virtualized I/O Technology

In order to settle the restrictions from the traditional I/O framework, the manufacturer Xsigo [7] of the virtualized I/O has researched and developed a technology supporting 20/40 Gb Infiniband and 10 GB Ethernet, and meanwhile, it also supports the X86 heterogeneous racks and blade servers from different manufacturers, realizing any interconnections among the storages such as the gigabit or ten thousand network, fibre, iSCSI (internet Small Computer System Interface), etc. The Xsigo I/O Director has provided various technology innovations for the overall virtualization of IT fundamental framework.

Dynamic allocation bandwidth: The 40 GB bandwidth of each serve can be shared by the virtual machine dynamically. All the bandwidth can be allocated dynamically to the network or storage channel according to the demands.

Resource isolation: It can distribute special vNIC connection for specific VM and realize I/O isolation.

QoS guarantee: It should be guaranteed that the key applications can achieve stable bandwidth through hardware-level electrical isolation. Even in case of fights for resources, QoS can allocate guaranteed bandwidth resource to specific application, thus expected application performance can be guaranteed. Throughput rate is realized by hardware, controlled by custom setup by the user, it includes as follows:

Committed Information Rate (CIR) ensures the smallest bandwidth.

Peak Information Rate (PIR) restricts the largest bandwidth consumed by the sources.

107.4 Construction of I/O Virtualization Framework for Data Center

107.4.1 Overall Frame

In the conception of I/O virtualization in TSNPP, the framework deployed is a virtual application scenario of X86 rack server, blade server and Storage Area

Table 107.1 Equipment disposition

Item	Configuration	Type	Amount
Environment configuration of the blade	Blade server	Blade Center	21
	Blade chassis		2
	InfiniBand switch model.	4X QDR IB Switch	4
	40 GB HCA card	4X QDR IB CX2 Dual Port Mezz HCA	21
X86server configuration	20/40 GB Infiniband card	IBM 4X DDR/QDR IBboard	46
Xsigo configuration	IO Director	Xsigo VP780 DDR Chassis	2
	Storage IO switching in module (8 GB)	4 8 GB I/O ports in each machine (2 × 2 Port 8 Gigabit Fibre Channel Module)	4
	Network IO switching in module (10 GB)	2 10 GB I/O switching in module in each machine	4

Network (SAN) [6], fibre network and ten thousand network switching. In this framework, the key system equipment includes: two sets of IBM Blade Center E IBM Blade Center H blade server, 46 X86 rack servers, two sets of IBM DS5300, IBM DS4800 storage, a set of IBM TS3310 tape base and two SAN switches with IBM B5K32 port and IBM B40 port for each. The new core equipment of the server is VP780 of Xsigo (Table 107.1).

107.4.2 Configuration Scheme

The topology figure and configuration table above have the following advantages in TSNPP.

Since TSNPP is in construction, the continuous increase of servers can add server much more conveniently.

X 86 server greatly decrease the network connections, which saves a lot of space and is convenient for management.

Configure InfiniBand switching module, replacing the transmission I/O of IP network and SAN network. Therefore, the transmission speed can reach 40 GB theoretically.

Several technologies such as the dynamic bandwidth allocation, resource isolation, QoS guarantee (hardware-level electrical isolation) have been adopted, which can improve the flexibility and controllability of the whole framework [1].

Optimize the backup [5] of the system framework, increase the backup network bandwidth and speed, and shorten the backup time window for 4 times.

Set up speedy and isolated vMotion exclusive network, increase the transfer speed of the virtualized platform, and decrease the setting up of separate server network port of vMotion (Table 107.2).

From the comparison table above, it can be seen that the amount of I/O ports in the framework have been reduced after the adoption of new technology, which saves the port resource of the network/fibre switch, and doubles the transmission bandwidth (Fig. 107.2).

107.4.3 Application Significance of I/O Virtualization in TSNPP

107.4.3.1 Resource Integration and Sharing, Realization of A2A Connection

Due to the fact that TSNPP is in construction stage, the server and storage [2] equipment has involved several hardware server manufacturers, and virtualization technology (VMware, HyperV, OVM, KVM, Xen and etc.). When the resource

Table 107.2 I/O Comparison table for old and new technologies

Original technology	Amount of X86 rack server	Amount of I/O port	Illustrations	New technology	Amount of X86 rack server	Amount of I/O port	Illustration (benefit)
	46	46 × 4	Calculated that each server has 2 Ethernets and 2 SAN fibre networks. The bandwidth of single-path is 1 GB and the storage band width is 4 GB		46	46 × 2	Decrease the I/O port by 17–19, single-path bandwidth is raised by 40 GB at most
		Blade server	Amount of I/O port		Blade server	Amount of I/O port	Illustration (benefit)
	2 sets/21 blade computing	21 × 4	Calculated that each blade has 2 Ethernets and 2 SAN fibre networks. The bandwidth of single-path is 1 GB and the storage band width is 4 GB		2 sets/21 blade computing	21X2	Decrease the I/O port by 46, single-path bandwidth is raised by 40 GB at most

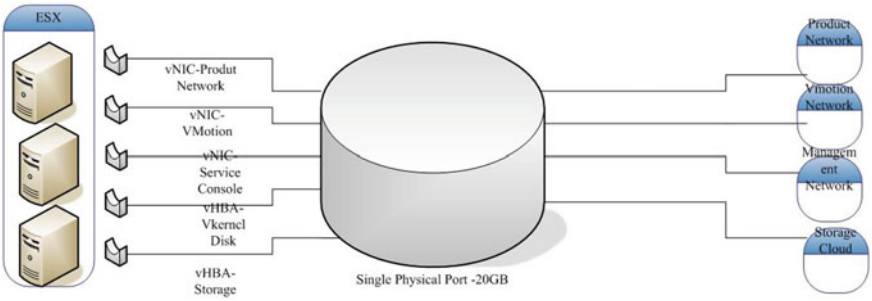


Fig. 107.2 Logic diagram of the virtualized I/O is shown in the following figure

integration in the I/O virtualization platform is realized, it can provide A2A (Any server to any network and storage) connection for the server and storage system [4] of TSNPP. The realization of A2A connection can solve the inter-platform rapid transmission in the nuclear power environment.

107.4.3.2 Fusion Environment for the Compatible Rack and Blade Server

The conception of I/O virtualization has provided a unified I/O management for the heterogeneous server [3] environment of TaiShan nuclear power data center, which connects heterogeneous networks and storage resource of different types. Especially the blade server with limited I/O extension ability, the I/O performance dramatically limits the capability of the blade server to operate several virtual machines on a single host computer.

107.4.3.3 Realizing the Fast Transfer of Server Failure Without Re-Mapping or Re-Wiring

The virtual I/O provides important support for the fast transfer of server.

1. Disaster recovery: The I/O resource some site can be reset on another site rapidly, increasing the switching speed.
2. Replacement of server: When one server breaks down, its complete I/O template (including WWN and MAC address) can be transferred to another server.
3. Safety: In order to guarantee the security, each server can only allocate the current connection needed. When there are changes in demand, there is no need to enter into the data centre to alter the configurations.

107.4.3.4 Strengthening the I/O Security of the Virtual Computing Environment for TSNPP

Through the electrical isolated virtual I/O framework, independent vNIC can be allocated for the ESX server of the virtual machine, and connection to firewall Demilitarized Zone (DMZ) can be added and removed in time so that the potential safety hazard can be eliminated by controlling the risks. If there are more DMZ connection demands, the connections to other ESX server could be set up at any time. Such deployment is better to prevent the existing hidden danger in the complex network environment of the contractor in TSNPP, to guarantee the security of transmission among various applying servers, and to remarkably enhance the safety performance.

107.4.3.5 Setting up Speedy and Isolated vMotion Exclusive Network for TSNPP

1. Isolate the vMotion transmission from other transmission of the data centre, to improve the safety.
2. Set up speedy network specialized in vMotion transmission, and improve the performance and reliability of vMotion.
3. Control all the vjMotion transmission within Xsigo I/O Director, to reduce the burden of LAN.
4. Lower the demand for expensive speedy network and server NIC.

107.4.3.6 Flexible VM Transfer during Service Period without Reallocating Network and Storage Setting

In the current stage of TSNPP, due to the demand from construction and commissioning, the server will be transferred frequently, but the demand from security will sometimes limit the using of vMotion. When transferring VM from one server to another, the two servers should see the same network and storage synchronously. Therefore, all the servers should be opened for visiting in the vMotion collection, and such configuration cannot satisfy the security demands in some IT environment. I/O virtual agency provides another choice so that the transfer of virtual machine can be realized without opening visit. Through I/O transfer, a VM can be hanged on one server, then recover on another server, and there is no need to open storage of the two servers, which can be realized from three steps.

Hang the virtual machine on server A.

Transfer the application storage and network (vHBA and vNIC) to server B.

Recover virtual machine on server B.

The realization of the simple process relies on that WWN keeps unchanged during the vHBA transferring process. And then the target LUN and related partition transfer with vHBA so that the server managers can restore the configuration without any alteration and recover the virtual machine on a new server.

107.4.3.7 Lowering the Total Cost of I/O and Energy Consumption

I/O virtual technology has eliminated the limitations in the traditional I/O connections so that the server can be widely deployed into the virtual environment. 32 full redundant vNIC ans 12 full redundant vHBA can be deployed on one server. The simplification in the amount od server I/O card also means the decrease in power consumption. The number of edge networks, fibre switches and cables also decrease, and the server I/O equipment cost can be saved by 50–70 %, while the energy consumption decreases by 35 %.

107.5 Conclusion

I/O virtualization framework in this paper solves the I/O bottleneck problems of the server in TSNPP largely. Meanwhile, due to the unified framework employed, there is no need to purchase various I/O boards for each platform server, which greatly saves the cost in prophase investment. For various small-scale applications, with the I/O virtualization, it will be convenient to transfer, backup and restore, which guarantee the convenience in maintenance to the full. And it can adapt to the fast-paced application changes during the period of construction.

References

1. Xie, X.: Computer Networking Technology, vol. 4, pp. 382–384. Electronic Industry Press, Beijing (2003)
2. Haojing, H., Wei, G.: Practice Course For Server And Storage Project. China water conservancy and electric power press, Beijing (2012)
3. Liu, H.: IBM Minicomputers into the World. Electronic Industry Press, Beijing (2010)
4. Bryant, R.E., O'Hallaron, D.R.: Computer Systems: A Programmer's Perspective, 2nd edn. Prentice Hall, Englewood Cliffs (2012)
5. Preston, W.C.: Unix Backup and Recovery, pp. 136–138. O'Reilly, Cambridge (2003)
6. Peterson, L.L., Davie, B.S.: Computer Networks: A Systems Approach, pp. 75–76, 3rd edn. Elsevier, Singapore (2005)
7. Xsigo Technique Manual, pp. 5–12 (2012)

Chapter 108
Application of Virtual Reality Techniques for Simulation in Nuclear Power Plant

Junjun Zhang and Xuan Zhang

Abstract Three-dimensional (3-D) virtual reality has proved to be an effective and efficient training tool to impart plant knowledge. In order to enhance users' understandings of the nuclear reactor principles, a virtual reality system based on simulator has been developed to interface with the scenarios in nuclear power plant (NPP). Physical characteristics are calculated by simulation codes, such as pressure, temperature, flow and void. The simulator transmits physical characteristics to virtual reality system for three dimension dynamic visualization of these parameters. It is useful to help analyzers understand the conditions of nuclear power plant. This paper introduces the basic concept, framework of the virtual reality system and its functions.

Keywords: Nuclear power plant · Virtual reality · Simulator

108.1 Introduction

The safe and efficient operation of a nuclear power plant is highly dependent on the knowledge and skills of operators. Training of operators at nuclear facilities is even more necessary and important than at other facilities. It is significant to spend time and resource on education and training at nuclear facilities [1]. Reactor simulator is a key step towards enhancing the operator capability and significantly improving the safety of the plant. However, the simulator normal indicators and graphical interface may not provide sufficient information on the behavior of the power plant during the accident conditions and it is not enough for researchers due to a lack of interactivity [2–5]. In addition, the complexity of accident phenomenology, especially server accident, makes the respective simulators development a very demanding task. On

J. Zhang (✉) · X. Zhang
China Nuclear Power (Beijing) Simulation Technology Corp. Ltd, Shenzhen, China
e-mail: dcdbb@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_108,

the purpose of solving these problems, a training paradigm based on virtual reality can bring promise towards achieving the training goals.

With the development of computer technology, 3-D virtual reality has been more mature and widespread used in diverse fields. It is also being studied and used in many ways in nuclear fields. Nuclear radiation which does rates assessment based on virtual reality has been researched by some groups [6]. Robert and Joseph studied using 3-D visualization technology for training first responders to nuclear facilities [7]. A group in Brazil took use of virtual reality for verification of the human factor in nuclear facilities [8]. But these studies were limited to dose management, maintenance management, etc. They did not involve application of virtual reality to simulation NPP.

In this paper, we developed a virtual reality system based on simulator to analyze complex nuclear reactor phenomena. Users can experience and emulate the operational and emergency scenarios of NPP by interactive 3-D graphical models of nuclear facilities.

108.2 Methodologies

108.2.1 Framework

The system was based on virtual reality engine and the code was written in C++ and Microsoft Foundation Class (MFC). The framework was shown in Fig. 108.1.

The framework is composed of four components, virtual scene, calculations codes, and GUI and communication module. Virtual scene based on virtual reality engine consists of two parts: geometry model and physical model. Geometry model shows the structure of NPP, and physical model makes them truer. The engine encloses some basic functions of managing the models. Calculations codes were to simulate the conditions of NPP. GUI is an interface to the virtual scene. Users can control and view the state of NPP. The communication module provided

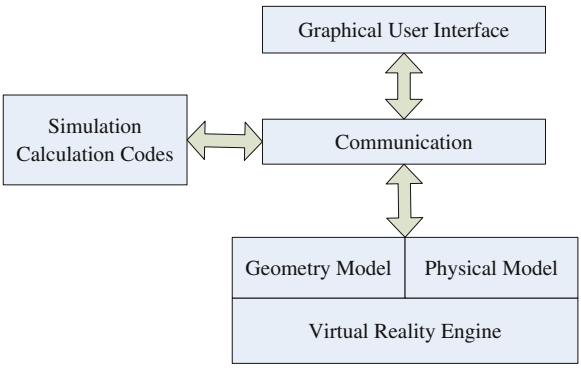


Fig. 108.1 Framework

mechanism of setting command and getting data among the other components. More details were given as follows.

OpenSceneGraph, or OSG, for short, was chosen from the beginning as a virtual reality engine. The OpenSceneGraph is an open source high performance 3D graphics toolkit, used by application developers in fields such as visual simulation, games, virtual reality, and scientific visualization and modeling. The OpenSceneGraph is now well established as the world leading scene graph technology, used widely in the space, scientific, oil-gas, games and virtual reality industries [9].

3D Max was used to create the geometry model, which described the important components of nuclear power plant. Furthermore, we developed an interface which can automatically convert the Computer- Aid-Design model from design data into 3D max model. Physical model provided the function of collision detection which made virtual roaming available.

Calculation codes compute the physical data which were predefined by analyzers. The computed result was transmitted to the virtual reality engine. So some physical phenomena were dynamically shown according to parameters. The calculation codes also accepted the commands from users for the purpose of controlling its run. Graphical user interface (GUI) was to present the scene and physical parameter values for users. The exchange data were performed by commutation model (shown in Fig. 108.2), which is introduced in detail in next section.

108.2.2 Communication Model

The communication model is key model, which connects all components in the system. The schema was shown in Fig. 108.2.

Dynamic link library is used to implement the communication model. It provides three options for various conditions. (a) Share memory: Shared memory is the simplest and the fastest protocol to use and has no configurable settings. Clients using the shared memory protocol can only connect to server instance running on the same computer. (b) TCP/IP: TCP/IP is a common protocol widely used over the Internet. It communicates across interconnected networks of computers that have diverse hardware architectures and various operating systems. Some calculation codes can only be run in Linux or UNIX operating system. So it is the most useful protocol that is used in our system. (c) PIPE: For named pipes,

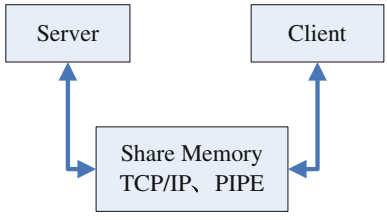


Fig. 108.2 Communication schema

network communications are typically more interactive. A peer does not send data until another peer asks for it using a read command. A network read typically involves a series of peek named pipes messages before it starts to read the data. These can be very costly in a slow network and cause excessive network traffic, which in turn affects other network clients [10].

In order to perform the communication between server and client, three main processes have been created for this purpose, one process receives and transmits the data calculated by simulator codes; another process publishes these data to OSG; the third process can serve as an interface between these two processes. All these processes can be run in one or more computers, and are accessible to each other through a local network.

108.3 Main Functions

Fig. 108.3 shows the GUI of virtual reality system, it has three important functions.

108.3.1 Import 3D model

The system can import various 3D models such as .3ds,.max etc., it also can convert CAD model into 3D model supported by the system. The CAD model involved IGES\SAT sourced from ProE, CATIA, UG, etc. This function enables powerful repair and simplification tools. The interactive repair feature assures that imported geometries are correct for rendering. And, for purpose of cutting down on unnecessary details in CAD models, it can remove fillets, small components or parts.

Fig. 108.3 GUI of virtual reality system

108.3.2 Rendering Scene

This function deals with how to render models generated from Sect. 3.1 cited before according to calculation results. A hierarchy graph of nodes was used to represent the spatial layout of graphic and state objects. It encapsulates the lowest-level graphics primitives and state combined to visualize primitive model. Object traversal, transform, culling of the scene, level-of-detail management, and other basic or advanced graphics characteristics [11] were provided to show the progress of accident. Particle effect was used to simulation explosion.

108.3.3 Data Validation

The computed data from NPP simulator were huge, so it was difficult to have a good command of it. This function provided a tool which can directly extract meaningful data from the results. Furthermore, users can easily understand it by 2D curve, 3D surface, table, etc. Fig. 108.4 shows the curves of validation of nuclear reactor power data between simulation and experimental design. The maximum error between them is less then 3 %.

108.4 Applications

There were many components in nuclear power plant, it was difficult to 3-D detailed visualize these components, and it was not necessary. In virtual reality

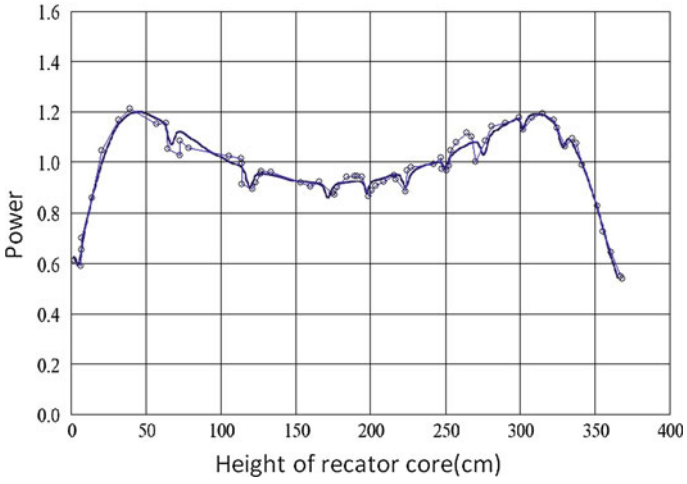


Fig. 108.4 Comparision curve of nuclear reactor power

environments, it was important to visualize the physical phenomenon of simulation of nuclear power plant. The first step was building virtual scenes of nuclear power plant in our virtual reality system. It can be performed by importing the CAD models from design data, and simplified some unnecessary details for rendering efficiency. Based on these virtual reality models, some physical phenomena which were driven by simulation codes can be shown, which can help the users easily comprehend the process of nuclear plant power.

108.5 Conclusion

Virtual reality environment (advanced algorithms, graphics hardware, etc.) has already been mature so that almost any single complex phenomenon could be solved. Based on OSG, we developed the VR prototype system to help analyzers understand the conditions of nuclear power plant. More functions are under development. We hope that it is a powerful tool for comprehending complex nuclear power plant phenomena in 3D environment in real time.

References

1. Xi, C., Wu, H., Joher, A., et al.: 3D virtual reality for education training and improved human performance in nuclear engineering. C. ANS NPIC HMIT 2009 Topical Meeting-Nuclear Plant Instrumentation, Controls, and Human Machine Interface Technology. ANS, Knoxville, Tennessee (2009)
2. Mol, A.C.A., Aghina, M.A.C., Jorge, C.A.F., et al.: Nuclear plants virtual simulation for online radioactive environment monitoring and dose assessment for personnel. J. Ann. Nucl. Energy **36**, 1745–1752 (2009)
3. Mol, ACA., Jorge, C.A.F., Couto, P.M., et al.: Virtual environments simulation for dose assessment in nuclear plants. J. Prog Nucl. Energy **51**, 382–387 (2009)
4. Mizuguchi, N., Tamura, Y., Imagawa, S., et al.: J. Dev. React. Des. Aid Tool Using Virtual Reality Technol. Fusion Eng. Des. **81**, 2755–2759 (2006)
5. Liu, P.-F., Yang, Y-H., Yang, Y.-M., et al.: Application of virtual reality to simulation in nuclear power plant. At. Energy Sci. Technol. 2008-S1 (2008)
6. Rodenas, J., Zarza, I., Burgos, M.C., et al.: Developing a virtual reality application for training nuclear power plant operators: Setting up a database containing dose rates in the refueling plant. J. Radiat. Protect. Dosim. **111**, 173–180 (2004)
7. Sanders, R.L., Lake, J.E.: Training first responder to nuclear facilities using 3-D visualization technology. In: Proceedings of the 2005 Winter Simulation Conference (2005)
8. Jose, I., dos Santos, A.L., et al.: The use of questionnaire and virtual reality in the verification of the human factors issues in the design of nuclear control desk.J. Int. J. Ind. Ergon. **39**, 159–166 (2009)
9. OpenSceneGraph, http://www.openscenegraph.org/
10. Chossing a Network Protocol, http://msdn.microsoft.com/
11. Wang, R., Qian, X.: OpenSceneGraph 3.0: Beginners Guide, PACKT publishing (2010)

Chapter 109
SCM-BSIM: A Non-Volatile Memory Simulator Based on BOCHS

Guoliang Zhu, Kai Lu and Xu Li

Abstract New storage-class memory (SCM) technologies, such as phase-change memory, are fast, non-volatile and byte-addressable. SCM provides a new realm for researchers to boost the performance of system. But most of SCM devices are not available on the market, which hindered further software research on leveraging the full feature of SCM. In this paper we design and implement a SCM device simulator on BOCHS named SCM-BSIM. SCM-BSIM can mimic full feature of SCM such as non-volatility and different access latency. Also it will gather life span statistics during simulation to support endurance relevant research. With a BOCHS-based interface, SCM-BSIM is easy to use.

Keywords Non-volatile · Endurance · Simulator

109.1 Introduction

Storage-class memory, which features non-volatility, low power consumption, fast access time and high density, is becoming both a heated academic topic and a fierce battlefield for storage industry. Although SCM is still under development, it is predicted that in storage system that uses SCM as disk driver replacement, the system will have random and sequential I/O performance that is orders of magnitude better than traditional disk-based systems [1]. As a promising storage technology,

G. Zhu (✉) · K. Lu · X. Li
National University of Defense Technology, ChangSha, China
e-mail: zgl8905@gmail.com

K. Lu
e-mail: kailu@nudt.edu.cn

X. Li
e-mail: lixu@nudt.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_109,

SCM has opened a new realm for storage research. Researchers are working on designing file system and hardware architecture that are specified around the properties of SCM [2–4]. Programmers can design new data structure on SCM [5].

In academic field, software research and development efforts are underway worldwide on SCM. Since some SCM devices are not available on the market, researchers have to gain a more specific model of SCM so that they can better leverage the performance of SCM.

However, nowadays there is no simulator that is capable of simulating the full feature of SCM. In this paper, we present an architecture-level simulator named SCM-BSIM which mimics the full properties of SCM device. SCM-BSIM is based on BOCHS that includes a non-volatile component, a device component and an endurance component. The non-volatile component simulates the non-volatile feature of SCM. The device component simulates the read/write latency of SCM devices as specified by users to mimic different devices. The endurance component is designed to gather the memory access statistics. Researchers can leverage the endurance component to develop new wear-leveling paradigm. SCM-BSIM also includes a GUI which makes it easy to use.

109.2 Background

109.2.1 Storage-Class Memory

As a class of storage technologies that features non-volatility, low power consumption, fast access time and high density, SCM not only complement the existing memory and storage hierarchy but also reduce the distinctions between memory and storage [1]. Currently, there is a fierce competition going on in the industry of SCM storage between large manufacturers like IBM, Intel and Toshiba. Phase Change RAM, NAND flash memory, Magnetic RAM and Magnetic Racetrack are all promising candidates of SCM [6] (Table 109.1). But currently some SCM products (such as Phase-Change RAM) are not available on the market. Existing simulators for SCM devices are either at device-level or based on CACTI and are not capable of providing the non-volatile feature of SCM [7]. Researchers have to adopt DRAM to simulate PCM [2, 5].

109.2.2 Bochs

BOCHS is a highly portable open source IA-32 (×86) PC emulator written in C++, that runs on most popular platforms. It includes emulation of the Intel ×86 CPU, common I/O devices, and custom BIOS. BOCHS can be compiled to emulate many different ×86 CPUs, from early 386 to the most recent ×86-64 Intel and AMD processors which may even not have reached the market yet.

Table 109.1 Comparison of access latency and endurance (in number of overwrites) of current and future dram, pcm, stt-ram, and flash memories. prospective characteristics are based on demonstrated prototypes

Category	Technology	Memory characteristics		
		Read	Write	Endurance
Current memory	DRAM	60 ns	60 ns	$> 10^{16}$
Storage-class memory	NAND Flash	25 µs	200–500 µs	10^4–10^5
	PCM sample	115 ns	120 µs	10^6
	PCM future	50–85 ns	150–1000 ns	10^8–10^{12}
	STT-RAM	6 ns	13 ns	10^{15}

BOCHS features portability. It runs on all the editions of Windows, Linux and Mac OS, while VMware cannot run on Mac OS and Virtual PC is not available on *BSD, BeOS and some editions of Windows.

BOCHS is written in high level language C++, thus making it easy to be extended and tuned to provide more function. And BOCHS has become an indispensible tool for debugging, so that programmer can perform security checks and analysis on the guest code and can build powerful tool based on BOCHS [8]. Furthermore, BOCHS is capable of handling the difference as endian, register width and address space sizes between the guest and the host.

These reasons make BOCHS a good platform to develop memory-relevant software. Although BOCHS is a purely interpreted execution virtualization product, research shows that BOCHS simulation boot time for Windows XP has already fall to 81 s on Core 2 Duo [8]. Therefore we chose BOCHS to develop an SCM simulator.

109.3 SCM-BSIM

To evaluate the emerging SCM and provide a better platform for developing higher level software mechanism to manage SCM, we developed a tool, called SCM-BSIM, which models SCM's characteristics.

109.3.1 Architecture

The architecture of SCM is shown in Fig. 109.1. BOCHS is responsible for simulating the hardware of a computer which uses both DRAM and SCM in its memory hierarchy. The non-volatile component is responsible for simulate the non-volatility of SCM. The endurance component accounts the access to SCM and the Device component provides different write and read delay of different SCM devices as configured by the user of SCM-BSIM.

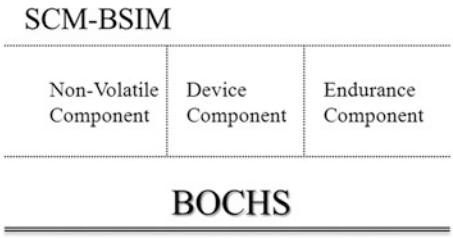


Fig. 109.1 The architecture of SCM-BSIM

109.3.2 Non-volatile Component

The non-volatile component is responsible (1) to manage virtual memory space for SCM (2) to provide interfaces for applications to access SCM, and (3) to deal with the BOCHS memory component to make SCM memory region non-volatile.

And we must guarantee that our simulator is capable of simulating SCM as universal memory and our non-volatile component is safe.

Our discussion is based on the address translation scheme which is illustrated in Fig. 109.2. SCM-BSIM operates between the guest OS's address space and the host PC's address space.

To mimic SCM's non-volatility, the first problem we have to address is that BOCHS simulates all the hardware of ×86 PC in DRAM memory of the host PC. We have to make sure that the data of the address space of SCM is non-volatile. So we dig into BOCHS's memory component and find that BOCHS has two different mechanisms for memory access. One is for prefetching instructions in CPU loop; another is for the assembly instructions simulated by BOCHS. By tracing the executions of these instructions we found that all writes fall to the same function called writePhysicalPage. And all reads fall into readPhysicalPage. In these functions BOCHS handles the logical address and perform the direct access to the host memory. So we modified it to isolate SCM's address space.

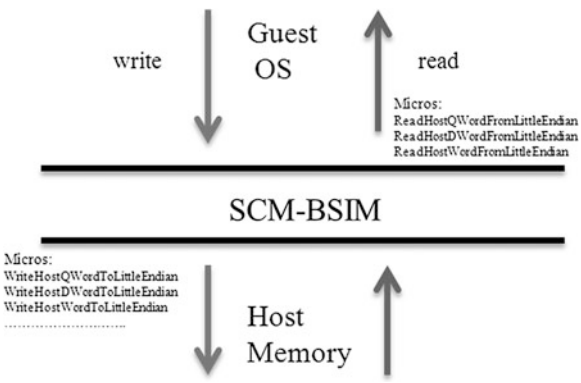


Fig. 109.2 The address translation of SCM-BSIM

To ensure our simulator is capable of simulating SCM as universal memory. We will judge every access to the logical address space until the address range of SCM is configured by user through the dialog we added to BOCHS interface.

To ensure the security of the non-volatile component, we have to guarantee the atomicity of the flush operation. Here we have two options, either to insert the flush operation into the CPU loop or to generate a synchronous event that will halt the simulation. Firstly, we inserted the flush operation case judgment into the CPU loop, it turned out to be a hazard to the performance since the CPU loop has taken around 50 % of the total execution time of BOCHS simulation [8]. Then we chose to generate a synchronous BX_SYNC_EVT_MY_FLUSH event to stall the CPU loop.

At last, we incorporate them as a button into the BOCHS simulation GUI, so if user want to test the non-volatility of our simulated SCM, it would be very convenient. Also, to further the research, we isolated the function of flushing and restricted it in a function so that further researchers can just call the function.

109.3.3 Device Component

By now, there are a lot of technologies competing to be the most viable SCM, thus the access latency are not identical. For some SCM devices, they still have long write latency compared to the DRAM memory of the host PC.

In device component, we simulate variety of SCM devices by providing different memory access latency. To handle different latency, we have to modify the memory access process of BOCHS. And since BOCHS handles read/write symmetrically so our discussion will focus on the write access here.

As mentioned before, BOCHS has two different mechanisms for memory access. One is for prefetching instruction and another is for executing instructions. According to the trace stack of BOCHS during simulation, we classified the assembly language instruction simulated by BOCHS and found that there are two different traces. The stack and data transfer instructions will call write_virtual_word while the bit, arithmetic, I/O, logic, and shift instructions will call write_rmw_virtual_word. As for the prefetching instruction mechanism, it will call write_virtual_word. Finally, we found that both the two functions mentioned above will fall to writePhysicalPage. In writePhysicalPage BOCHS will call several micros to accomplish the data copy process as depicted in Fig. 109.2.

Since we have already got access to the whole address space of BOCHS, we can catch every read/write access. Once the access latency is passed through by the GUI of SCM-BSIM, we will watch every memory access and if access is within the range of SCM. The access will be delayed.

Another issue we have to discuss here is the delay mechanism. Simply loop the host CPU will cause other thread starving which will result to performance loss. So we do a system call so that the host CPU will do a context switch which in turn solved this case. And our Device component provides latency in unit of millisecond.

109.3.4 Endurance Component

The endurance component is designed for endurance statistics gathering. To boost the research on either wear-leveling algorithms [9] or lifetime-enhancing techniques, the endurance component accounts access of SCM and gathers access distribution statistics.

In order to gather the information of the SCM wear status, first we have to discuss the granularity issue. Since some SCM devices are byte-addressable, we tried to track access to every byte but it turned out to consume excess memory.

Then we decided to track memory access at block granularity. The logical address space is divided into blocks. BOCHS will allocate memory in unit of block at runtime. As depicted in Fig. 109.2, to copy data from and to the host memory, a lot of micros are called, and along with the micros, the function BX_MEM_C::get_vector is called to get the exact block number of the desired data. So that can monitor the access of each block by keeping track of the calling of this function. We also defined an array to store the exact access count. Since the SCM address range will not be fixed until configured at runtime. The size of the array is fixed to be the number of blocks.

109.3.5 Interface

With the purpose of making SCM-BSIM easy to use, we build the interface of SCM-BSIM on BOCHS.

Similar to X-windows, BOCHS handles GUI with event mechanism, to handle the non-volatile configure command. So we trigger a synchronous event so that we can suspend the simulation of BOCHS. Then the callback FlushParamProc we defined will pass the configured parameter to the SCM-BSIM's non-volatile component. The runtime interface of SCM-BSIM is shown in Fig. 109.3.

As is shown in Fig. 109.3, for Device Component, by clicking on the toolbar, developer can specify the parameters. The read/write delay is designed for users to specify the latency of different SCM devices. And the non-volatile address range is the address range of SCM in the guest OS of BOCHS. When the configuration is done, user can push the OK button. Then SCM-BSIM will start the simulation and perform a flush immediately. Also in Fig. 109.3, for Endurance Component, we added on the toolbar, developer can get the endurance statistics at runtime of the simulation. If further more direct statistics is needed for developers, the developer can refer to the array block access data we defined.

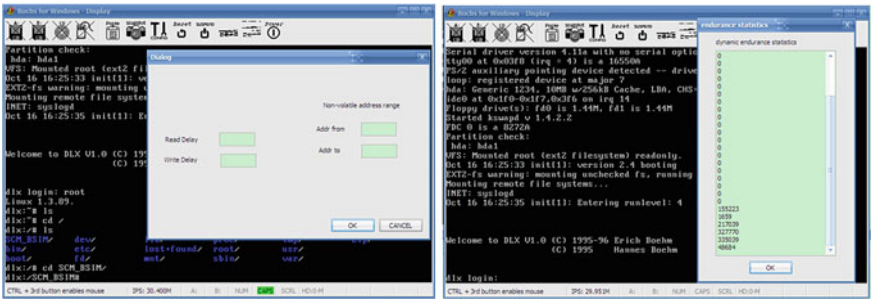


Fig. 109.3 Configure interface of parameters in device component and the runtime endurance stastics in Endurance Component

109.4 Verfication

To verify the validity of SCM-BSIM, we write specified data to the memory in the guest OS and check the consistency of the known data in the host file system. To acquire the location of the data in the host file system, we use Mnemosyne [5]. Mnemosyne implemented a non-volatile programming interface and mapped the virtual address to the physical address on SCM devices. By the map information from Mnemosyne we can get the physical address on the simulated Linux, i.e. the virtual address in BOCHS, so during the flush operation, we flush the respective data to host file system. By checking the consistency of the data specified by us, we can confirm the validation of our system.

109.5 Conclusion and Future Work

In this paper, we developed a SCM simulator called SCM-BSIM. By supporting the non-volatility feature and different latency of SCM, SCM-BSIM helps evaluation and design exploration for architecture-level research on leveraging SCM for better performance. Also, by collecting access statistics of SCM, SCM-BSIM helps developing wear balancing strategies. In the future, we plan to build a device parameter database that can be loaded at runtime to make SCM-BSIM easier to use. By combining SCM-BSIM with other OS level support, we aim to provide more simulation options for SCM-BSIM.

Acknowledgments This work is partially supported by NCET, and National Science Foundation (NSF) China under grant NO. 61272142, 61170261, 61103082, 61103193 and 61003075, and the National High-tech R&D Program of China (863 Program) under grant NO. 2012AA01A301 and 2012AA010901.

References

1. Freitas, R. F.: Storage class memory: technology, systems and applications. Proceedings of the 35th SIGMOD international conference on Management of data, ACM 985–986 (2009)
2. Wu, X., Reddy, A.L.N.: SCMFS: A file system for storage class memory. In: 2011 high performance computing networking, storage and analysis (SC'11) Seattle. Washington, USA (2011)
3. Nightingale, J.C.E.B., Frost, C., Lee, E.I.B., Coetzee, D.B.D.: Better I/O through byte-addressable, persistent memory. In: 22nd symposium on operating systems principles (SOSP'09) Big Sky, Montana, USA, ACM, pp. 133–146 (2009)
4. Coburn, J., Caulfield, A.M., Akel, A., Grupp, L.M., Gupta, R.K., Jhala, R., Swanson, S.: NV-Heaps: making persistent objects fast and safe with next-generation, non-volatile memories. In: 16th Architectural support for programming languages and operating systems (ASPLOS'11) Newport Beath, California, USA, ACM, pp. 105–118 (2011)
5. Volos, H., Tack, A.J., Swift, M.M.: Mnemosyne: lightweight persistent memory. In: 16th architectural support for programming languages and operating systems (ASPLOS'11) Newport Beath, California, USA, ACM, pp. 91–104 (2011)
6. Burr, G.W., Kurdi, B.N., Scott, J.C., Lam, C.H., Gopalakrishnan, K., Shenoy, R.S.: Overview of candidate device technologies for storage-class memory. IBM J. Res. Dev. **52**, 449–464 (2008)
7. Mihocka, D. and S. Shwartsman.: Virtualization without direct execution or jitting: Designing a portable virtual machine infrastructure. 1st Workshop on Architectural and Microarchitectural Support for Binary Translation in ISCA-35, Beijing (2008).
8. Qureshi, M.K., Srinivasan, V., Rivers, J.A.: Scalable high performance main memory system using phase-change memory technology, in 36th Annual International Symposium on Computer Architecture (ISCA'09) Austin, Texas, USA, ACM, pp. 24–33 (2009)
9. Qureshi, M. K., et al.: Enhancing lifetime and security of pcm-based main memory with start-gap wear leveling. Microarchitecture, 2009. MICRO-42. 42nd Annual IEEE/ACM International Symposium on, IEEE pp. 24–33 (2009)

Chapter 110
Model of Horizontal Technological Alliance Based on Energy Efficiency

Chunxin Yu and Qing Zhan

Abstract According to the life cycle of technological alliance, a three-stage model about energy efficiency of horizontal technological alliance is presented in this paper. The influence of cooperative technology innovation on technological alliance and the influence of operation mode of technological alliance on its energy efficiency are discussed. The necessary conditions of improving energy efficiency of technological alliance are proved by researching the relation between energy efficiency of enterprise member and energy efficiency of echnological alliance. And here are the conclusions: the ratio between differences of income and differences of cost, the operation mode of technological alliance are important factors for improving the member's energy efficiency. Energy efficiency of technological alliance decreases when member's ratio between income increment and cost increment is lower than energy efficiency of other members set. The result of it can be applied to establish technological alliance of industrial groups in China.

Keywords Horizontal technological alliance · The energy efficiency · The business model · Technological innovation

110.1 Introduction

The concept of the strategic alliance was firstly proposed by Hopland and Nigel [1]. Strategy alliance is classified into vertical alliance and horizontal alliance by Professor Michael E.Poter. In fact, more than 85 % of strategy alliances among

C. Yu (✉)
Shandong Yingcai University, Jinan, China
e-mail: gjd730210@163.com

Q. Zhan
CVIC Software Engineering Co., Ltd.(CVIC SE), Jinan, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_110,

enterprises are related to technological innovation activities. Therefore, the strategic alliance is also more popularly known as technology alliance by some foreign scholars [2].

After the twentieth century, in order to enhance the competition capacity of China's enterprises with foreign enterprises and improve the efficiency of innovation, sharing the risk of innovation and interest, several Chinese enterprises established a long-term and stable strategy alliance under the guidance of Chinese government. Strategic alliance is mainly classified into technological research cooperation alliance, industry chain cooperation alliance and technical standard cooperative alliance [3], where technological research cooperative alliance is the most common modes and its main purpose is to solve industrial generic technology problem. This alliance mode is usual horizontal alliance, which is suitable for technical innovation activities. From the standpoint of innovation theory, the efficiency of technological research cooperation alliance is studied in this paper. Based on the research results of this paper, enterprises can decide whether to establish alliance or not.

In this paper, the energy efficiency of technological alliance is defined as the ratio of actual income and cost [4–6]. By using notion of energy efficiency, member can efficiently utilize their resources. Energy efficiency is not only used as the basis of establishing technology alliance, but also convenient for comparing when member face several investment chooses. Under the guidance of the government, the affect of the technological innovation to energy efficiency of technological alliance, the model of energy efficiency, sufficient condition of improving energy efficiency and the correlation analysis between improvements of energy efficiency of the technological alliance with other key factors are provided in this paper. Result from this paper can be used to providing theoretical foundation for establishing technological alliance.

110.2 Model of Energy Efficiency

Energy efficiency is defined as ratio of income and cost in this paper, i.e. $e = E_m/E_i$. The life cycle of technological alliance is divided into three stages, which is illustrated in Fig. 110.1. In the first stage, every member which prepares to establish technological alliance respectively produce according to its own maximum energy efficiency. After the technological alliance is established, the process of technological innovation needs indeterminate time; new technology isn't applied into productions at this stage. Every member applies new technology into its own production in the third stage. According to the contract, the enterprise members produce in the second and third stage, their target is to maximize the energy efficiency of the technological alliance. Based on their own interests, every member produces according to actual market demand, because the constraint of contract no longer exists after the third stage. The energy efficiency in the second stage may be lower than the first stage; this result is from the enterprise's

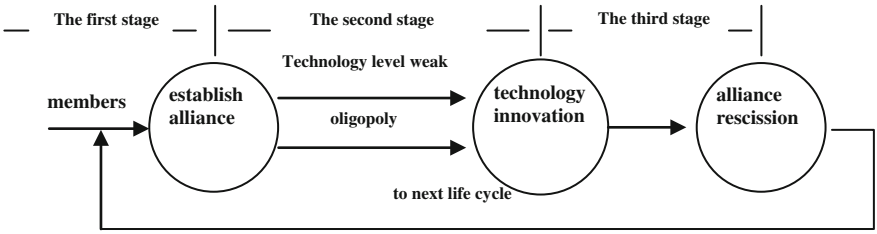


Fig. 110.1 The life cycle of technological alliance

compromise to achieve higher energy efficiency in the third stage, better technology platform and higher competitive ability (Fig. 110.1).

There are three roles: members of technological alliance $g \in U$, Chinese government C and competition enterprise D, where $U = \{1, 2, \ldots n\}$ is set of member of technological alliance, $V = U - \{g\}$ is subset after removing g from U. Technological alliance is established under the guidance of C. To decrease the dependence of D's production, according to industrial development strategy planning of a country or an area, C formulates several preferential policies.

Let E_{mg} and $E_m = \sum_g E_{mg}$ are respectively the actual income of g and technological alliance; E_{ig} and $E_i = \sum_g E_{ig}$ are respectively the actual cost of g and technological alliance; e_g and e are respectively the energy efficiency of g and technological alliance. The target of establishing technological alliance is maximization of the energy efficiency of technological alliance about member's yield vector $q = (q_1, q_2, \ldots, q_n)$ in the whole innovation cycle, as shown in Eq. (110.1),

$$e = \frac{E_m(q)}{E_i(q)} \tag{110.1}$$

where $q_g = D_g(p_g)$ is g's the market demand function when it has average technology content, p_g is price of g's production.

In order to explain our model and application environment, several hypothesis are made as following.

Hypothesis 1. The market demand function of g's production is only affected by relative average technology content of similar industry production.

Let T_g^k is impact factor of technology content, where k denotes the stage. $k\left(T_g^k\right)$ is impact factor function and it is an increasing function, where $k(1) = 1$. Market demand of g's production is $k\left(T_g^k\right)q_g$ and g arranges production according to this market demand. Let $C_g(q_g)$ is the transaction cost function.

Hypothesis 2. There are only one production using innovation technology in g. The raw material and production price are known to every member of technological alliance.

Let $s_i, p_{Ii}, i = 1, 2, \ldots, N_g$ are respectively quantity and unit price of raw material for a g's production. Let R_g is g's the technological innovation cost.

Hypothesis 3. Period of validity of innovation technology is limited. Increment of income (result of use the innovation technology) would decline over time.

Hypothesis 4. The total demand of market, the unit price of raw material and unit demand quantity are invariant in the whole life cycle of technology alliance.

110.3 The Effect of Technological Innovation to Energy Efficiency of Horizontal Technological Alliance

From the hypothesis above, in the first stage according to maximizing its own energy efficiency, $e_g^* = \max e_g = E_{mg}^1\left(q_g^*\right)/E_{ig}^1\left(q_g^*\right)$ denotes g's energy efficiency, where q_g^* is g's optimal production and $x_g^* = k\left(T_g\right)q_g$ is the actual production in this stage. $E_{mg}^1 = p\left(q_g^1\right)k\left(T_g^1\right)q_g^1$ and $E_{ig}^1 = C_g\left(q_g^1\right) + \left(\sum_{i=1}^{N_g} s_i p_{Ii}\right) k\left(T_g^1\right)q_g^1$ are respectively g's income and cost in the first stage, its feasible domain is $q_g \geq 0, g \in U, \sum_g q_g > 0$.

In the second stage, the operation mode of technological alliance can be divided into 2 types according to the position of members in the market. In the following discussion, the second superscript denotes operation modes, for example, $E_m^{2,1}$ is the actual income of the first operation mode in the second stage. According to its operation mode, the actual income and cost of the technological alliance are analyzing as following.

(1) The technological alliance increases market share by lowering price and enlarging yield. This mode is suitable for situations that the technological level of technological alliance is weaker than D. One of the targets of technological alliance is competing with D. The income and cost of technological alliance is formulated in Eqs. (110.2) and (110.3).

$$E_{mg}^{2,1} = \left(p\left(q_g^2\right) + \Delta p_g^2\right)\left(k\left(T_g^2\right)q_g^2 + \Delta q_g^2\right) \tag{110.2}$$

$$E_{ig}^{2,1} = C_g\left(q_g^2\right) + \left(\sum_{i=1}^{N_g} s_i p_{Ii}\right)\left(k\left(T_g^2\right)q_g^2 + \Delta q_g^2\right) + M_g \tag{110.3}$$

where $\Delta q_g^2 > 0$ is variation in product, $\Delta p_g^2 < 0$ is variation in price, $M_g > 0$ is the net cost of merger and acquisition. Compared with g's income and cost in first stage, the increment of g's income and cost are separately $\Delta E_{mg}^{2,1} = E_{mg}^{2,1} - E_{mg}^1$ and $\Delta E_{ig}^{2,1} = E_{ig}^{2,1} - E_{ig}^1$.

(2) Based on Cournot's oligopoly model, after establishing the horizontal technological alliance, the market monopoly power would be enlarged and the unit price of production would be increased [7]. Let $\Delta p_g^2 > 0$ is the price increment of g's production. The member$g$'s actual income and cost are separately $E_{mg}^{2,2} = \left(p\left(q_g^2\right) + \Delta p_g^2\right)k\left(T_g^2\right)q_g^2$ and $E_{ig}^{2,2} = C_g\left(q_g^2\right) + \left(\sum_{i=1}^{N_g} s_i p_{Ii}\right) k\left(T_g^2\right)q_g^2 + M_g$. Compared with g's income and cost in first stage, the increment of g's income and cost are separately $\Delta E_{mg}^{2,2} = E_{mg}^{2,2} - E_{mg}^1$ and $\Delta E_{ig}^{2,2} = E_{ig}^{2,2} - E_{ig}^1$.

$E_{mg}^{2,2}$ and $E_{ig}^{2,2}$ are special case of Eqs. (110.2) and (110.3), so the formulations of two operation models can be represent by same equations, but their feasible domains are difference. If the operation model is the first model in the second stage, then $\Delta p_g^2 < 0,\ \Delta q_g^2 > 0$, else $\Delta q_g^2 = 0,\ \Delta p_g^2 > 0$.

The technological levels don't change in the second stage, i.e., $T_a^2 = T_a^1,\ T_b^2 = T_b^1$. After applying the new technology, the technology content in the third stage is $T_g^3,\ T_g^3 > T_g^2 = T_g^1$. Otherwise, the member g would not invest in technological innovation. In general, because of market share, the market demand in previous stage is greater than or equal to it in third stage. The market demand in third stage would be decreasing from hypothesis 3. To simplify the problem, suppose the model is linear. So the actual market demand in the third stage is sum of market demand and half of the maximum of demand increment, where market demand is determined by pure technology level.

$E_{mg}^3 = \left(p\left(q_g^3\right) + \Delta p_g^3\right)\left(k\left(T_g^3\right)q_g^3 + \frac{\Delta q_g^2}{2}\right)$ is g's actual income and $E_{ig}^3 = C_g\left(q_g^3\right) + \left(\sum_{i=1}^{N_g} s_i p_{Ii}\right)\left(k\left(T_g^3\right)q_g^3 + \frac{\Delta q_g^2}{2}\right)$ is g's actual cost.

In general, after the new technologies are applied, the technological alliance would be in monopoly stage, so the unit price of their production in this stage will be higher than the previous. $\Delta p_g^3 \geq 0$ denotes price increment in the third stage. In the third stage, the increment of g's income and cost are separately $\Delta E_{mg}^3 = E_{mg}^3 - E_{mg}^2$ and $\Delta E_{ig}^3 = E_{ig}^3 - E_{ig}^2$.

After the expiration of contract, members of technological alliance go back to stage of free competition. Every member of technological alliance would maximize own energy efficiency according to the market demand. Entering the next life cycle of alliance, technological impact factor is $T_g^{1+} = T_g^3$, where the superscript $1+$ denotes the first stage of the next life cycle. $\Delta E_{mg}^{1+} = E_{mg}^{1+} - E_{mg}^3$ and $\Delta E_{ig}^{1+} = E_{ig}^{1+} - E_{ig}^3$ are separately the increment of the first stage of the current life cycle with the third stage of the previous life cycle of g's income and cost, where

$E_{mg}^{1+} = p\left(q_g^{1+}\right)k\left(T_g^3\right)q_g^{1+}$ and $E_{ig}^{1+} = C_g\left(q_g^{1+}\right) + \left(\sum_{i=1}^{N_g} s_i p_{Ii}\right)k\left(T_g^3\right)q_g^{1+}$. In the whole alliance life cycle, the increment of g's income and cost are formulated in Eqs. (110.4) and (110.5).

$$\Delta E_{mg}^1 = \Delta E_{mg}^2 + \Delta E_{mg}^3 + \Delta E_{mg}^{1+} = p\left(q_g^{1+}\right)k\left(T_g^3\right)q_g^{1+} - p\left(q_g^1\right)k\left(T_g^1\right)q_g^1 \quad (110.4)$$

$$\begin{aligned}\Delta E_{ig}^1 &= \Delta E_{ig}^2 + \Delta E_{ig}^3 + \Delta E_{ig}^{1+} = E_{ig}^{1+} - E_{ig}^1 \\ &= C_g\left(q_g^{1+}\right) - C_g\left(q_g^1\right) + \left(\sum_{i=1}^{N_g} s_i p_{Ii}\right)\left(k\left(T_g^3\right)q_g^{1+} - \left(k\left(T_g^1\right)q_g^1\right)\right)\end{aligned} \quad (110.5)$$

According to maximizing the energy efficiency of the alliance, each member of alliance would yield. Because interest subject is inconsistent, every member of the alliance hopes that own energy efficiency can be improved; else the technological alliance would not be established. So we have Hypothesis 5.

Hypothesis 5. $(q_1^*, q_2^*, \ldots, q_n^*)$ is the optimal point of maximizing Eq. (110.1); at this point, the ratio of increment of income-cost is greater than energy efficiency, that is, $\frac{\Delta E_{mg}^*}{\Delta E_{ig}^*} > \frac{E_{mg}^*}{E_{ig}^*}$, $g \in U$ where $\Delta E_{mg}^* = E_{mg}^{1+} - E_{mg}^1$, $\Delta E_{ig}^* = E_{ig}^{1+} - E_{ig}^1$ are respectively the increment of actual income and cost after g applies new technology.

Theorem 1 if Hypothesis 1 and 5 hold true, and the energy efficiency of member g's increased investment is greater other member h's energy efficiency, that is, $\Delta E_{mg}^* E_{ih}^* > E_{mh}^* \Delta E_{ig}^*, g \neq h$, then the technological innovation strictly improves energy efficiency of the technology alliance.

Proof from the conditions of Theorem 1, Eq. (110.6) can be proved.

$$\sum_{g,h,g\neq h} E_{ih}^* \Delta E_{mg}^* > \sum_{g,h,g\neq h} E_{mh}^* \Delta E_{ig}^*, \forall g, h \in U \quad (110.6)$$

Equation (110.7) holds true because of Hypothesis 5.

$$\sum_{g\in U} E_{ig}^* \Delta E_{mg}^* > \sum_{g\in U} E_{mg}^* \Delta E_{ig}^* \quad (110.7)$$

By adding Eqs. (110.6), (110.7), and (110.8) is obtained.

$$\sum_{g\in U}\sum_{h\in U} E_{ih}^* \Delta E_{mg}^* > \sum_{g\in U}\sum_{h\in U} E_{mh}^* \Delta E_{ig}^* \quad (110.8)$$

Since $\Delta E_m^* = \sum_g \Delta E_{mg}^* = \sum_g E_{mg}^{1+} - E_{mg}^1$, $\Delta E_i^* = \sum_g \Delta E_{ig}^* = \sum_g E_{ig}^{1+} - E_{ig}^1$ and Eqs. (110.8), (110.9) is obtained.

$$\frac{\Delta E_m^* + E_m^*}{\Delta E_i^* + E_i^*} > \frac{E_m^*}{E_i^*} \tag{110.9}$$

Eq. (110.10) is obtained by Eq. (110.9).

$$\frac{E_m^{1+}}{E_i^{1+}} > \frac{E_m^*}{E_i^*} = \max_{q_g \geq 0, \sum_g q_g > 0} \frac{E_m}{E_i}, g \in U \tag{110.10}$$

110.4 Conclusion

Under the guidance of C, the problem of several domestic enterprises that yield similar production and establish the horizontal technological alliance is discussed in this paper. The life cycle of alliance is divided into three stages and the energy efficiency in every stage is formulated. From the perspective of the whole life cycle, improvements of g's energy efficiency are closely related with the ratio between differences of income and differences of cost in the first stage of the adjacent life cycle and the operation mode in the second stage of alliance. Only from the perspective of energy efficiency, efficiency of the first model is lower than the second model's. According to Chinese actual situation, the feasibility of the first operation model in the second stage can be drawn.

$E_{ig}^* \Delta E_{mg}^* > E_{mg}^* \Delta E_{ig}^*$ holds true in Hypothesis 5. If $\Delta E_m^* E_i^* < \Delta E_i^* E_m^*$ is derived from $\Delta E_{mg}^* \sum_{g \neq h} E_{ih}^* < \sum_{g \neq h} E_{mh}^* \Delta E_{ig}^*$, then energy efficiency of alliance decreases. This result comes from the fact that the member$g$'s ratio of income increment and cost increment is lower than energy efficiency of subset V.

The total market demand is invariant from hypothesis 4. Because of the production commonness of technological alliance, the increasing of g's yield impacts to other member h's market share. The potential competitions among members lead to instability of the horizontal technological alliance increases. The member g obtained its interest from the technological alliance at the cost of its long-item interest obtained when g isn't in innovation. Under the guidance of C, if the former is greater than the latter, then g selects the former, because the former is more attractive than the later. Purchase discount of raw material, accelerating up the process of technological innovation and shortening the third stage of technological alliance are all methods that can make g more favour of innovation and improving the stability of the horizontal technological alliance.

References

1. Zhang, Y., Liu, Y., Li, H.: Value creation and distribution analysis of strategy alliance. J. Indus. Eng. Eng. Manag. **17**(2), 20–23 (2003)

2. Shi, Z.: Enterprise Strategic Alliance, pp. 168. ShangHai University of Finance & Economics Press, ShangHai (2001)
3. Gao, G.: Development and enlightenment of the international industrial technology innovation alliance. Sci. Technol. Dev. Res. **12**(5), 1–8 (2008)
4. Xiao, T., Sheng, Z.: Resolutions and effects of technological innovations on energy efficiency of industrial groups. J. Manag. Sci. China **4**(6), 1–5 (2001)
5. Xiao, T., Sheng, Z.: Strategy on optimal energy efficiency of the merger and acquisitions of industrial groups. Chin. J. Manag. Sci. **8**(4), 63–67 (2000)
6. Chen, L., Zhang, Z., Gu, L.: Effects of technological innovations on energy efficiency of technological alliance [J]. Chin J Manag **1**(1), 125–129 (2004)
7. Farrel, J., Shapiro, C.: Horizontal mergers: an equilibrium analysis. Am. Econ. Rev. **80**(1), 107–126 (1990)

Chapter 111
Evaluating Life-Cycle Matrix Model for Mobile Social Network

Guo-feng Zhao, Bing Li, Juan Wang and Hong Tang

Abstract Most of Mobile Social Network (MSN) sites may go through a process from emergence to disappearance, and have their own life-cycle in the market. To evaluate the life-cycle of MSN, a life-cycle model with four growth stages for MSN was proposed on the basis of BCG-like matrix in this article. Instead of employing the traditional market shares and growth rates, model in this paper is innovative for consulting two parameters which can be easily and rapidly derived from MSN user behavior pattern. Therefore, it provides a simple but effective way to identify which life-cycle stage an MSN could be, and helps choose prospective market stars for commercial purpose. Based on real click-stream data set collected from a Mobile Telecom Carrier in Chongqing, the authors evaluate the model and show the performance analyses.

Keywords Mobile social network · Life-cycle · BCG matrix · User behavior pattern

111.1 Introduction

Human, plant, animal, and other organisms all have their own life-cycle process in the nature. Therefore, we argue that Mobile social network site (MSNS) which is a special kind of products, would have its own life-cycle process in nature. During past several years, a flock of MSNSs sprouted up. However, most of them vanished with a relative transient life process from birth to death. Can we find a simple but effective way to analyze and to select some future market stars from these MSNSs for commercial purpose?

G. Zhao · B. Li (✉) · J. Wang · H. Tang
Institute of Mobile Internet Technology, Chongqing University of Posts and Telecommunications, Chongqing, China
e-mail: chengyinglib@126.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_111,

BCG matrix is perhaps the most renowned and widely used portfolio analysis method which is generally used in the case of business portfolio analysis based on the combination of two dimensions: business growth and market share. Although the BCG matrix method has little theoretical support and cannot give precise decision on which stage a market competitor could be in, it provides a helicopter view and may lead managers to make decisions that are less irrational than those they make when using unaided judgment [1]. The method is used as an analytical tool in many fields: brand marketing, service quality management [2], petrochemical strategic management [3], sensor technologies [4], etc.

In recent years, many researchers have concentrated on mobile social network (MSN). However, research on life-cycle of MSN is little. Our work was innovative: Based on BCG-like matrix, a simple but effective method was creatively proposed. It can identify which life-cycle stage an MSNS could be in and help choose better investment targets in the market. To evaluate our model, a week's user click-stream data from a WAP gateway of a Mobile Telecom Carrier in Chongqing Province was collected, and it showed the results of life-cycle stages of some popular MSNSs in China.

111.2 User Behavior Pattern on MSNS

Before building up the model to portray life-cycle of MSNS, a user's behavior pattern on Mobile Social Network will be presented. The user's behavior pattern is to profile how users accessing to the MSNS server, describes the interactions between the MSNS content server and its users from the angle of server view. As shown in Fig. 111.1, the user's behavior pattern has two layers: session level and transaction level.

- Session level

A session is a procedure that a user logins an MSN and launches browsing; then issues a series of accessing requests; and finally logs out or leaves [5]. A session is composed of several requests in time sequence generated by a user when accessing to MSNS server. Session duration refers to the time span between the first and the last request in a session.

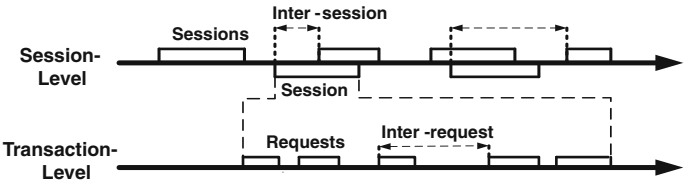


Fig. 111.1 User behavior pattern on MSNS

From the perspective of website server, inter-session refers to the time span between two successive sessions arrivals though the two sessions may belong to different users because sessions from different users may overlap in time axis. As is mentioned in [6], shorter average inter-session (AIS) reflects that the content server is more active. Generally speaking, the more users are attracted by the MSNS server, the more active the server is. Therefore, shorter AIS implicate more users an MSNS possessed.

- Transaction level

A transaction process consists of several successive requests in a session. Similar to inter-session, inter-request refers to the time span between two successive requests during a session. The shorter average inter-request (AIR) implicates the users are more active, when they are conducting in their sessions [6]. In other words, the shorter AIR indicates the users have more interest in the content which provided by the MSNS; moreover, content is king in Internet world.

In next Section, a life-cycle model of MSNS based on AIS and AIR will be introduced.

111.3 Life-cycle Model of MSNS

As shown in Fig. 111.2, the life-cycle process of an MSNS can be classified into 4 different stages: Infancy, Adolescence, Adulthood and Aging, while each stage represents a different profile of risk and return.

Infancy. This stage starts with the release of an MSNS in Mobile Internet market and hopefully ends with stepping into its adolescence. It is characterized by the uncertainty of the acceptance. In most cases, there is well prospective growing market share, but the market share is still small.

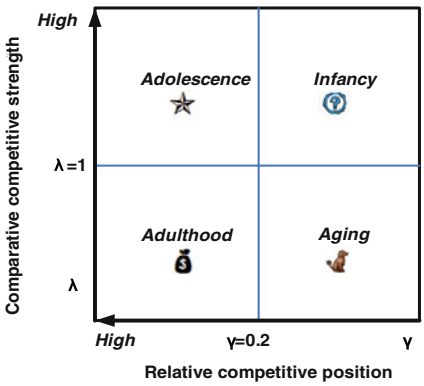


Fig. 111.2 BCG-like matrix model

Adolescence. Adolescence is the best stage of the life-cycle: good perspectives about the market share growth. Although it requires a great effort for supplying the greedy need of investment, providers would be rewarded by the coming benefits.

Adulthood. In adulthood stage, there is less prospective market share growth; however, there are much more benefits than costs, because by then the MSNS is getting its high benefits ratio. Therefore, more actions should be taken to prolong the period of this stage.

Aging. In this stage, an MSNS should be merged to a new one or be abandoned because it leads to great costs and less benefits.

Basically, the BCG matrix was constructed based on market share and market growth. However, the use of the BCG matrix is often inhibited by difficulties in measurement of market growth rates and relative market shares. Nowadays, due to the immature development in the MSNS market, it is hard to get the market share and the market growth rate for an MSNS by normal market investigation means. As our model shown in Fig. 111.2, we employ relative competitive position γ (competitive position relative to that of its largest competitor) to represent the market attractiveness, and comparative competitive strength λ (competitive strength relative to the average of all competitors) to reveal the compete-ability of an MSNS within the market.

In the business or product market, the annual market growth rate of a brand above 10 % (absolute value) is considered high, so the cut-off point on vertical axis is usually chosen as 0.1; and the threshold on horizontal axis is experientially set to 0.2 in many works. However, in our model, the parameter λ means a relative ratio to the average, and its value above 1.0 indicates a stronger compete-ability than that of most MSNSs in the market. Therefore, 1.0 (the average compete-ability) is chosen as the cut-off point on vertical axis and 0.2 as that on horizontal axis as previous works did.

Then we present how to obtain the two parameters, i.e. λ and γ. As described in Sect. 111.2, AIS and AIR are respectively close related to the activeness of the MSNS server and interestingness of users for the content on the MSNS server. Considering the negative correlation between AIS and server activeness which means the shorter AIS is, the more active a content server is and more users may a MSNS possess. While the competitive position of a MSNS is positively correlated with the number of users it had, so we argue that the AIS can mirror the competitive position of a MSNS quite well.

Similarly, there exists a negative correlation between AIR and users' interestingness. When users are conducting in their accessing sessions, shorter AIR indicates that the users are more active and the contents on the MSNS server are more attractive that implies the MSNS may have stronger competitive strength. Therefore, the AIR can mirror the competitive strength of an MSNS.

Assume that any one MSNS within an observation period T, it has received N sessions. Suppose ith session arrived at time ti, then the inter-session between ith session and (i + 1)th session is ti + 1 – ti. The AIS within period T for the MSNS can be written as

$$t_s = \frac{\sum_{i=1}^{i=N-1} (t_i + 1 - t_i)}{N - 1} \tag{111.1}$$

Similarly, assume that in session si, it had Mi requests and the jth request arrived at time tj, so the AIR for all N sessions within period T for the MSNS can be written as

$$\overline{t_q} = \frac{\sum_{i=1}^{i=N} \sum_{j=1}^{j=Mi-1} (t^i_j + 1 - t^i_j)}{\sum_{i=1}^{i=N} M_i - 1} \tag{111.2}$$

Hence, two parameters $\alpha = 3600/\overline{t_s}$ and $\beta = 3600/\overline{t_q}$ is obtained. Parameter α is the average number of sessions per hour arrived at an MSNS server during period T, and β is the per hour average number of requests within a session. In our notations, we denote by αi the competitive position and by βi the competitive strength of the ith MSNS respectively.

Therefore, the other two parameters, i.e. λ and γ, can be obtained by following formulas.

$$\gamma_{\mathrm{i}} = \alpha_i / \max(\alpha 1 \alpha 2, \cdots \alpha N) \tag{111.3}$$

$$\lambda_i = \beta_i \bullet N / \sum_{j=1}^{N} \beta_j \tag{111.4}$$

For the ith MSNS, γi denotes its relative competitive position which means its competitive position relative to that of the market leader (whose $\gamma = 1$). Also, in Eq. (111.4), λi denotes its comparative competitive strength which means its competitive strength relative to the average while the average indicates $\lambda = 1$.

111.4 Evaluation

111.4.1 Dataset Description

Our raw data was collected from a WAP gateway of one main Mobile Telecom Carriers in Chongqing Province, where more than 30 million people dwell, for a week from Apr. 5, 2010 to Apr. 11, 2010. Before a user can access MSNS server, his/her mobile device must connect to the WAP gateway first through the wireless cellular network. Moreover, all requests and responses between the MSNS server and its users are transferred and logged by the WAP gateway. Additionally, the gateway dynamically assigns client IP address for each connection.

Table 111.1 Statistics of the click-stream data for MSNSs

MSNS	# Of users	# Of sessions	# Of requests	Uplink(MB)	Downlink (GB)
Qzone	9,142	294,383	2,089,745	374.77	9.6953
Kaixin001	868	29,851	2,14,699	1.08	1.3052
Renren	997	6,243	1,35,253	1.46	0.3841
51	169	2,755	13,974	0.21	0.0734
Total	11,176	333,232	2,453,671	377.52	11.458

The log file collected has a total size of 130 GB and contains 17,316,616 records of 80,690 individual users. An record includes the information of Time, Calling Number, Client IP Address, User Agent, URL, Content Type, Domain, In-Status, Uplink, Downlink length, etc. Therefore, the new session will be identified by the change of client IP address while the user discriminated by the calling number. Together with other items, e.g. URL, users' requests can be resolved.

In China, the records of several sample MSNSs was extracted, for example, *Qzone*, *kaixin001*, *51,* and *renren*. The detailed statistics for these MSNSs is shown in Table 111.1. From the viewpoint of individual users having accessed in the week, *Qzone* had the majorities of users, and *51* had a little. *renren* had more users than those of *kaixin001*. Therefore, it may conclude that *renren* would be the possible future market star compared to *kaixin001*. However, we wonder if that would be true.

111.4.2 Performance Evaluation

Based on the dataset, for each sample MSNS, (1) with observation period $T = 1$ week, first we computed AIS and AIR; (2), then we obtained four parameters – α, β, γ and λ, as shown in Table 111.2.

To identify the corresponding life-cycle stage for each sample MSNS, the two parameters(γ, λ) were calculated according to formulas (1)–(4) as shown in Table 111.2. Using parameter values in the table, we have the diagram shown in Fig. 111.3. We find: *kaixin001* is at the stage of infancy; *renren* and *51* both suffer their aging; but *Qzone* is enjoying its adulthood. The results show that *kaixin001*

Table 111.2 Parameters for identifying life-cycle of MSNSs

MSNS	*Qzone*	*Kaixin001*	*Renren*	*51*
AIS (s)	2.61	20.26	80.68	219.08
AIR (s)	18.33	5.37	12.52	12.58
A	1,378.8	177.6	44.4	16.2
β	202.2	670.2	288.6	286.2
γ	1	0.13	0.03	0.01
λ	0.56	1.85	0.80	0.79

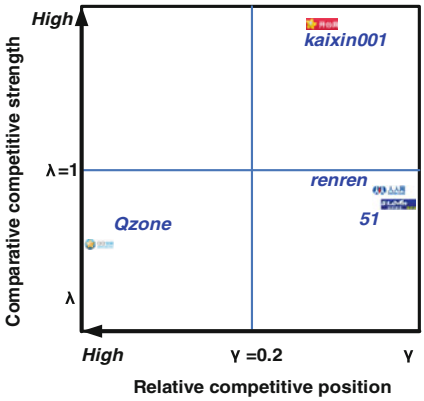


Fig. 111.3 Identified life-cycle stages of sample MSNSs

has the most powerful competitive strength ($\lambda = 1.85$) than others, and its competitive position ($\gamma = 0.13$) is second only to the market leader. Therefore, *kaixin001* will be chosen as the future market star by our model.

Our dataset does not embody the click-stream data of whole nation-wide users but that of a province instead; however, the province has more than 30 million people, we conjecture that the evaluation results derived from so huge people sample can reveal [7] the real life of the MSNSs.

111.4.3 Results Analyses

In this part, some deep analyses of the effectiveness of the model will be given. First, it needs to analyze unique visitors during the week, and Fig. 111.4 shows the percentage of users of each MSNS. It is shown that the users having visited *Qzone* are 81.8 % of the total; we know it dominates the MSNS market. Furthermore, traffic volume, including uplink and downlink, from *Qzone* is about 85 % of the total, as shown in Table 111.1. It is proved that *Qzone* is the market leader.

However, compare the average logins and requests per user, interestingly shown in Fig. 111.5, it shows that the average login number and average request

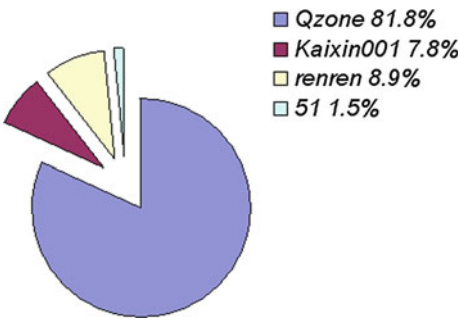


Fig. 111.4 Percentage of unique visitors

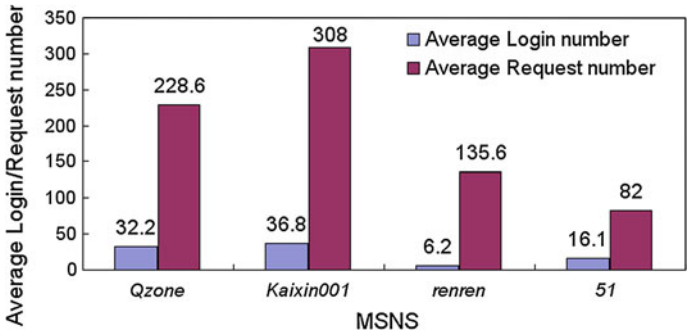


Fig. 111.5 Average login/request number per user of MSNSs accessing to each MSNS

number per user on *Qzone* are both less than those on *kaixin001*. We think the contents on *Qzone* are less attractive to visitors than those on *kaixin001*, and *Qzone*'s competitive strength would be weaker than that of the latter, in spite of its overwhelming users.

As shown in Fig. 111.3, *Qzone* is in its adulthood while *kaixin001* in its infancy. That means the *kaixin001* is more attractive, but has lower market share at present. Probably, *kaixin001* would soon step into its adolescence if it can acquire great number of users with its high competitive strength. Maybe it is a good chance to invest *kaixin001* now.

Historically, *Qzone*, *renren* and *51* communities were all launched in 2005, while *kaixin001* in 2008. The *renren* and *51* lag far behind *Qzone* in the market now, and even have not surpassed the *kaixin001*. We can see from Fig. 111.4 that they have lower competitive positions and weaker competitive strength from Fig. 111.5, though they both have their loyal users.

In Fig. 111.3, *renren* and *51* are both undergoing their aging. The *renren* has better competitive position than *51*, but the latter has nearly identical competitive strength although it has much less users. However, it is better not invest to any of them at present.

111.5 Conclusion

One of the best-known business portfolio models to identify the "yesterday's has-beens" and "tomorrow's breadwinners" is BCG growth-share matrix model. It motivates researchers to use the BCG-like matrix to model the life-cycle of MSNS. The model provides a simple but effective method that can help one choose some prospective market stars for commercial purpose. Instead of employing the traditional market shares and growth rates which difficult to obtain, researchers model the life-cycle using two parameters derived from user behavior pattern on MSNS. The method only needs to collect and analyze measurement data.

Therefore, it is very easy to apply. If the method can be used multiple times with longer observation period, the helicopter view of the market will be more precise. In addition, the method can be extensively applied to evaluate different websites, blogs, etc. Future work will be concentrated on analyzing differences of user habits in different phases of life-cycle.

Acknowledgments This work is supported by National Basic Research Program of China (2012CB315806) and Natural Science Foundation of Chongqing (CSTC.2009BA2089 and CSTC.2012JJB40008).

References

1. Scott Armstrong, J., et al.: Effects of portfolio planning methods on decision making: Experimental results. Int. J. Res. Mark. **11**(1), 73–84 (1994)
2. Ioana, A., et al.: Analysis of service quality management in the materials industry using the bcg matrix method. Amfiteatru. Econ. **11**(26), 270–276 (2009)
3. Al-Sharrah, G.K., et al.: Planning an integrated petrochemical business portfolio for long-range financial stability. I.E.C.R, **41**(11), 2798–2804 (2002)
4. Singh, J.P.: Development trends in the sensor technology: A new bcg matrix analysis as a potential tool of technology selection for a sensor suite. IEEE Sens. J. **4**(5), 664–669 (2004)
5. Schneider, F., Feldmann, A., Krishnamurthy, B., Willinger, W.: Understanding online social network usage from a network perspective. SIGCOMM on Internet Measurement Conference, pp. 35–48 (2009)
6. Benevenuto, F., et al.: Characterizing user behavior in online social networks. SIGCOMM on Internet Measurement Conference, pp. 49–62 (2009)
7. Gvarmati, L.: Measuring user behavior in online social networks. IEEE Netw. **24**(5), 26–31 (2010)

Part VI
Management Science

Chapter 112
Regulation and Environmental Innovation: Effect and Regional Disparities in China

Qingjiang Ju, Tianli Feng and Ya Ding

Abstract Based on improved Griliches-Jaffe knowledge production function, this paper analyzes the effect of government regulation on environmental innovation by employing a panel dataset that covers 30 Chinese provinces from 1998 to 2006. We find that the regulation pressure (as measured by the investment in the treatment of industrial pollution) has significant and positive impact on environmental innovation (as measured by environmental patent applications) at the national level. We also find that there are distinct regional disparities between effects of environmental regulation on innovation. Only the regulation pressure in eastern China has positive and statistically significant impact on environmental innovation, while the regulation pressure in western and central China has insignificant effect on environmental innovation. More importantly, we show that the regional innovation conditions such as innovation input, export pressure, economic growth rate, and educational expenditure share are more important factors to affect environmental innovation than environmental regulation.

Keywords: Environmental regulation · Environmental innovation · Patent · Regional disparity

112.1 Introduction

As a developing country in process of industrialization and urbanization, China is facing severe challenges of resource shortages and environmental degradation. Chinese government has been speeding up the strategic transition of environmental protection. In addition to traditional administrative measures, legal, economic, and

Q. Ju (✉) · T. Feng · Y. Ding
School of Management and Economics, University of Electronic Science and Technology of China, Chengdu, China
e-mail: juzzy@uestc.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_112,

technical measures are also adopted. The industrial firms are facing increasing pressure from environmental regulations. The total investment in the treatment of environmental pollution reached 665.42 billion yuan in 2010, about five times higher than that in 2001 (China Statistical Yearbook 2011).

The technical measures are important to alleviate the contradiction between economic development and environmental protection. The advanced environmental technologies are helpful for firms to realize the win–win. Many researchers have studied the effects of environmental regulation on innovation and found different results in different countries. Based on these research methods, this paper uses Chinese panel data to study the relationship between environmental regulation and technological innovation at national and regional level respectively. This research contributes to the identification of influencing factors on environmental innovation in China, and our results could help improve government policies aiming at enhancing the effectiveness of environmental regulation on innovation.

112.2 Literature Review

The relationship between environmental regulation and technological innovation was firstly discussed by Michael Porter. He suggested that environmental regulation may have a positive effect on the performance of domestic firms relative to their foreign competitors by stimulating domestic innovation [1, 2]. Lanjouw and Mody [3] used a patent data set of US, Japan, and Germany from 1972 to 1986, and found a correlation between pollution abatement expenditures and innovation. Jaffe and Palmer [4] used a panel data set for US manufacturing industries and found that higher lagged abatement costs would lead to higher levels of R&D expenditures but have no relations with patents. Also with US panel data, Brunnermeier and Cohen [5] found that environmental innovation responded to increases in pollution abatement expenditures, while increased monitoring and enforcement activities related to existing regulations did not provide any additional incentive to innovation. Based on new institutional economics and resource-based theory, Pascual et al. [6] studied the effects of environmental policies on corporate innovation and competitiveness. Yarime [7] studied the effects of different policies on innovation by case studies in the Chlor-Alkali Industry in Japan and Europe.

In China, Shen and Liu [8] used panel data of China during 1992–2009 to analyze the relationship between environmental regulation and technological innovation, and used nonlinear threshold panel model to study "threshold effect" between environmental regulation and innovation. Wang and Wang [9] employed China's eastern and central regional panel data of 1999–2007 to study the relationship between environmental regulation and technological innovation.

One major limitation of these empirical studies is that most of them measured innovation by using overall R&D or patents, and did not use environmental innovation data. It may lead to biased empirical results. This paper aims to

examine the relationship between regulation and environmental innovation by using regional panel data in China. Regional disparities of regulation effect are also considered.

112.3 Model Variables and Data

Knowledge production function is first proposed by Griliches in 1979 to study the contribution of R&D and knowledge spillover to the growth of productivity [10]. Jaffe [11] modified this model, extended its application, and made it an important empirical tool and framework to study knowledge production, innovation, and regional innovation. The improved Griliches-Jaffe knowledge production function in C-D form is as follows:

$$Q = AK^{\alpha}Z^{\beta} \tag{112.1}$$

where Q denotes innovation output, K denotes innovation inputs, and Z denotes other economic and social factors affecting innovation. Based on the study of Brunnermeier and Cohen [5], we put forward the following reduced equation:

$$\log(PATT_{i,t}) = \alpha_{i,t} + \beta_1 \log(INVT_{i,t}) + \beta_2 \log(RND_{i,t}) + \beta_i \log(X_{i,t}) + \mu_{i,t} \tag{112.2}$$

where i denotes region, t denotes time. PATT denotes the environmental innovation while INVT denotes the environmental regulation pressure. RND is the expenditure on R&D, denoting innovation input. X is a vector of control variables capturing the possible influence on innovation activities. $\alpha_{i,t}$ is intercept term reflecting unobservable regional heterogeneity, $\mu_{i,t}$ is a residual error term capturing all other effects. β is regression parameter reflecting the effects of independent variables on environmental innovation. Based on Griliches-Jaffe knowledge production function, the variables are defined as in Table 112.1.

By comparing different indicators, Acs [12] found that it was reliable to use patents to measure innovation. Under Chinese patent system, the data of patents in different regions are comparable and can reflect innovation level. Unlike previous studies, we use environmental patent applications (PATT) as a proxy for innovation of environmental technology. The data come from State Intellectual Property Office of China. By setting "abstract" as search field, we use "environmental pollution" and "environmental protection" as keywords to search for environmental patents. The time and region of a patent are determined by its application data and principal applicant address respectively. The Province of Tibet is not included for missing data. A total of 27,888 environmental patents are identified for 30 provinces of China over the period of 1998–2006.

The variable of environmental regulation is measured by the investment in the treatment of industrial pollution (INVT). Unlike the standards of pollution reduction that may not be fully enforced, this indicator can directly reflect policy

Table 112.1 Variables and definitions

Variable	Measuring indicators and definitions	Variable symbol
Environmental innovation	Number of environmental patent applications	PATT
Environmental regulation	Investment in the treatment of industrial pollution	INVT
Innovation input	Intramural expenditure on R&D	RND
Export intensity	Ratio of total value of exports by location of exporters to gross regional product	EXPT
Economic growth rate	Growth rate of gross regional product	GRP
Educational intensity	Ratio of educational investment to gross regional product	EDUR
Region dummy	D1 = 1 if the sample is from eastern region, 0 otherwise	D1
Region dummy	D2 = 1 if the sample is from central region, 0 otherwise	D2

pressure. In practice, an increase of investment in the treatment of industrial pollution indicates a greater level of pressure in reducing pollution, which may stimulate more innovation activities and bring about more environmental patent applications. So the coefficient of INVT, β_1, is expected to be positive. The data come from *China Statistical Yearbook* and are converted to 1998 constant price using price index for investment in fixed assets by region.

In knowledge production function, innovation inputs are classified into two parts, labor and capital. It is hard to find the input data of environmental innovation, so they are replaced by the full-time employed R&D personnel and intramural expenditure on R&D. However, the correlation coefficient of these two variables is 0.95, therefore only the intramural expenditure on R&D (RND) is included. We expect a positive sign of β_2.

Following the industrial organization literatures on innovation, we also include three important control variables affecting innovation activities. The variable EXPT takes into account the influence of international trade on environmental innovation [5]. Compared with domestic markets, foreign markets are more competitive and the consumers demand more green products, which may promote environmental innovation to enhance export competitiveness. Thus, we expect a positive sign on the coefficient of EXPT. The data of EXPT is obtained from *China Statistical Yearbook*. The variable GRP measures the effect of economic growth rate on innovation. It may reflect the environment and capability of regional innovation, so we expect its coefficient to be positive. The data of GRP is obtained from *China Statistical Yearbook*. The variable EDUR measures the investment intensity on human capital in a region and higher educational intensity could improve a region's innovation potential. So its coefficient is expected to be positive. The data of EDUR is obtained from *China Educational Expenditure Statistical Yearbook*.

Table 112.2 Descriptive statistics for all variables

Variable	Measurement	Mean	Standard deviation	Maximum	Minimum
PATT	Number	103	141.29	1132	1
INVT	10,000 yuan	80,657	85,459	5,08,986	804
RND	10,000 yuan	4,53,577	6,25,298	3,942,368	6,757
EXPT	Ratio	15.75	18.97	92.93	2.24
GRP	Ratio	10.87	2.39	23.80	5.10
EDUR	Ratio	4.66	1.51	11.72	2.56

To analyze the regional effect of regulation on environmental innovation, we include two regional dummy variables D1 and D2, with D1 representing eastern region and D2 representing central region. By adding the mediating variables of INVT multiplying by D1 and D2 respectively, we can estimate the different effect of regulation among the eastern, central, and western regions.

112.4 Results

Table 112.2 presents the descriptive statistics across all regions and years for each variable used in our study. The data cover 30 provinces over the 9-year period 1998–2006 and have a total of 270 observations.

We exploit the time-series and cross-sectional nature of the data by using panel data estimation techniques and conduct multivariate regression analysis with Eviews 5.1. Main regression results are reported in Table 112.3. It is assumed that regional heterogeneity varies randomly across regions, so the random effects models (Model 1–3) are estimated. A different method is used in Model 4 to test the robustness of our estimation. As a whole, the fitness of the four models is good, with adjusted R^2 higher than 0.75 and Pseudo-R^2 of 0.93.

We first estimate the effect of expenditure on R&D (RND) and regulation (INVT) on environmental innovation in Model (1). The results indicate that in addition to fundamental innovation input, the pressure under environmental regulation, measured by the investment in the treatment of industrial pollution, has statistically and significantly positive effect on environmental innovation.

Three control variables are considered in Model (2). The coefficient on INVT is still 0.13, significant at the 1 % level. Other things held constant, it means that for each 1 % increase of investment in the treatment of industrial pollution, the number of environmental patent applications will grow by 0.13 %. The coefficients on GRP and EDUR in Model (2) are significant at the 1-percent level, consistent with our expectation. But the EXPT coefficient is insignificant. Our explanation is that although the export trade has greatly increased as a result of the implementation of reform and opening policies of China since 1978, the regional disparity of export intensity is large. As shown in Table 112.4, this indicator (EXPT) of eastern provinces is about 6 times higher than that of central and

Table 112.3 Regression results

Independent variable	Dependent variable: patent (PATT)			
	Model (1) Random effects	Model (2) Random effects	Model (3) Random effects	Model (4) Negative binomial
Constant	−8.09(−19.28)***	−9.50(−20.51)***	−8.96(−13.39)***	−8.08(−16.27)***
INVT	0.13(2.79)***	0.13(2.91)***	0.05(0.72)	0.25(6.23)***
RND	0.87(20.92)***	0.75(15.93)***	0.72(14.25)***	0.59(18.28)***
EXPT		0.07(1.00)	0.13(1.27)	0.03(0.79)
GRP		0.64(4.26)***	0.65(4.15)***	0.80(4.85)***
EDUR		0.84(4.75)***	1.04(5.22)***	0.16(1.16)
D1			−1.48(−1.71)*	
D2			0.10(0.10)	
INVT*D1			0.14(1.78)*	
INVT*D2			0.03(0.30)	
Adjusted R^2	0.76	0.78	0.78	–
F-statistic	416.62	194.24	108.83	–
LR index (Pseudo-R^2)	–	–	–	0.93

Note a. D1, D2 in Model (3) and INVT in Model (4) are not in logarithmic form
b. The regression coefficients are before the brackets and t-statistics are in brackets
c. ***, **, * indicate significance at levels of 1, 5 and 10 % respectively

Table 112.4 Comparison of the mean of indicators in different regions

Indicator	Measurement	National total	Eastern	Central	Western
PATT	Number	103	190	76	37
INVT	10,000 yuan	80,657	1,31,131	68,508	39,018
INVT Intensity	Ratio	0.203	0.198	0.180	0.226
RND	10,000 yuan	4,53,577	8,80,155	2,64,208	1,64,723
EXPT	Ratio	15.75	32.87	5.83	5.80
GRP	Ratio	10.87	11.59	10.38	10.51
Educational Expenditure	10,000 yuan	1,783,065	2,717,167	1,700,837	1,001,159

Note a. Absolute indicators are adjusted to 1998 constant price
b. INVT intensity is the ratio of investment in the treatment of industrial pollution to gross regional product

western provinces. So at the national level, it may be hard to observe the positive impact of EXPT on innovation. By adding the mediating variables of EXPT multiplying by D1 and D2 respectively, it shows that the EXPT of eastern region has significant and positive effect on environmental innovation, while the EXPT of both central and western region has no significant impact on innovation (it is not reported in the Table 112.3 to save space). This is consistent with our expectation.

Considering the dependent variable, the number of environmental patent applications is a non-negative integer which is helpful to use count data model to test the robustness of estimation results [13]. In our study, the raw patent count

data are highly over-dispersed with a sample mean of 103 and a variance of 19,964, so it is suitable to choose the negative binomial model. As shown under Model (4) in Table 112.3, all other regression results are consistent with that under Model (2) except the variable EDUR.

In Model (3), the estimated coefficient on INVT*D1 is positive and significant at the 10 % level, while the coefficients on INVT and INVT*D2 are positive and insignificant. It means that there are huge regional disparities in the relationship between regulation and environmental innovation in China. The regulation has played a statistically significant role in environmental innovation in eastern China. In contrast, the regulations of western and central regions have insignificant effect on environmental innovation. As shown in Table 112.4, the investment intensity of pollution treatment in eastern region is not high, but the data of EXPT, GRP and educational expenditures that capture regional innovation conditions are better than the other two regions. In contrast, the regulation pressure in western region is the highest, but unfavorable conditions hinder the process of environmental innovation.

112.5 Conclusion and Discussion

Based on improved Griliches-Jaffe knowledge production function, this paper analyzes the relationship between the regulation and environmental innovation by using a panel dataset of 30 Chinese provinces from 1998 to 2006. It is found that while other things held constant, environmental regulation (as measured by the investment in treatment of industrial pollution) have significant and positive impact on environmental innovation at the national level, which is consistent with our expectation. At the region level, we further find there are huge regional disparities in the effect of government regulation on environmental innovation. Only the regulation pressure in the eastern region of China has positive and statistically significant impact on environmental innovation, while the regulation pressure in western and central China have insignificant effect on environmental innovation. Our results also suggest that regional innovation conditions, such as innovation input, export pressure, economic growth rate and educational investment intensity, have greater effect than regulation pressure on environmental innovation. Therefore, in order to improve the innovation effect of regulation, different measures should be taken according to regional background and conditions.

Acknowledgments This work is supported by National Social Science Foundation of China (12CJL040) and Humanities and Social Science Youth Foundation, Ministry of Education of China (10XJC630004).

References

1. Porter, M.E.: America's green strategy. Sci. Am. **264**(4), 168 (1991)
2. Porter, M.E., Van der Linde, C.: Toward a new conception of the environment competitiveness relationship. J. Econ. Perspect. **9**(4), 97–118 (1995)
3. Lanjouw, J.O., Mody, A.: Innovation and the international diffusion of environmentally responsive technology. Res. Policy **25**, 549–571 (1996)
4. Jaffe, A.B., Palmer, K.: Environmental regulation and innovation: a panel data study. Rev. Econ. Stat. **79**(4), 610–619 (1997)
5. Brunnermeier, S.B., Cohen, M.A.: Determinants of environmental innovation in US manufacturing industries. J. Environ. Econ. Manag. **45**, 278–293 (2003)
6. Pascual, B., Liliana, G., Andrea, F., et al.: Can institutional forces create competitive advantage? an empirical examination of environmental innovation. IESE Business School Working, pp. 723 (2007)
7. Yarime, M.: Promoting green innovation or prolonging the existing technology: regulation and technological change in the Chlor-Alkali industry in Japan and Europe. J. Ind. Ecol. **11**(4), 117–139 (2007)
8. Shen, N., Liu, F.: Can intensive environmental regulation promote technological innovation? Porter hypothesis reexamined. China. Soft. Sci. **4**, 49–59 (2012)
9. Wang, G., Wang, D.: Porter hypothesis, environmental regulation and enterprises technological innovation: the comparative analysis between central China and eastern China. China. Soft. Sci. **1**, 100–112 (2011)
10. Griliches, Z.: Issues in assessing the contribution of R&D to productivity growth. Bell J. Econ. **10**, 92–116 (1979)
11. Jaffe, A.B.: Real effects of academic research. Am. Econ. Rev. **79**(5), 957–970 (1989)
12. Acs, Z.J., Anselin, L., Varga, A.: Patents and innovation counts as measures of regional production of new knowledge. Res. Policy **31**, 1069–1085 (2002)
13. Fritsch, M., Slavtchev, V.: Universities and innovation in space. Ind. Innov. **14**(2), 201–218 (2007)

Chapter 113
The Organizational Innovation Path Formation Mechanism of Innovative-Oriented Enterprises Based on Effect Elements

Peng Wang and Chunsheng Shi

Abstract The effective implementation of the organizational innovation path is the key element to promote the enterprise performance value. But how to locate the accurate path for organizational innovation within innovative-oriented enterprises is the urgent issue. This paper is based on the effect elements of the innovative-oriented enterprises organizational innovation path formation period; it constructs the theoretical model of the effect elements impact on the formation of the innovative-oriented enterprises organizational innovation path. And then it uses the partial least squares structure equation model (PLS-SEM) method to verify the model path and hypothesis. The research can fulfill the organizational innovation theories and provide the guidance for the enterprise to precede the organizational innovation path activities.

Keywords Organizational innovation path · Innovative-oriented enterprises · Formation mechanism · Effect elements · Partial least squares structure equation model (PLS-SEM)

113.1 Introduction

The organizational innovation can promote the enterprise core capabilities, at the same time it can also promote the enterprise performance value. So the formation of the organizational innovation path has the great meaning for promoting the enterprise performance. The scholar Chen Chumming thought the main body of the organizational innovation path is path gateway, path character, and the path foundation. The effect elements of the organizational innovation path forming

P. Wang (✉) · C. Shi
School of Management, Harbin Institute of Technology, Harbin, Heilongjiang, China
e-mail: 18746040895@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_113,

include the resource elements, strategy destination and the nation environment [1]. The scholar Yang thought the organizational innovation path had the evaluations character; he concluded three evolution types of the organizational innovation path evolution [2]. The scholar Wang Yamuna use structure equation model to verify the hypothesis of the organizational innovation path forming mechanism, but he didn't give concrete analysis of the forming effect elements [3]. The scholar Wong T.C thought we should take stable and dynamic perspective to analysis the forming of the organizational innovation path [4]. The scholar Lin Ming thought the innovation conscious, innovation environment and the innovation performance are the main influence factors of the organizational innovation path forming [5]. And after the above scholars research, the scholar Shi Chunsheng suggest the strong force of the organizational innovation path forming is the innovation conscious because the innovation conscious can push the technical innovation and other innovation behavious forward, then promote the enterprise economic performance [6]. Therefore, according to the above discussion, we can conclude that the innovative enterprises organizational innovation is guiding by the technology innovation, and it takes the system innovation as the hypothesis, then it considers the management innovation as the protection, and it takes the continued innovation as the method and its final destination is to promote the enterprise performance value. But our nation implementing the organizational innovation path efficiency is very low though the organizational innovation delay character can have the influence on the innovation efficiency, but with the complex changing of the effect elements and it's very difficult for the enterprise to locate the accurate organizational innovation path, these embarrassing elements lead to more risk for the enterprise to proceed the organizational innovation path activities. This research is based on the organizational innovation theories, and then constructs a theoretical model of the effect elements impact on the formation of the organizational innovation path. We hope through the following research, we can provide realistic suggestions to innovative-oriented enterprise managers.

113.2 The Construction of the Innovative-Oriented Enterprises Organizational Innovation Path Formation Effect Elements

113.2.1 The Definition of the Organizational Innovation

The definition of the organizational innovation is the enterprise promotes the organization economic efficiency which based on the behavior scientific and the organization destination. The main body of the organizational innovation includes the organizational innovation elements of the organization level, organizational innovation elements of the individual level and organizational innovation elements of the group level. And the organizational innovation elements of the organization

level include the structure-oriented organizational innovation, culture-oriented organizational innovation, strategy-oriented organizational innovation, and the knowledge-oriented organizational innovation [7]. The organizational innovation elements of the individual level include the innovation quality, innovation tendency and the innovation management ability. The organizational innovation elements of the group level include group construct elements, group innovation environment and the group cohesive force. And the above elements are the enterprise innovation behavior. The behavior can promote the organization efficiency.

The origination of the organizational innovation is the enterprise use the manage ability to enhance the innovation performance. The Fig 113.1 is the effect elements of the innovative-oriented enterprises organizational innovation path forming.

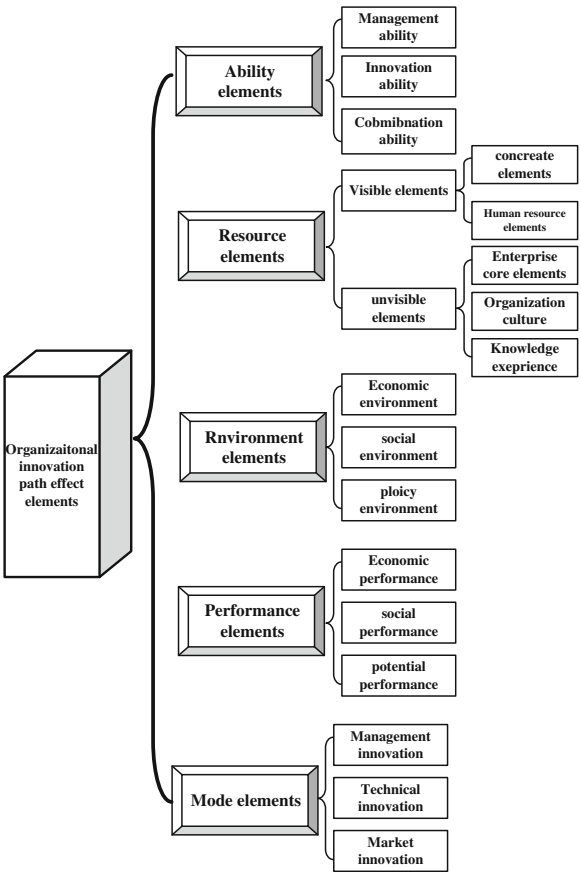


Fig. 113.1 The effect elements of organizational innovation path forming mechanism

113.3 Hypotheses and the Theoretical Model Construction

Through the above discussion, we have the following relate hypothesis.

H1: the better of the ability elements, the better of the innovation behavior;

H1a: the better of the ability elements, the better of the technical innovation value;
H1b: the better of the ability elements, the better of the management innovation value;
H1c: the better of the ability elements, the better of the market innovation value;

H2: the better of the resource elements, the better of the innovation behavior;

H2a: the better of the visible resource elements, the better of the technical innovation value;
H2b: the better of the visible resource elements, the better of the market innovation value;
H2c: the better of the visible resource elements, the better of the management innovation value;
H2d: the better of the invisible resource elements, the better of the technical innovation value;
H2e: the better of the invisible resource elements, the better of the market innovation value;
H2f: the better of the invisible resource elements, the better of the management innovation value;.

H3: the better of the behavior elements, the better of the innovation performance;

H3a: the better of the technical innovation elements, the better of the economic performance;
H3b: the better of the technical innovation elements, the better of the social performance;
H3c: the better of the technical innovation elements, the better of the potential performance;
H3d: the better of the management innovation elements, the better of the economic performance;
H3e: the better of the management innovation elements, the better of the social performance;
H3f: the better of the management innovation elements, the better of the potential performance;
H3g: the better of the market innovation elements, the better of the economic performance;
H3h: the better of the market innovation elements, the better of the social performance;
H3i: the better of the market innovation elements, the better of the potential performance;

H4: The environment support has the positive regulation for the other elements;

H4a: the better of the environment support, the resource elements have bigger impact on the technical innovation behavior;
H4b: the better of the environment support, the resource elements have bigger impact on the market innovation behavior;
H4c: the better of the environment support, the resource elements have bigger impact on the management innovation behavior;

H5: The environment uncertain has the regulation for the other elements;

H5a: the weaker of the environment uncertain, the technical innovation have bigger impact on the organizational innovation performance;
H5b: the weaker of the environment uncertain, the market innovation have bigger impact on the organizational innovation performance;
H5c: the weaker of the environment uncertain, the management innovation have bigger impact on the organizational innovation performance;

113.4 Research Methods

We construct the SEM path figure according to the theoretical model of the organizational innovation path forming principle model and relate hypothesis [8]. We need to take the potential variables and the measurements variables into consideration. The external variables include ability elements and resource elements. The internal variables include technical innovation, market innovation, management innovation and organizational innovation and economic performance and the social performance and the potential performance [9]. The Fig 113.2 is the initial PLS-SEM model.

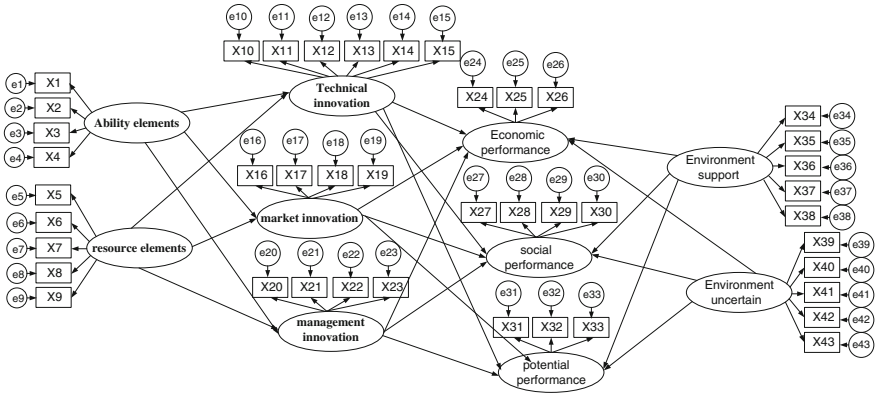


Fig. 113.2 The initial PLS-SEM model

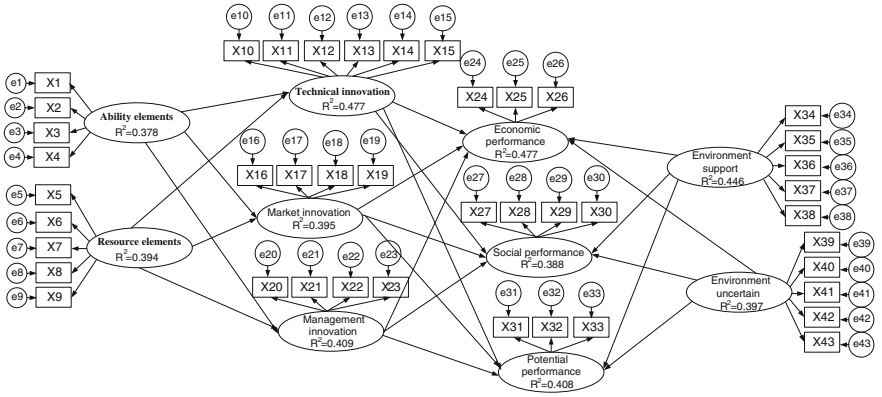


Fig. 113.3 The modification of the PLS-SEM model

113.5 Research Verification

In order to test and verify the correctness of theoretical model, this paper uses AMOS software to build model structure [10], then this paper test and Verify the match degree between the theoretical model and the realistic data. The following research answer means the match degree between the theoretical model and the realistic data can accept. The Fig 113.3 is the verification of the PLS-SEM model [11].

From Fig 113.3 we can see the model R^2 value is bigger than 0.649, that inflect the whole model illustrate the half of the innovation performance value, that means the organizational innovation theory has high illustrate level. The ability elements, resource elements impact on the innovation behavior R^2 value is 0.447, 0.395, and 0.709 that means the above elements have strong illustrate force. The ability elements impact on the technical innovation, market innovation, and management innovation path coefficient are 0.339, 0.492, 0.372, that means the impact effect is obvious, and the entire hypothesis pass the verification.

113.6 Conclusion

This chapter is a study on the different effect elements impact on the innovate-oriented enterprise organizational innovation path forming which based on the organizational innovation related theories. The forming of the organizational innovation is completed by the different elements influence and the environment elements are the regulation factor.

This research has the following theoretical contributions: firstly, this chapter enlarges the organizational innovation path forming factors scale, and takes the environment elements as the regulate factor. Secondly, this chapter analyzes the

ability elements and the resource elements impact on the organizational innovation path forming, and then uses the SEM method to verify the ability element, resource element, behavior element and the environment element impact on the performance function.

This research has the following realistic contribution for the enterprise management. The enterprise should take the environment uncertainty and the support into consideration when the enterprise makes decision. The enterprise should cultivate the ability and combine the resource of the enterprise to promote the enterprise value and performance.

This resource has some limitations. The important character of the organizational innovation is the dependence, and researchers haven't taken the dependence into consideration. The next work is combining the research on the organizational innovation path forming with research on the different elements impact on the dependence character. The effect elements from this research are limited since the data researchers collected may have little difference from the enterprise realistic situation. The next work is to insert the enterprise investigation part into the verify part, and make the asking paper match for the enterprise realistic situation. During the verification work, researchers take more innovation-oriented enterprise into consideration to enlarge the research universal character.

References

1. Chen, C.-M.: Study on the organizational innovation based on chaos theory. In: Proceedings of 2006 International Conference on Management Science and Engineering, pp. 102–106 (2007)
2. Yang, Y.: Research on organizational innovation mode in coal logistics enterprise. In: International Conference on Information Engineering and Applications, pp. 388–424 (2012)
3. Wang, Y.: Factors affecting enterprise boundaries an organizational innovation perspective. In: 2nd International Conference on E-Business and E-Government, pp. 519–534 (2011)
4. Wong, T.C.: A neural network-based approach of quantifying relative importance among various determinants toward organizational innovation. Expert Syst. Appl. **38**(10), 03065–13070 (2011)
5. Lin, M.: How does organizational commitment affect organizational innovation. In: 1st International Conference on E-Business and E-Government, pp. 14–20 (2010)
6. Shi, C.-S.: The impact of structure-oriented organizational innovation on technological innovation. In: 2006 International Conference on Management Science and Engineering, pp. 228–234 (2007)
7. Chen, C.-M.: Study on the function mechanism of enterprise culture in the enterprise organizational innovation. In: 2007 International Conference on Management Science and Engineering, pp. 148–152 (2008)
8. Wu, J.: LSP-client relationship a new angle of view on logistics firms' organizational innovation. In: 4th IEEE International Conference on Management of Innovation and Technology, pp. 229–241 (2008)
9. Guaglianone, M.T.: MNEMO (methodology for knowledge acquisition and modeling): definition of a global knowledge management approach combining knowledge modeling techniques. Adv. Inf. Sci. Serv. Sci. **7**(6):161–162 (2012)

10. Yiqun, Li.: Information services platform of international trade based on E-commerce. Adv. Inf. Sci. Serv. Sci. **9**(2), 79–85 (2011)
11. Lo, L.Y.-S.: The effect of price presentation, sales restrictions, and social networks on consumer EWOM intention. Adv. Inf. Sci. Serv. Sci. **3**(2):2–5 (2011)

Chapter 114
Assessment of S&T Progress' Share in China Provincial Economy Growth

Qiang Li

Abstract The analysis based on neoclassical growth theories shows that calculations of Contribution Rate of S&T Progress on China provincial economy growth could not provide exact explanation on economic effect of S&T progress. This study established a new growth accounting model based on endogenous growth theory with output of GDP and inputs of capital, labor, human capital and S&T progress. Empirical study based on panel data shows that this method can get better understanding on practical contribution of S&T progress on China's provincial economy growth.

Keywords Neoclassical growth theory · Economy growth · Growth accounting model growth accounting model

114.1 Introduction

Calculation of Contribution Rate of S&T Progress on Chinese Economy is an active area in economic research. Experts estimate S&T progress' contribution in Chinese economic growth science 1980s. However, there is no unified standard and theory in existing literature, and there are big differences on researches among experts and institutes. Methods based on Solow growth model still need to be discussed further.

More and more experts started to take empirical methods to find evidences of endogenous growth with the development of the endogenous growth theory. For example, Jones [1] analyzed economy growth in the U.S. from 1950 to 1993. Cécile Denis [2] reviewed productivity change in USA and EURO as well as some studies in China. However, there are few studies in growth accounting based on

Q. Li (✉)
Institute of Policy and Management, Chinese Academy of Sciences, Beijing, China
e-mail: lq@casipm.ac.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_114,

endogenous growth theory, especially studies taking quantitative technical progress and human capital measurements into consideration.

Theoretic analysis and empirical study based on Solow growth model in this paper demonstrated that S&T Progress' share in Chinese economy growth is just a kind of residual between observational output increasing rate and input increasing rates, which could not provide explanation to the exact effect of S&T progress on economy growth.

This study established a new growth accounting model based on endogenous growth model with output of GDP and inputs of capital, labor, human capital and technology-value-equivalent. Concept of technology-value-equivalent and its estimating method are given to provide observational S&T progress in the economic body itself. Empirical study on provincial panel data shows that China is experiencing endogenous growth from 2000 to 2006 than that in year 1994–1999.

114.2 Defects of Solow Residual Framework

According to Solow growth model [3], the product function of constant returns-to-scale can be expressed as $f(\lambda K, \lambda L, t) = \lambda f(K, L, t)$, in which Y for outputs, K and L for capital and labor separately. Under conditions of perfect competition of products and input factors, the growth rates can be expressed as

$$\frac{\dot{Y}}{Y} = \alpha \frac{\dot{K}}{K} + (1-\alpha)\frac{\dot{L}}{L} + \frac{\partial Y}{\partial t}\frac{1}{Y} \tag{114.1}$$

where α stands for output elasticity of capital.

In 1992, previous State Planning Commission of China, National Bureau of Statistics of China jointly delivered *Notice on Calculation of S&T Progress' Contribution to Economy Growth*, giving computation formula as

$$E_a = \frac{a}{y} = 1 - \alpha\frac{k}{y} - (1-\alpha)\frac{l}{y} \tag{114.2}$$

where

$$a = \tfrac{\dot{A}}{A} \quad y = \tfrac{\dot{Y}}{Y}; \quad k = \tfrac{\dot{K}}{K}; \quad l = \tfrac{\dot{L}}{L}$$

The file also gives instructions on output elasticity of capital (0.35) and labor (0.65) separately. Data can get from China Statistical Yearbook except for capital stock, which can be calculated as

$$K_t = K_{t-1} + \frac{(CF_t - D_t)}{P_t} \tag{114.3}$$

where CF_t is gross fixed capital formation in year t, D_t is depreciation of fixed capital in year t, P_t is price indices of investment in fixed assets in year t, Other variables are presented in 1990 fixed price. Current capital stock can be estimated as

$$K_{jT_0} = \frac{I_{jT_0}}{0.036 + g_{jT_0-T}} \tag{114.4}$$

where K_{jT_0} stands for current capital stock in province j at base year, I_{jT_0} stands for total amount of current capital formation in province j at base year, and g_j stands for total capital formation growth speed. Thus we can have Table 114.1 based on Eq. (114.2).

Results in Table 114.1 manifest that there are significant positive correlation between provincial Solow residuals and GDP increase, and large Solow residuals exist in developed provinces such as Beijing, Tianjin, Shanghai, Guangdong, Jiangsu, Zhejiang, Shandong and so on. However, the result indicates that

Table 114.1 Contribution rates of S&T progress

Provinces	Growth rates			Solow residual	Contribution rates		
	GDP	*K*	*L*		*K*	*L*	*TFP*
Beijing	0.15	0.14	0.02	0.09	31.9	10.0	58.1
Tianjin	0.13	0.12	–0.02	0.10	31.1	–7.5	76.3
HeBei	0.12	0.14	0.00	0.06	43.4	2.4	54.1
Shanxi	0.12	0.13	0.00	0.07	38.2	0.8	61.0
Neimenggu	0.13	0.14	0.00	0.08	36.6	0.1	63.4
Liaoning	0.09	0.08	–0.01	0.06	31.7	–3.7	71.9
Jilin	0.11	0.13	–0.01	0.07	43.2	–8.6	65.4
Heilongjiang	0.09	0.09	0.01	0.05	35.3	4.8	59.9
Shanghai	0.12	0.11	0.00	0.08	32.1	1.6	66.4
Jiangsu	0.12	0.12	–0.01	0.08	34.3	–1.1	66.7
Zhejiang	0.13	0.14	0.01	0.07	37.1	6.1	56.8
Anhui	0.10	0.12	0.01	0.05	41.1	6.8	52.0
Fujian	0.11	0.15	0.02	0.05	47.8	9.0	43.2
Jiangxi	0.11	0.13	0.00	0.06	40.5	1.7	57.9
Shandong	0.12	0.13	0.01	0.07	37.2	4.4	58.5
Henan	0.12	0.14	0.02	0.05	42.8	11.9	45.3
Hubei	0.09	0.14	–0.01	0.04	53.5	–2.3	48.9
Hunan	0.10	0.13	0.00	0.05	43.9	2.7	53.5
Guangdong	0.13	0.10	0.02	0.08	27.8	10.0	62.2
Guangxi	0.09	0.10	0.01	0.04	39.8	9.9	50.3
Hainan	0.08	0.06	0.01	0.05	25.9	7.8	66.4
Chongqing	0.10	0.16	0.00	0.04	56.5	1.9	41.6
Sichuan	0.10	0.13	–0.01	0.05	46.0	–0.6	54.6
Guizhou	0.09	0.14	0.02	0.03	54.2	13.1	32.6
Yunnan	0.08	0.08	0.01	0.05	34.1	9.0	56.8
Tibet	0.13	0.18	0.02	0.06	47.0	8.4	44.6
Shannxi	0.12	0.12	0.01	0.07	37.8	4.3	57.9
Gansu	0.11	0.12	0.01	0.06	40.4	7.9	51.7
Qinghai	0.10	0.15	0.02	0.03	55.3	10.1	34.6
Ningxia	0.11	0.14	0.02	0.04	46.6	13.6	39.9
Xinjiang	0.11	0.10	0.01	0.06	32.2	7.0	60.7

underdeveloped provinces such as Guizhou, Qinghai, Guangxi, Yunnan, Hainan, Ningxia, and Anhui are experiencing higher S&T progress' share in economy development which means the faster GDP grows the fewer S&T progress shares in provincial economic growth.

For example, GDP growth rate in Beijing is the highest in China. Solow Residual is higher than Shanghai, Guangdong, Jiangsu and other inshore provinces, also higher than Xinjiang, Heilongjiang, Jilin, and Liaoning, but S&T progress' contribution rate in GDP growth is less than above mentioned province and the average level of China as well.

114.3 A New Growth Accounting Model

Take the market economy version of the Romer Model into consideration [4]. This economy is composed of three sectors: The research sector takes human capital and a given stock of knowledge (A) as input factors produces new knowledge. This sector behaves competitively.

The intermediate sector uses the knowledge produced by the research sector as an input factor and produces intermediate goods (x_i), which are used as input factors by the final good sector. The firms of the intermediate sector cannot be perfect competitors but must have some market power since they use the non-rival factor knowledge as input.

The final good sector employs intermediate goods together with labor (L) and human capital (H) as input factors and produces the final good.

Production function for final output can be given as

$$Y = (H - H_A)^{\alpha} L^{\beta} \int_0^A x(i)^{1-\alpha-\beta} di \tag{114.5}$$

where $(H - H_A)$ is human capital employed in the production of the final good and L is unqualified labor. H is total human capital in the economy and H_A denotes human capital employed in the research sector. $(1 - \alpha) \in (0, 1)$ denotes the capital share, and β stands for output elasticity of labor.

Assuming that total stock of physical capital can be written as $K = \eta A x$ with η the amount of foregone consumption necessary to produce one unit of the intermediate good. The aggregate production function follows as:

$$Y = \eta^{\alpha+\beta-1} A^{\alpha+\beta} (H - H_A)^{\alpha} L^{\beta} K^{1-\alpha-\beta} \tag{114.6}$$

Logarithmic form can also be given as:

$$\begin{aligned} \ln Y - \ln K =& \alpha(\ln A + \ln(H - H_A) - \ln K) \\ &+ \beta(\ln A + \ln L - \ln K) + \ln \eta \end{aligned} \tag{114.7}$$

The differential coefficient form can be given as:

$$\frac{\dot{Y}}{Y} = (\alpha + \beta)\frac{\dot{A}}{A} + \alpha(\frac{\dot{H} - \dot{H}_A}{H - H_A}) + \beta\frac{\dot{L}}{L} + (1 - \alpha - \beta)\frac{\dot{K}}{K} \tag{114.8}$$

Similarly as Eq. (114.2), the S&T progress' contribution to economy growth can be given as

$$E_a = \frac{a}{y} = 1 - \alpha\frac{h}{y} - (\beta)\frac{l}{y} - (1 - \alpha - \beta)\frac{k}{y} \tag{114.9}$$

It is clear that in Eq. (114.8) we can get aggregate output elasticity $1 + \alpha + \beta > 1$, denotes the Increasing returns to scale in endogenous growth economy.

To carry out regression analysis according to Eqs. (114.7) and (114.8), we will have to establish measurements of knowledge. Human capital employed in the production of the final good sector can get from China statistical yearbook on science and technology, defined as the difference value of "Provincial Population by Educational Level of College & Higher Level" and "Provincial S&T Personnel".

114.4 Measure of Technology-Value-Equivalent

Reconsidering the market economy version of the Romer model and giving the concept of technology-value-equivalent as: Technology-value-equivalent is the fixed value of knowledge input in the research sector. It is an index relative to base year.

Thus the R&D activity can be described as a production process: In the research sector, technology-value-equivalent gains knowledge spillover with human capital inputs. The new knowledge produced in the research sector is the same of technology progress in practical economy running.

And the new knowledge, as well as technology progress is presented as three kinds of overlapped and interacted elements: the first element is scientific discovery, which means to find scientific regularity in experimental observations. Scientific discovery is often presented in forms of papers, books and so on; The second element is technological innovation, it's the first commercial use of invention and is often presented in forms of patents and know-how; The third element is technology diffusion, a process of adopting "new-to-the-firm" processes and products [5].

So we can establish the index of technology-value-equivalent increment with primary indexes of scientific discovery, technological innovation and technology diffusion. Secondary indexes are given as papers published in SCI journals (I_1); patent under the trilateral of USA, JP and EU (I_2); and technology expenditures with equal weight (I_3). Providing that the overall increment of technology-value-equivalent in China at base year is 100, the share of province j is:

$$\dot{P}_j^{T_0} = \frac{100}{3}(\frac{I_{1j}}{\sum_{j=1} I_{1j}} + \frac{I_{2j}}{\sum_{j=1} I_{2j}} + \frac{I_{3j}}{\sum_{j=1} I_{3j}}) \tag{114.10}$$

Table 114.2 Provincial technology-value-equivalent

Years	1993	1995	1997	1999	2001	2003	2004	2005	2006
Beijing	110	118	123	130	161	241	282	366	467
Tianjin	17.1	19.0	20.3	22.5	26.6	35.9	40.3	49.6	63.1
HeBei	12.7	14.9	15.0	15.5	18.2	26.4	29.8	35.4	44.7
Shanxi	7.9	8.7	9.0	9.4	11.6	18.2	21.6	28.0	35.5
Neimenggu	2.7	3.2	3.2	3.5	4.6	8.5	10.3	12.5	16.6
Liaoning	36.6	37.7	40.4	42.3	47.9	67.0	75.1	89.4	115
Jilin	15.1	17.1	17.4	18.5	22.0	29.9	33.8	39.5	51.7
Heilongjiang	15.0	16.3	17.3	18.2	21.1	29.2	32.1	38.3	47.7
Shanghai	48.1	51.1	52.3	55.9	68.9	105	127	166	225
Jiangsu	31.8	35.8	38.4	45.7	55.7	75.5	89.0	110	14
Zhejiang	17.6	21.3	22.2	24.5	30.4	45.3	54.7	70.9	95.6
Anhui	11.3	11.7	12.1	13.0	16.4	23.4	28.1	33.7	39.2
Fujian	7.5	8.3	8.8	9.7	12.9	19.4	21.8	26.1	33.4
Jiangxi	5.1	5.4	5.4	5.8	6.9	10.0	12.0	14.6	16.5
Shandong	17.9	24.0	27.4	29.7	37.4	57.0	68.1	87.2	111
Henan	14.1	14.9	14.9	16.0	18.5	27.7	30.2	35.0	41.8
Hubei	21.3	23.4	24.3	25.5	30.5	44.8	51.5	65.1	88.2
Hunan	14.0	16.1	17.7	20.4	25.4	35.9	40.1	51.7	63.2
Guangdong	20.2	23.7	25.4	28.8	40.0	62.1	74.0	98.4	151
Guangxi	4.1	4.8	4.9	5.3	6.0	9.2	9.8	12.0	14.7
Hainan	1.0	1.6	1.5	2.1	2.8	3.9	3.7	4.0	5.3
Chongqing	1.5	1.7	1.5	2.5	6.7	14.5	18.1	24.5	29.6
Sichuan	30.9	32.5	33.1	33.9	35.4	44.3	48.3	56.0	69.7
Guizhou	3.2	3.4	3.2	3.1	3.3	5.6	7.2	8.9	12.8
Yunnan	5.9	6.8	7.9	9.7	14.8	22.6	24.8	29.1	34.4
Tibet	0.2	0.2	0.2	0.2	0.2	0.5	0.6	0.8	0.9
Shannxi	14.5	17.4	20.0	21.5	25.7	36.5	42.1	49.0	64.2
Gansu	9.4	9.9	9.9	10.8	12.5	17.2	20.3	23.5	27.9
Qinghai	1.0	1.1	1.0	1.1	1.2	2.0	2.4	2.8	3.2
Ningxia	0.6	0.8	0.9	0.8	1.2	2.2	2.7	3.9	4.8
Xinjiang	2.1	2.7	3.3	4.0	5.5	10.6	12.4	14.7	16.0

Providing that knowledge depreciation rate is 20 %, we can have technology-value-equivalent (as well as the value of knowledge stock, standing for technology level) of province j at base year (T_0) like formula (114.4):

$$P_j^{T_0} = \frac{\dot{P}_j^{T_0}}{20\,\%} = 5\dot{P}_j^{T_0} \tag{114.11}$$

And the technology-value-equivalent of province j in year (T_i) can be computed as follows:

$$P_j^{T_i} = 0.8P_j^{T_{i-1}} + \dot{P}_j^{T_i} \tag{114.12}$$

Provincial time serial data of (I_1), (I_2), and (I_3) can be obtained from China statistical yearbook on science and technology. Put the data into formula (114.10)–(114.12), we can get provincial time serial data of technology-value-equivalent in Table 114.2.

114.5 S&T Progress' Share in Provincial Economy Growth

Provincial time serial data of (L) can be obtained from "Number of Employed Persons" in China Statistical Yearbook 1994–2007. Similarly, output (Y) in 1990 fixed price can be obtained from "Gross Domestic Product" at current prices and "Consumer Price Indices". Human capital ($H - H_A$) can be derived from China statistical yearbook on science and technology 1994–2007. Capital stock (K) can also be estimated with formula (114.3) and (114.4). And technology level (A) is given in Table 114.2. Put provincial time serial data of (Y), (K), (A), ($H - H_A$) and (L) into formula (114.7), we can get α and β. And also y, a, h, l and k. Growth rate of output and inputs are given in Table 114.3.

Table 114.3 Output and inputs growth rate

Years	1994–1999					2000–2006				
	y	a	h	L	k	y	a	h	l	k
Beijing	0.14	0.03	0.08	–0.01	0.15	0.15	0.22	0.10	0.08	0.13
Tianjin	0.12	0.05	0.09	–0.01	0.13	0.16	0.17	0.11	0.01	0.11
HeBei	0.12	0.03	0.20	0.01	0.16	0.14	0.18	0.10	0.00	0.12
Shanxi	0.11	0.03	0.07	0.01	0.12	0.16	0.23	0.10	0.01	0.14
Neimenggu	0.11	0.04	0.12	0.01	0.14	0.18	0.28	0.15	0.00	0.17
Liaoning	0.07	0.03	0.13	0.00	0.08	0.10	0.17	0.09	0.02	0.09
Jilin	0.08	0.04	0.14	–0.01	0.15	0.15	0.16	0.06	0.00	0.13
Heilongjiang	0.09	0.03	0.07	0.03	0.09	0.11	0.16	0.09	0.00	0.10
Shanghai	0.12	0.02	0.05	–0.01	0.14	0.13	0.24	0.14	0.04	0.10
Jiangsu	0.11	0.06	0.07	0.00	0.12	0.15	0.18	0.11	0.01	0.12
Zhejiang	0.12	0.05	0.16	0.00	0.16	0.16	0.24	0.16	0.03	0.13
Anhui	0.11	0.03	0.12	0.02	0.13	0.12	0.19	0.16	0.01	0.10
Fujian	0.13	0.04	0.12	0.01	0.18	0.11	0.20	0.12	0.02	0.12
Jiangxi	0.10	0.02	0.17	0.01	0.13	0.14	0.18	0.13	0.01	0.13
Shandong	0.11	0.09	0.06	0.01	0.14	0.17	0.22	0.10	0.02	0.12
Henan	0.12	0.02	0.09	0.02	0.16	0.15	0.17	0.09	0.01	0.13
Hubei	0.09	0.03	0.11	0.01	0.16	0.10	0.21	0.10	0.01	0.11
Hunan	0.10	0.06	0.09	0.01	0.13	0.12	0.18	0.09	0.01	0.13
Guangdong	0.11	0.06	0.26	0.02	0.11	0.15	0.27	0.10	0.03	0.10
Guangxi	0.07	0.04	0.05	0.02	0.11	0.13	0.18	0.18	0.01	0.10
Hainan	0.04	0.09	0.15	0.00	0.07	0.11	0.09	0.12	0.02	0.06
Chongqing	NA	NA	NA	NA	NA	0.10	0.30	0.17	0.00	0.16
Sichuan	NA	NA	NA	NA	NA	0.10	0.05	0.08	–0.03	0.13
Guizhou	0.07	–0.01	0.09	0.02	0.09	0.12	0.27	0.10	0.02	0.18
Yunnan	0.10	0.08	0.00	0.02	0.08	0.10	0.19	0.15	0.01	0.10
Tibet	0.13	–0.07	–0.22	0.01	0.18	0.14	0.31	0.01	0.02	0.18
Shannxi	0.09	0.08	0.02	0.01	0.13	0.15	0.18	0.14	0.01	0.14
Gansu	0.11	0.02	0.07	0.01	0.11	0.12	0.17	0.12	0.03	0.13
Qinghai	0.07	0.01	0.11	0.01	0.15	0.13	0.21	0.15	0.02	0.16
Ningxia	0.10	0.06	0.07	0.03	0.13	0.13	0.31	0.14	0.02	0.16
Xinjiang	0.10	0.11	0.14	0.01	0.11	0.12	0.25	0.08	0.02	0.10

Table 114.4 Provincial growth accounting results

Provinces	Parameters		Output elasticity			
	α	β	a	h	l	k
Beijing	0.058	0.295	0.35	0.06	0.59	0.50
Tianjin	0.161	0.139	0.30	0.16	0.14	0.70
HeBei	–0.077	0.390	0.31	–0.08	0.39	0.69
Shanxi	–0.082	0.351	0.27	–0.08	0.35	0.73
Neimenggu	–0.049	0.275	0.23	–0.05	0.28	0.77
Liaoning	0.075	0.010	0.09	0.08	0.01	0.92
Jilin	0.094	0.042	0.14	0.09	0.04	0.86
Heilongjiang	0.052	0.035	0.09	0.05	0.04	0.91
Shanghai	0.082	0.536	0.62	0.08	0.54	0.38
Jiangsu	0.023	0.307	0.33	0.02	0.31	0.67
Zhejiang	–0.047	0.310	0.26	–0.05	0.31	0.74
Anhui	–0.117	0.374	0.26	–0.12	0.37	0.74
Fujian	–0.284	0.450	0.17	–0.28	0.45	0.83
Jiangxi	–0.050	0.313	0.26	–0.05	0.31	0.74
Shandong	–0.081	0.536	0.46	–0.08	0.54	0.55
Henan	–0.161	0.473	0.31	–0.16	0.47	0.69
Hubei	–0.307	0.570	0.26	–0.31	0.57	0.74
Hunan	–0.193	0.361	0.17	–0.19	0.36	0.83
Guangdong	0.135	0.349	0.48	0.07	0.41	0.52
Guangxi	0.172	–0.069	0.10	0.17	–0.07	0.90
Hainan	0.015	0.234	0.25	0.02	0.23	0.75
Chongqing	0.351	–0.770	–0.42	0.35	–0.77	1.42
Sichuan	–0.049	0.419	0.37	–0.05	0.42	0.63
Guizhou	–0.429	0.373	–0.06	–0.43	0.37	1.06
Yunnan	–0.026	0.056	0.03	–0.03	0.06	0.97
Tibet	–0.152	0.378	0.23	–0.15	0.38	0.77
Shannxi	0.059	0.162	0.22	0.06	0.16	0.78
Gansu	–0.146	0.173	0.03	–0.15	0.17	0.97
Qinghai	–0.312	0.670	0.36	–0.31	0.67	0.64
Ningxia	–0.445	0.582	0.14	–0.45	0.58	0.86
Xinjiang	–0.073	0.248	0.18	–0.07	0.25	0.83

Parameters of α, β and output elasticity of inputs increments are given in Table 114.4.

Put data of output elasticity in Table 114.4 into formula (114.9) and we can get S&T Progress' and other inputs' increment's share in provincial economy growth. The results are given in Table 114.5.

From Table 114.5 we can see that China is still an investment oriented economy. From 2000 to 2006, investment growth shares 69.5 % of economy growth, and it could be much higher in period of 1994–2000, as high as 93.2 %. In the meantime, we can see the emerging of knowledge-based economy in China.

Table 114.5 Provincial growth accounting results

Provinces	Share of input increment in economy growth							
	1994–1999				2000–2006			
	a	*h*	*L*	*k*	*a*	*h*	*l*	*k*
Beijing	25.3	14.5	3.2	57.0	49.2	8.1	6.0	36.6
Tianjin	12.2	12.5	–1.1	76.4	35.0	12.5	0.5	52.0
HeBei	9.6	–15.1	4.4	101	43.8	–6.3	0.5	62.0
Shanxi	9.9	–6.5	2.4	94.1	41.2	–5.2	1.7	62.3
Neimenggu	7.8	–5.8	1.9	96.1	36.0	–4.3	0.4	67.9
Liaoning	2.8	11.7	0.0	85.6	14.4	6.7	0.2	78.8
Jilin	2.7	4.0	–0.4	93.7	26.0	10.0	1.4	62.6
Heilongjiang	3.4	4.5	1.0	91.1	13.6	4.4	–0.1	82.1
Shanghai	26.5	14.7	–5.5	64.2	66.1	3.1	0.6	30.2
Jiangsu	20.9	1.7	–0.9	78.3	42.9	1.8	2.6	52.7
Zhejiang	12.4	–6.8	0.6	93.8	40.7	–4.9	6.6	57.6
Anhui	6.9	–14.5	7.5	100	45.3	–18.1	2.6	70.2
Fujian	6.2	–28.5	5.5	117	32.1	–33.3	10.0	91.2
Jiangxi	5.8	–9.1	4.5	98.7	35.9	–4.7	2.9	65.9
Shandong	35.8	–4.5	5.0	63.6	62.0	–5.0	4.8	38.2
Henan	6.3	–14.6	11.2	97.1	41.8	–11.7	3.6	66.3
Hubei	9.8	–41.2	3.4	128	54.5	–31.5	4.4	72.6
Hunan	10.6	–17.1	4.3	102	26.5	–14.7	2.6	85.6
Guangdong	35.8	3.2	2.9	58.4	48.0	3.8	4.2	44.0
Guangxi	2.2	10.1	0.4	87.3	15.7	11.8	–3.9	76.4
Hainan	31.7	3.1	0.7	64.5	32.1	2.5	8.0	57.4
Chongqing	NA	NA	NA	NA	–85.7	41.6	–1.6	146
Sichuan	NA	NA	NA	NA	26.0	–5.1	–17.5	96.6
Guizhou	1.2	–64.2	11.7	151	–12.5	–35.6	5.4	143
Yunnan	3.3	0.0	1.2	95.6	5.9	–4.1	0.7	97.5
Tibet	–11.2	22.3	3.3	85.6	33.8	–0.9	4.3	62.8
Shannxi	11.7	13.8	2.6	71.9	41.1	3.8	3.7	51.3
Gansu	0.6	–10.1	1.8	108	4.1	–15.8	4.1	108
Qinghai	3.8	–52.1	14.6	134	54.8	–34.1	10.0	69.3
Ningxia	8.5	–34.2	16.3	110	35.6	–51.5	8.9	107
Xinjiang	19.1	–10.3	3.3	87.9	36.0	–4.6	4.8	63.8
Average	11.9	–8.3	4.1	93.2	35.6	–8.5	3.5	69.5

In years from 2000 to 2006, the S&T progress' share in china economy growth is 35.6 %, as much as 3 times of years from 1994 to 1999.

Results of this assessment have proved that China is still running an investment oriented economy [6]. However, with the execution of "Rejuvenate Our Country through Science and Education", knowledge-based economy is emerging in China and the developed provinces have been experiencing higher share of S&T progress in Economic growth (See Fig. 114.1).

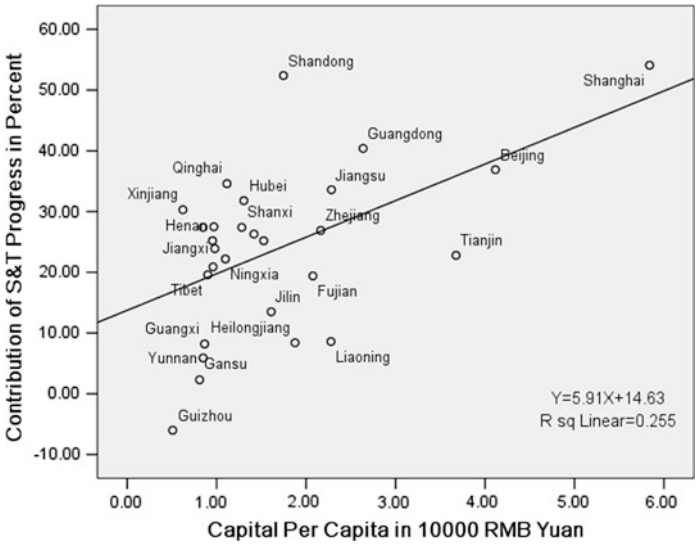


Fig. 114.1 Coefficient of capital per capita and S&T progress' contribution on economy growth

114.6 Conclusion

Science and technology made significant contributions to the economic progress and society development, just as Technology in the National Interest says that about 50 percent of the long-run economic growth in the U.S. can be attributed to S&T Progress. Focusing on the disputes in growth accounting on Chinese economy growth, this paper established a framework based on endogenous growth and assessed S&T progress' share in China provincial economy growth.

This assessment have proved that although knowledge-based economy is emerging in China, and the developed provinces have been experiencing higher share of S&T progress in Economic growth, China is still running an investment oriented economy.

References

1. Jones, C.I.: Introduction to Economic Growth, 2nd edn. W.W. Norton and Company, New York (2002)
2. Denis, C., McMorrow, K., Röger, W.: An analysis of EU and US productivity developments. European Communities (2004)
3. Solow, R.M.: Technical change and the aggregate production function. Rev. Econ. Stat. **39**(3), 312–320 (1957)

4. Greiner, A., Semmler, W., Gong, G.: Endogenous Growth: Theory and Time Series Evidence. Princeton Press, New York (2003)
5. Krugman, P.R., Obstfeld, M.: International Economics: Theory and Policy, 6th edn. Addison Wesley Publishing Company, Boston (2002)
6. Martin, P., Ottaviano, G.I.P.: Growth and agglomeration. Int. Econ. Rev. **42**(4), 947–968 (2001)

Chapter 115
Inter-Firm Innovation Networks: The Impact of Scale-Free Property on Firm Innovation

Xiaolong Lu, Wen Zhou, Yan Zhao, Ying Zhu and Shengnan Fei

Abstract Enterprise alliance network is considered to have scale-free property. This paper aims to study the correlation between scale-free property and enterprise innovation performance. Networks of alliances in automobile industry are constructed, and then researchers utilize Maximum Likelihood Estimation and KS-test to calculate the p value of the network's scale-free property. Negative binomial regression model is used to analyze the relevance between enterprise innovative performance and the scale-free property of alliance network, with patents number that firms have applied as the indicator of firms' innovative performance. Experiment result shows that scale-free attribute of enterprise alliance network has quite a significant positive effect upon enterprise innovative performance.

Keywords: Innovative network · Scale-free network · Innovation performance

115.1 Introduction

Knowledge and technology innovation is the core competitive force of the enterprise, while most of the enterprises, especially for enterprises from technology-intensive industries, are unable to satisfy the competition requirements for technology and knowledge. As a result, they seek technology and knowledge aids from the outside, which leads to the emergence of enterprise alliances. Based on this demand, when an enterprise prepare to ally, it will consider external enterprises which control the leading technology or marketing resources as priorities,

X. Lu · W. Zhou (✉) · Y. Zhu · S. Fei
School of Computer Engineering and Science, Shanghai University, Shanghai, China
e-mail: zhouwen@shu.edu.cn

Y. Zhao
School of Management, Shanghai University, Shanghai, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_115,

and such a preferential mechanism is the dynamic force which leads to the formation of scale-free property.

Most of the previous papers focus on the existence of scale-free attribute and what it means in some specific field, while in this paper, we calculate the characteristic indicator of scale-free attribute of the network, and analyze the relation between this structural characteristic of the alliance network and the innovation performance of the enterprises quantitatively. We collect information about domestic enterprise alliances between 2000 and 2008, set 3 years as a time period, and then for enterprise alliances formed within each time period we build a network graph, upon which we calculate the indicative value of scale-free property. We represent the innovation ability of an enterprise by the number of patents that this enterprise has applied during the time period. At last, we use negative binomial regression model to analyze the relationship between the two indicators quantitatively. The experiment demonstrates the correlation between the structural characteristic of scale-free property and innovation capability of the enterprises within the alliance.

115.2 Related Work

Complex network is a new-rising research direction originated in the 90 s, which is based on graph theory. Relevant applications can be found in many areas [1, 2]. Scale-free property is one of the network characteristics which focus on the degree distribution of the nodes in the network. Barabási and Albert [3] are the first to present this characteristic, which are followed by many researchers in several areas. Researchers find that quite a part of the networks in these areas have scale-free property, such as the Internet [4] and the spreading computer virus [5], etc. Based on the model proposed by Barabási and Albert (the BA model), new models and methods are presented. Klemm and Eguiluz [6] put forward a model which renders a restriction to the age of nodes. Clauset and Newman [7] provide a quantitative research of scale-free property, which can be used to calculate the extent of scale-free attribute. In this paper, we will use this method to generate a quantitative indicator value which can describe to what extent the network is scale-free.

Researches of the enterprise innovation alliance mainly focus on regarding the enterprise innovation alliance as a network and analyzing the structural characteristics of the network. Furthermore, we can analyze the relationship between these structural characteristics and the real world consequences [8]. It is revealed that the structural characteristics of the network could affect information flow and knowledge sharing of the enterprise alliance [9, 10], so there exists a direct correlation between the network structure and the goal of enterprise alliances, which is to share knowledge and technologies. It has been confirmed that networks of enterprise alliances in some specific industries have scale-free property [11, 12]. This can be considered to be the result of the preferential attachment (i.e., researchers always choose to cite the best papers in their research fields), as is demonstrated in some researches [3].

115.3 Method

Visually, we can plot the frequency distribution on doubly logarithmic axes. And if we find that the spots are 'nearly' on a straight line by our eyes, we can boldly assert that the network is scale-free. However, this approach could generate serious errors, since it requires the researchers to check whether the regression is good or not, and there isn't a benchmark which we can use to judge to what extent the data obey the power law. Furthermore, deviation of the data near the start point could be very big, which is normal. So there should be a lower bound from which we judge the data, which can make the estimate more accurate. In this paper, we generate a p value to indicate the scale-free property of a network through Maximum Likelihood Estimation (MLE) and Kolmogorov–Smirnov test (KS-test) [7].

Then negative binomial regression model is utilized to do the analysis work. The dependent variable is the count of total number of patents that a particular firm has applied for in a given year, which varies from zero to several or even many. Empirical model has to accommodate the nature of these counts as positive, integer values. We consider count models like Poisson regression model, OLS regression model and negative binomial regression model as the possible applicant models. Poisson regression is often used as a starting point for modeling count data. However, Poisson regression has a very strong assumption, which is that the conditional variance equals to conditional mean. OLS regression is usually utilized to deal with log-transformed count outcome variables. But an issue arises when undefined values generated by taking the log of zero (which is frequently occurred in our data). Most importantly, since our patents data are over dispersed, and Poisson regression or OLS regression lacks the ability to model over dispersed data, so we employ negative binomial regression model to do the analysis, which can deal with the over dispersion issue successfully. Negative binomial distribution is a discrete probability distribution of the number of successes in a sequence of Bernoulli trials before a specified number of failures occur. And it perfectly fit into our patent counts model [13].

115.4 Experiments and Result Analysis

We excerpt all the alliance data in China's automobile industry from 2000 to 2008, and a database is built using these data. Then networks are constructed of all the alliances for every 3 years during the time period. We select five control variables for the analysis and calculate the p value to indicate the scale-free property of the networks. Through the analysis of the correlation between scale-free p value and the number of patents enterprise has applied, we can reach our conclusions. Here, almost all the control variables are calculated by UCINET, and the scale-free p value is generated by MATLAB.

115.4.1 Preparing for the Experiment

115.4.1.1 The Network

We collect the alliance data of automobile industry from SDC Platinum database from 2000 to 2008. Any enterprises that have allied during the time period are sent to our alliance database. Before 2000 the domestic automobile industry had just started to rise and few companies allied with others; Furthermore, what we want to analyze is the change of the number of patents with the network formed a few years before, so the information that we required about the enterprise innovation ability indicator—the number of patents, should be excerpted a few years after the forming of the alliance.

We set the enterprise alliance network as an undirected graph with all the weight on the edge to be 1. We assume that the influence of the alliance should be significant during 3 years after it formed. So we divide the time period (2000–2008) as seven separate sub time periods: from G_{2002}(2000–2002) to G_{2008}(2006–2008).

Here, we just present a network of G_{2002}(2000–2002). As is shown in Fig. 115.1, there exist a few nodes (with different colors), which indicates the existence of the attachment mechanism to some extent, although evidence is not enough just by our eyes.

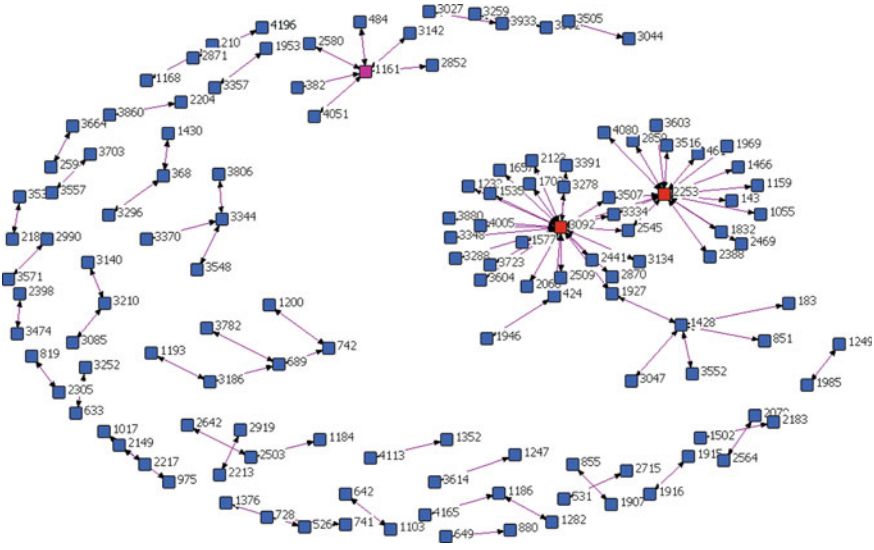


Fig. 115.1 Network of G_{2002}(2000–2002)

115.4.1.2 The Variables

Controlling variables. Some of the control variables, such as Network Density and Betweenness Centrality, can be calculated immediately through UCINET, while other variables can be obtained intermediately with UCINET, during which some other operations are needed. We use network density, betweenness, structural holes and centralization as the controlling parameters, which are the main factors that have impact on innovative performance of firms. Furthermore, pre_patent count (firm has applied in years before the alliance is formed) are used to control the heterogeneity of the application of patents.

Dependent variables. Since auto companies are mostly used to apply for patents to keep their rights, we use patents data to indicate firm's innovation ability [14]. Information about patents is collected from domestic Patent Office. For each network, the patents applied in 3 years after the formation of the alliance are gathered. We deal with the situation when a patent is applied by a group of enterprises by counting the application for each of the group member. Because most of the alliances do not announce their termination, it is important for us to determine when an alliance fades. In literature, the choice ranges from 3 to 5 years [14, 15]. Based on the situation and data in China automobile industry, we choose 3 years, and assume after this time period the effect fades.

115.4.2 Result and Analysis

To calculate the scale-free p value, we need to count the degree of every node in the network, and figure out the probability distribution of the degrees. The probability distribution and doubly logarithmic plots of G_{2002} network is depicted in Fig. 115.2.

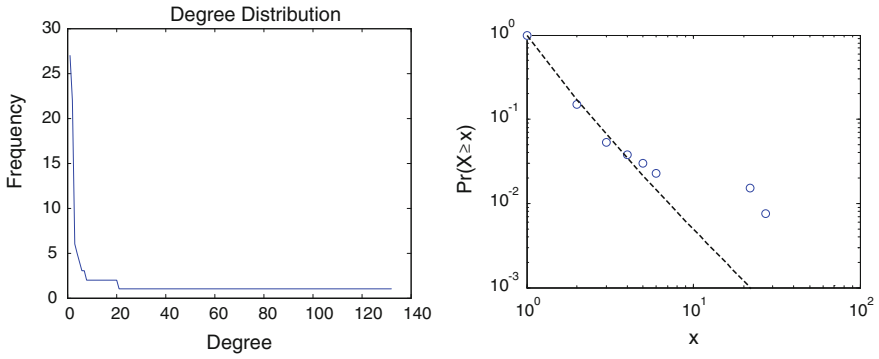


Fig. 115.2 Degree distribution and doubly logarithmic plots of G_{2002}(2000–2002)

Table 115.1 Scale-free *p* value during 2002–2008

K	2002	2003	2004	2005	2006	2007	2008
400	0.0075	0	0	0.0	0.0	0.0075	0.2925
600	0.0050	0	0	0.0017	0.0	0.0083	0.2817
800	0.0075	0	0	0.0013	0.0013	0.0088	0.2850
1000	0.0060	0	0	0.0010	0.0010	0.0090	0.2710

We can find that the large degrees on the right side have very small frequency of happening, while the frequency of the small degrees on the left side is very big. This is the tendency to be scale-free. Since the range of degrees is small, the fit of MLE does not look perfect, but most of the points lie around the line.

We present the indicative p value of scale-free property of every network with respect to the time period, since every network has a scale-free indicator. In Table 115.1, K is the iteration times of the algorithm (for details, see [7]). We find that after 800 iterations, most of the p values converge. The p value ranges from 0 to 1. When p equals to 0, the network has no trend to be scale-free. While p gets closer to 1, the scale-free property of the network gets stronger.

With the scale-free p values of all the networks, we construct panel data about the alliances. And then we use negative binomial regression model to analyze the correlation between p value and the number of patents, with other factors such as network centrality controlled. The result presented in Table 115.2 is generated by Stata, where patent_i is the number of patents that the enterprise has applied in the ith year after formation of the alliance. As is shown in the table, we examine the effect during 3 years after the alliance's formation, which is a usual way to check the influence [14]. The result values in Table 115.2 are indicatives of how the data obey the negative binomial regression model. In statistics, it shows the significance of a regression model (not the scale-free indicator p value).

From Table 115.2, we find that the patents applied 1 year after the formation of the alliance have a stronger correlation with the scale-free property of the network than in other 2 years. This could be explained that the first year after the formation of the alliance is mainly used to exchange knowledge and information, and the second year is key time period for enterprises to present their innovative work, which in our situation, to apply for patents. But after these two time period, the effect of the alliance fades.

Table 115.2 Correlation significance

K	Patent_0	Patent_1	Patent_2
800	0.790	0.012*	0.646
1000	0.799	0.013*	0.648

† $p < 0.10$, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

115.5 Discussion and Conclusion

This paper mainly analyzes the relationship between the scale-free property of the alliance network and the innovation performance of the enterprise quantitatively. In statistics, it is seen that there exists correlation between scale-free and firm's innovative performance, and it happens a year after the alliance is formed (Table 115.2). In fact, the scale free property is depicted by the tendency to attach to the most influential firm in the industry. When new firms enter the alliance, they want their allied partners to be strong in technology, marketing or other aspects. So, the strong ones become stronger since almost all the other firms want to ally with them. This is the reason why the scale free property exists in the alliance network. It's like the Mathew Effect.

In a real alliance network, this is a situation when firms are allied to a few stronger ones because these few firms have more advanced technology or marketing. This situation is better in some way than the situation where firms are scattered to ally with some random firms, since in the former situation, weak firms can learn more from their allied partners. This is the reason for the significant impact of scale free property on firm innovation performance. To look at this question from another aspect, why cannot all alliances become very scale free since all of the companies want to ally with the strongest firm in the alliance? If so, the network should be starlike. The reason is that the stronger ones will consider their own conditions and do their own choices. They have this privilege to choose who they want to ally with, and they will choose the ones that will benefit them the most. Another reason is, that each company allies with a specific demand, the demand differs in aspects such as some specific technology the firm want to acquire, which may be very particular. Based on these reasons, a real alliance network is never so starlike. However, it can be judged that, if a real network is very scale free, it should be beneficial.

From the discussion above, it is known that the alliance action is quite a game alike thing, during which all the firms are chasing the most beneficial strategy. Further research will focus on the strategic balance among the alliance members, and attempt to reveal the factors that a company will take into consideration when they participate into the alliance.

Acknowledgments This work is supported by three projects of National Science Foundation of China (NSFC No. 71003069; 40976108) and Shanghai Leading Academic Discipline Project (J50103).

References

1. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**, 47 (2002)
2. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. **45**, 167–256 (2003)

3. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**, 509 (1999)
4. Barabási, A.L., Crandall, R.: Linked: the new science of networks. Am. J. Phys. **71**, 409 (2003)
5. Newman, M.E.J., et al.: Email networks and the spread of computer viruses. Phys. Rev. E **66**, 035101 (2002)
6. Klemm, K., Eguiluz, V.M.: Highly clustered scale-free networks. Phys. Rev. E **65**, 036123 (2002)
7. Clauset, A., et al.: Power-law distributions in empirical data. SIAM Rev. **51**(4), 661–703 (2009)
8. Liu, C.H.: The effects of innovation alliance on network structure and density of cluster. Expert Syst. Appl. **38**(1), 299–305 (2011)
9. Du, R., et al.: Relationship between knowledge sharing and performance: a survey in Xi'an China. Expert Syst. Appl. **32**, 38–46 (2007)
10. Jiang, X., Li, Y.: An empirical investigation of knowledge management and innovative performance: the case of alliances. Res. Policy **38**, 358–368 (2009)
11. Verspagen, B., Werker, C.: Keith Pavitt and the invisible college of the economics of technology and innovation. Res. Policy **33**, 1419–1431 (2004)
12. Gay, B., Dousset, B.: Innovation and network structural dynamics: study of the alliance network of a major sector of the biotechnology industry. Res. Policy **34**, 1457–1475 (2005)
13. Hausman, J.A., et al.: Econometric models for count data with an application to the patents-R&D relationship. National Bureau of Economic Research, Cambridge (1984)
14. Schilling, M.A., Phelps, C.C.: Interfirm collaboration networks: the impact of large-scale network structure on firm innovation. Manag. Sci. **53**, 1113–1126 (2007)
15. Lin, C., et al.: The alliance innovation performance of R&D alliances—the absorptive capacity perspective. Technovation **32**, 282–292 (2012)

Chapter 116
Dynamic Analysis on Significant Risk of Innovative Enterprise During the Strategic Transformation Period

Zejian Li and Hongwu Zuo

Abstract Sustainable innovation is very important for innovation enterprise's development and also difficult to obtain for the risk which is complex and changeful, especially in strategic transformation period. In order to help the innovation enterprise get deeper understanding about dynamic changeable significant risks in the process of enterprise sustainable innovation, this paper constructs innovative enterprise's sustainable innovation strategic transition process model process model, analyzes the main factors of significant risks during the strategic process of innovative enterprise and summarizes the significant risks of innovative enterprise into three categories—innovation strategic risk, manager risk and major innovation projects risk. By dividing the stages of strategic management into strategic analysis phase, strategic selection phase and strategic implementation phase, the paper analyzes the significant risk of innovative enterprise, and reveals the dynamic variation regulation of the three significant risks during the strategic transformation period.

Keywords Innovative enterprise · Significant risk · Strategic transformation · Dynamic analysis

Z. Li (✉)
Faculty of Management & Economics, Kunming University of Science & Technology, Kunming, China
e-mail: lucy_lizejian@hotmail.com

Z. Li
Institute of Technology, Kunming University of Science &Technology, Kunming, China

H. Zuo
Faculty of Adult Education, Kunming University of Science & Technology, Kunming, China
e-mail: dbminer2@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_116,

116.1 Introduction

Internationally renowned scholars Ikujiro Nonaka once pointed out, "The fundamental task of the innovative enterprises is to achieve sustainable innovation". The Enterprise Sustainable Innovation (ESI) process is such a process that has (had) continually implemented innovation projects of introducing new products/new process techniques/developing new markets/acquiring new materials sources/realizing new organization or new institution and/or their inside diffusions in the enterprise for a long period and made a continually significant economic development of the enterprise during the period [1]. However, the ESI process is full of risks and barriers. There are many risks causing ESI process interruption and much more serious consequence. Once significant risk get out of control and outbreak, the significant innovation projects, and even the implementation of innovative strategies will result to huge losses, and may even cause the interruption or termination of sustainable innovation process the innovative enterprises.

To face the challenges of global competition and the complex financial, economic, technology and resource environment, Chinese local enterprises are doing the implementation of the strategic transformation to seek sustainable survival and development. It is a strategic transition period now, a transition period for product upgrading to brand, and a regional market transition period towards international as well [2]. The risk Chinese enterprises are facing is becoming more complex and changeable.

Based on this background, this paper analyzes the factors, mutual interaction, and dynamic variation regulation of significant risk during the strategic process of innovative enterprise, and do help to control with significant risks in strategic transformation and sustainable innovation.

In this paper, innovative enterprise is defined as some enterprise has independent innovation strategy and has a powerful and long-lasting power of innovation. Significant risk of innovative enterprise refers to the risks which have a significant negative impact on the sustainable innovation process of innovative enterprises.

116.2 The ESI Strategic Transition Process

Due to the Enterprise Sustainable Innovation (ESI) process is a continuous process in time and outputs for a long period. It has its life-cycle. Sustainable innovation evolution characteristics lies on the ESI process is dynamic integration of a variety of types of innovation, a number of innovative projects cluster. During the period, the ESI process can be divided to some stages and has different innovation strategy in every stage.

W. J. Abernathy and J. M. Utterback thought enterprise innovation is doing in a continuous mode, which can be divided into unstable stage, transition stage and mature stage [3]. X. Gang built a sustainable innovation process stage development

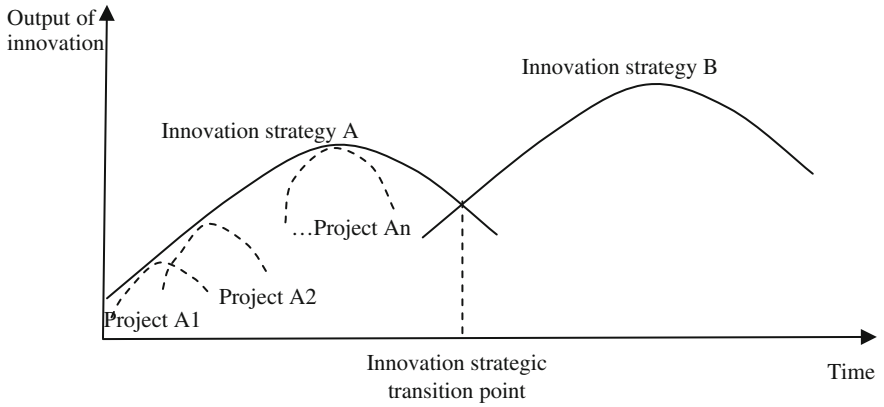


Fig. 116.1 ESI strategic transition process

model and divided sustainable innovation process into unstable stage, transition stage and stable development stage [1]. Sustainable innovation process was divided into the start-up stage, the stage of large-scale development and the stage of international development by Y. Yijie [4].

Refer to the above research results, as well as the ESI practical experience in China, this paper agrees that ESI process composed by three stages—the initial stage, the large-scale development stage and the international development stage. Each stage has its strategy. The ESI process of innovative enterprises can be understood for the continuous successful implementation of the innovation strategy to bring sustainable economic benefits growth, which is shown in Fig. 116.1. There are many innovation projects to support every innovation strategy. At some point, it will transfer to a new innovation strategy when it entered a new stage. It can be called "Innovation strategic transition point". Actually, it is maybe a period—"the strategic transformation period".

The ESI process of innovative enterprises can be understood for the continuous successful implementation of the innovation strategy. That is, innovative enterprises will face several innovation strategies' formulation, implementation and control in the ESI process. In fact, each strategic stage is including a new strategic decision-making phase and strategic implementation phase. So we can divide each strategic stage into strategic transformation period (1–3 years) and strategic stability implementation period (3–5 years). Strategic transformation period can be divided into new strategic formulation phase and the early stage of new strategic implementation, with many uncertainties, and which is full of risk. When a new strategy was implemented for some time, through validation, modify, supplement and perfect from the practice, it will be a gradual transition to the strategic stability implementation period.

116.3 The Factors Analysis of Significant Risk of Innovative Enterprise

Innovation has become an enterprise essential function rather than happenstance, which is the basis for sustainable innovation of innovative enterprises. For creating a long-term sustainable innovation system, there must be a clear strategy which the whole company is widely recognized [5]. Some studies show that the lack of leadership of sustainable innovation system is very dangerous [6].

Based on the above analysis, this paper summarized the significant risks of innovative enterprise into three categories—innovation strategic risk, manager risk and major innovation projects risk which is shown in Fig. 116.2. As a guiding directional outline of enterprise, a wrong strategy will only lead to further and further away in the wrong direction. So innovation strategy is a very important factor to determine innovation activities success or failure in innovative enterprises. The major innovation projects relying on enterprise innovation strategic objectives are a core part of the enterprise strategy implementation. In addition, due to the selection and implementation of the strategy and major innovation project are both guided by enterprise managers, which led to manager risk not only directly affect the enterprise, but also indirect effect the enterprise by the impact of innovation strategic decision-making, or by the impact of the implementation and control of major innovation projects, to cause significant losses.

The three types of significant risks are not independent. They contact each other and showed a non-linear interaction mechanism, which constitutes a significant innovation risk system during the ESI process.

It needs to be pointed out that, the above three significant risks interaction of innovative enterprise exist in the whole process of ESI, exist in every strategic development phase, and even exist in strategic transformation period which is extremely critical in every strategic development phase.

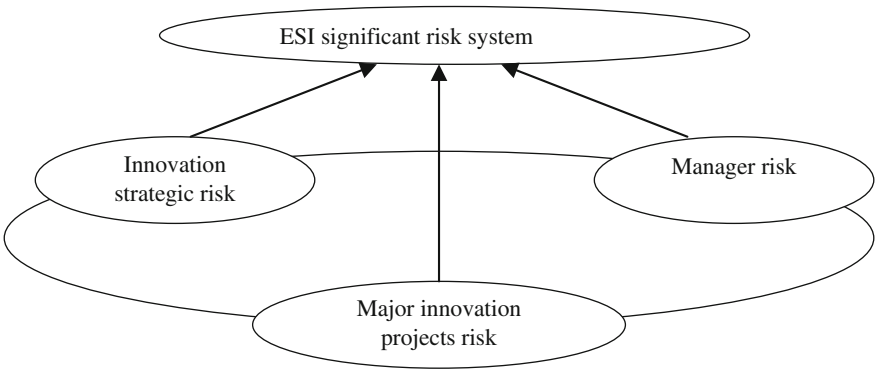


Fig. 116.2 Factors analysis of ESI significant risk system

Table 116.1 Significant risks in phases of the strategic management process

Strategic analysis phase	Strategic selection phase	Strategic implementation phase
Information risks	Manager risk Information risks Risk appetite	Changes of the external environment
Strategic risks	Strategic risks	Significant innovation projects risks

116.4 Dynamic Variations of the Significant Risks of Innovative Enterprise During the Strategic Transformation Period

As mentioned above, in each stage of ESI process, it can be assumed that each strategy includes strategic transformation period (strategic analysis phase, strategic selection phase and strategic implementation initial phase) and strategic stability implementation period. Strategic risks are closely linked with the phases of the strategic management. The significant risks of innovative enterprise during the strategic transformation period could be analyzed according phases with the strategic management process, which is shown in Table 116.1.

116.4.1 Significant Risks Analysis in Phases of Strategic Management

116.4.1.1 Significant Risks Analysis in Strategic Analysis Phase

Mainly by the impact of strategic analysis organization and strategic analysts as well as their own quality impact, it mainly exists on information risks, personnel risks and strategic objectives risks. By the impact of comprehensiveness, reliability and accuracy of strategic analysts on the full grasp degree of information, information risk may exist in strategic analysis phase. And by the impact of strategic analysts' own structure of knowledge, ability, experience, there may be personnel risks of strategic analysts [7]. These two risks ultimately affect the realization of strategic objectives through strategic selection and implementation.

116.4.1.2 Significant Risks Analysis in Strategic Selection Phase

The essence of strategic selection is strategic decision-making, it mainly exists manager risk and strategic objectives risks affected by the impact of the strategic decision-making information and strategic decision-makers. Due to the selection and implementation of strategy are guided by manager, there may be information risks affected by the manager's grasping of information. The risk appetite of the

managers, decision-making style, as well as insight into the environment, the law of industry development, and the ability of its own resource awareness, there may be the risk of manager ability. Both risks eventually affect the realization of strategic objective through strategic selection and implementation.

116.4.1.3 Significant Risks Analysis in Strategic Implementation Phase

Perfect strategy which has not been effective implementation will also affect the strategic objectives. The success of innovation strategy relying on innovative projects, significant innovation project is a core part of strategy implementation. The uncertainty of the process of strategy implementation may come from changes of the external environment, the significant innovation projects and the strategy implementation personnel.

116.4.2 Dynamic Variation of the Significant Risks During the Strategy Transformation Period of Innovative Enterprise

Through the above analysis it can be revealed that the three types significant risk of innovative enterprise during strategic transformation are innovation strategic risks, manager risks and significant innovation projects risks. The dynamic variation regulation is shown in Fig. 116.3.

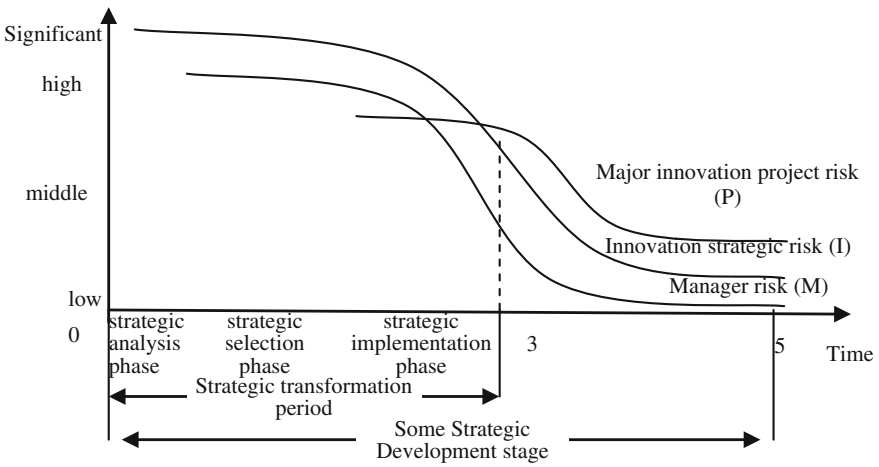


Fig. 116.3 Dynamic changes curve of three significant risk during the strategic transformation period

(1) In the stage of development of a sustainable innovation strategy, the variation of innovation strategy (own) risk is shown in curve I of Fig. 116.3. In the stage of strategic transformation period of a sustainable innovation strategy (here, the strategic transformation period to develop a new innovative strategy, as well as the period of the early years of the implementation of this new strategy, generally refers to the stage of an first 1–3 years innovation period), innovation strategic risk maintain the state of the high-level. Affected by the strategic decision-making information and strategic decision-makers ability, the strategic decision-making errors may bring huge losses to the enterprise. At the early beginning of a new strategy implementation, the enterprise needs an adaptation period, so innovative strategic risk maintains at a high state. With the continuous progress of the implementation process of innovation strategy, enterprise adapt to the new strategy gradually, and gradually into the strategic stability period, when the innovation strategy risk will also gradually declined from a high level to low-risk status.
(2) The variation of manager risk is shown in curve M of Fig. 116.3. In the stage of development of a sustainable innovation strategy, manager risks maintain the state of the high-level. Since the analysis, selection and implementation of innovation strategy during this period are led by the managers, it may bring huge losses to the enterprise if the strategic decision-making error occurs. With the continuous progress of the implementation process of innovation strategy, managers can better overall grasp the enterprise, when the manager's ability risk will at moderate level which is shown in Fig. 116.3 that curve M gradually declined from a high level to low-risk status.
(3) The variation of major projects risks is shown in curve P of Fig. 116.3. The risk of major innovation projects starting at the initial stage of the implementation of the strategy has been maintained at a high risk status. The major innovation project is a core part of the implementation of the strategy due to the implementation of innovative strategy is ultimately fall to the projects. The decision-making mistakes early in a new major project will lead to innovative strategies failure and enterprise's losses. In the implementation stage and the final stage, risks will gradually decrease as shown in curve P of Fig. 116.3 gradually declined from a high-risk state to low-risk state.

116.5 Conclusion

Chinese innovation enterprises are doing the implementation of the strategic transformation to seek sustainable development. The ESI process of innovative enterprises can be understood for continuous successful implementation of the innovation strategy. The significant risks of innovative enterprise during strategic transformation period include innovation strategic risk, manager risk and major innovation projects risk.

It is expected to have more cases to verify. By dividing the phase of strategic management into strategic analysis phase, strategic selection phase and strategic implementation phase, the significant risks of innovative enterprise were analyzed, and the dynamic variation of the above three types of significant risk were summarized and revealed on its basis.

Acknowledgments This paper is the production financed by Yunnan Provincial educational bureau program (Grant No. 2010Y411) and National Natural Science Foundation program of China (Grant No. 70862002).

References

1. Gang, X.: Enterprise Sustainable Innovation. Science Press, Beijing (2006)
2. Chunyu, P., Ge, Z.: Brand transformation: Chuijin Huangsha Shidaojin. Sales Mark. **2**, 22–32 (2012)
3. Abernathy, W.J., Utterback, J.M.: Patterns of industrial innovation. Technol. Rev. **80**(6–7), 40–47 (1978)
4. Yijie, Y.: Research on financial risks identification dynamically and countermeasures of sustainable innovation process. Master's degree thesis of Kunming University of Science and Technology, Kunming (2010)
5. Shanshan, Z., Yulin, Z.: Sustainable innovation: the new trend of innovative research abroad. Stud. Dialectics Nat. **1**, 68–71 (2007)
6. Lindgren, P., Bohn, K., Sorensen, B.V.: Network based product development leadership and management—the impact on short and long term continuous innovation. In: The Proceeding of 5th International CINet Conference, **81**(9), 456–470 (2004)
7. Yingqiu, S.: Strategic risk identification model construction. Technoecon. Manage. Res. **1**, 69–73 (2011)

Chapter 117
The Development of State-Level Science and Technology Industrial Parks in China: Industry Clustering or Enterprises Gathering?

Qiang LI

Abstract This paper provided a model framework to differentiate industry clustering. Based on CD production function with external science and technology progress efficiency, the study investigated geographical concentration of production factors and increasing return to scale in High-tech Zones in China. Result shows that there is industry clustering in the zones. However, the returns to scale of industry clusters differ greatly between the zones.

Keywords High-tech zone · Industry cluster · Production factors · Increasing return to scale

117.1 Introduction

As China aggressively seeks to transform R&D achievements into real productivity in reliance on its own S&T capability and economic strength, 53 state-level science and technology industrial parks (STIPs) are established in knowledge-intensive cities. Considerable resources are being devoted to science parks as policy instruments aimed at promoting R&D-based as well as innovation activities. Advanced foreign multinational company production facilities are also attracted to nurture native enterprises by promoting geographic proximity. From 1995 to 2005, the number of enterprises and employee in STIPs increased at annual rate of 12.46 and 18.18 % separately. At the same time, the revenue of enterprises in STIPs reached US $42.65 billion, about 22.5 times that of 1995.

STIPs in China have attached much importance to linkages between high-tech development and market demand both at home and abroad. The effeteness of

Q. LI (✉)
Institute of Policy and Management, Chinese Academy of Sciences, Beijing, China
e-mail: lq@casipm.ac.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_117,

Chinese STIP model lies in its mix of local and foreign forms of investment and in the role of universities in nurturing native companies through information networks and entrepreneurship training which is bridging the gap [1]. For example, serving as the base for high-tech industrialization, R&D, incubator and high-tech talents training, Shenzhen Science and Technology Industrial Park (SHIP) has attracted a large number of hi-tech firms both from domestic and overseas. A high degree of hi-tech firm concentration, principally multinationals in IT and optical-electronics, contributes significantly to economic growth [2]. Located northwest Beijing in proximity with over 30 universities and 200 more national research institutes geographically, Zhongguancun Science & Technology Industrial Park (ZGCSTIP) is regarded as one of the typical high-tech clusters in the world [3]. The park has 626,000 staff employed in 17,000 firms with revenue of 60 billion US$ which is one-seventh of all 53 STIPs (2005).

Much literature supports the argument that a positive relationship can be found between Industrial cluster and productivity in economies [4]. Fan and Scott showed that many kinds of manufacturing sectors are characterized by a strong positive relationship between spatial agglomeration and productivity in China, especially sectors and regions where liberalization has proceeded rapidly [5]. Further study by Liang and Zhan based on the 1998–2003s data of 16 cities in Yangtze River delta proved that industry clusters are experiencing higher increase rate of technology progress and its contribution ratio to economy, where regional specialization could promote industry from a labor-intensive one to a technology-intensive one, driving industry's technology progress [6].

The "cluster" concept has become a new idea and policy tool in regional economic and technological development of China. However, the rush to employ "cluster" idea in the planning and policy-making is problematic [7]. In pursue of industry clustering, 4,210 development zones (many of them are high-tech zones like STIP) are established by provincial, municipal, county and even township governments, which makes STIPs facing failure in attracting foreign high-tech firms and overseas high-tech personnel into these parks. According to Xia's comprehensive assessment, only four of 53 STIPs are in good performance category and others performed poorly or very poorly [8]. Further study by Han Bo-tang and Li Qiang proved more polarized development of STIPs since late 1990s [9, 10]. A comparison of innovation capacity at Shanghai Zhangjiang Science & Technology Industrial Park (SHZJSTIP) and Hsingchu Science-based Industrial Park (HSIP) showed that HSIP was significantly preferable to SHZJSTIP for the four drivers of factor conditions, demand conditions, related and supporting industries and context for firm strategy, the determinants of innovation orientation of national industry cluster [11]. Macdonald and Deng argued that science parks in China could not promise in terms of growth and employment achieved through providing new, high technology companies with an ideal location, and there is little evidence that clustered in pleasant surroundings alongside a university or research centre, entrepreneurs would be able to transform their ideas into innovations [12].

Commented as the most important creation in industrialization of science and technology, STIPs are expected to combine scientific & technological activities with industrial development and solve the problem of separation of science & technology from economy and make the discoveries and inventions of humankind smoothly convert into economic and social benefits. A literature review shows that there have been much assessment and descriptive discussion about the development of STIPs. However, not much research attention has been paid to quantitative analysis of industry cluster in STIPs and their economic effects directly. Based on panel data from 53 Chinese STIPs and their mother cities for the 1991–2004 periods, this paper will start with a survey on the process of geographical concentration of production in STIPs, and then discuss whether this concentration can contribute to increasing returns to scale-to distinguish industry clustering from enterprises gathering. The rest part of this paper will focus on the relationship between industrial clustering and technology development and try to compare industry-clustering standards among STIPs.

117.2 Geographical Concentration of Production

To examine the geographical concentration of Production in STIPs, we can suppose an economy of two regions (STIP and its mother city), both capable of producing manufacturing goods and providing a variety of services at a certain level of technology. Suppose that all goods and services are produced with the two primary inputs, capital (K) and labor (L). Providing that labor may move between the regions in response to differences of wages and the invested capital (investment) is fixed. New investment and fresh labor will move into the regions for utility maximization. The invested capital is split between both regions into shares $\lambda(t)$ for STIP and $(1 - \lambda(t))$ for the mother city as follows:

$$\lambda(t) = \frac{K_{STIP}(t)}{K_{STIP}(t) + K_{mother\,city}(t)} \tag{117.1}$$

In Eq. (117.1), $K_{STIP}(t)$ and $K_{mother\,city}(t)$ stands for annual average total assets of enterprises registered in STIP and its mother city separately. According to *Statistics on China Torch Program* (1991–2004) and statistical data of the mother cities, we can get the shares of capital in STIPs and their time series from 1996 to 2004 (see Table 117.1).

Hierarchical Cluster Analysis (HCA) is an unsupervised technique that combines individual samples into clusters according to their similarities to each other. The similarity between two samples is determined in terms of their nearness in the multidimensional space. In this work, the distance matrix of λ (2004) and $\Delta\lambda/\Delta t$ was calculated in Euclidean distances using the SPSS 13.0 software (Fig. 117.1).

We can identify four major clusters from Fig. 117.2. Cluster A contains four STIPs of Chengdu, Changsha, Xi'an and Beijing. Cluster B Contains nine STIPs of

Table 117.1 Time series of $\lambda(t)$ in STIPs

Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	$\Delta\lambda/\Delta t$
Beijing Zhongguancun	0.156	0.127	0.180	0.258	0.354	0.456	0.476	0.503	NA	0.061
Chengdu	0.077	0.083	0.128	0.158	0.205	0.304	0.313	0.437	NA	0.051
Xi'an	0.118	0.193	0.223	0.286	0.319	0.393	0.405	0.450	NA	0.047
Changsha	0.149	0.193	0.272	0.317	0.360	0.409	0.421	0.438	NA	0.043
Zhuzhou	0.072	0.119	0.145	0.171	0.208	0.297	0.334	0.352	NA	0.042
Nanning	0.070	0.097	0.136	0.180	0.203	0.463	0.279	0.258	NA	0.039
Changchun	0.133	0.167	0.214	0.215	0.300	0.313	0.369	0.382	NA	0.037
Zhuhai	NA	0.025	0.028	0.060	0.081	0.115	0.217	0.253	0.230	0.037
Wuhan Donghu	NA	NA	0.106	0.146	0.191	0.231	0.260	0.278	NA	0.036
Luoyang	NA	NA	0.011	0.012	0.013	0.013	0.021	0.200	0.211	0.035
Haikou	0.090	0.130	0.208	0.243	0.234	0.302	0.309	0.323	NA	0.033
Nanchang	NA	NA	0.123	0.159	0.211	0.270	0.295	0.296	0.274	0.029
Guilin	0.161	0.193	0.234	0.225	0.288	0.304	0.313	0.368	NA	0.023
Jinan	0.073	0.125	0.149	0.156	0.148	0.188	0.268	0.253	0.252	0.023
Nanjing	NA	NA	0.120	0.162	0.156	0.186	0.216	0.234	NA	0.022
Zhengzhou	NA	NA	0.094	0.122	0.193	0.194	0.209	NA	0.228	0.022
Guangzhou	NA	0.032	0.034	0.045	0.063	0.099	0.125	0.152	0.155	0.021
Lanzhou	NA	NA	0.056	0.080	0.107	0.152	0.156	0.143	NA	0.020
Wulumuqi	0.032	0.057	0.056	0.065	0.076	0.079	0.267	0.101	NA	0.019
Wuxi	NA	NA	0.100	0.103	0.131	0.152	0.187	0.194	0.185	0.018
Mianyang	0.194	0.263	0.368	0.351	0.325	0.348	0.342	0.342	NA	0.016
Haerbin	0.191	0.217	0.254	0.247	0.284	0.316	0.297	0.318	0.319	0.016
Zibo	0.080	0.082	0.165	0.166	0.138	0.174	0.182	0.202	0.224	0.016
Hangzhou	NA	0.060	0.067	0.071	0.088	0.111	0.118	0.152	NA	0.015
Chongqing	NA	0.043	0.076	0.077	0.097	0.131	0.142	0.145	0.141	0.015
Qingdao	0.081	0.069	0.109	0.136	0.119	0.156	0.161	0.197	0.151	0.013
Huizhou	NA	0.095	0.112	0.200	0.143	0.122	0.169	0.186	0.210	0.013
Weihai	0.072	0.080	0.088	0.133	0.140	0.151	0.155	0.166	0.168	0.013
Tianjin	0.055	0.049	0.074	0.081	0.087	0.117	0.135	0.144	0.137	0.013
Changzhou	NA	0.042	0.060	0.084	0.097	0.107	0.125	0.132	0.132	0.013
Shanghai Zhangjiang	0.030	0.030	0.071	0.093	0.104	0.099	0.115	0.124	0.137	0.013
Hefei	0.169	0.204	0.456	0.169	0.167	0.287	0.293	0.302	0.329	0.012
Weifang	0.070	0.084	0.082	0.109	0.080	0.107	0.118	0.167	0.159	0.011
Shenzhen	NA	0.043	NA	0.098	0.132	0.141	0.143	0.107	0.116	0.009
Zhongshan	NA	0.131	0.113	0.171	0.150	0.169	0.159	0.178	0.186	0.008
Foshan	NA	0.051	0.042	0.087	0.093	0.083	0.094	0.087	0.100	0.007
Suzhou	NA	NA	0.122	0.110	0.109	0.112	0.114	0.140	0.126	0.003

Data source China Statistical Yearbook 1996–2004, *Statistics on China Torch Program* 1996–2004. The list is incomplete because certain data are not available or are incomplete for some STIPs or the mother cities.

Changchun, Guilin, Zhuzhou, Wuhan Donghu, Nanchang, Haerbin, Hefei, Haikou, and Mianyang. Except for Haikou STIP, all STIPs in Cluster A and B are located in inland China. Cluster C and D contain the majority of STIPs located in Circum-Bohai Economic Ring (Jinan, Zibo, Qingdao, Weihai, Weifang, Tianjing), Pearl

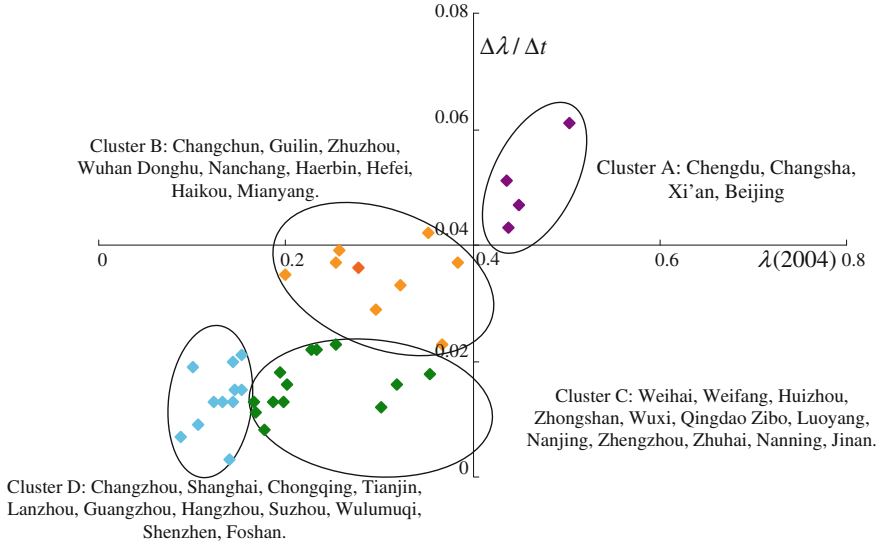


Fig. 117.1 Samples in the space defined by λ (2004) and $\Delta\lambda/\Delta t$

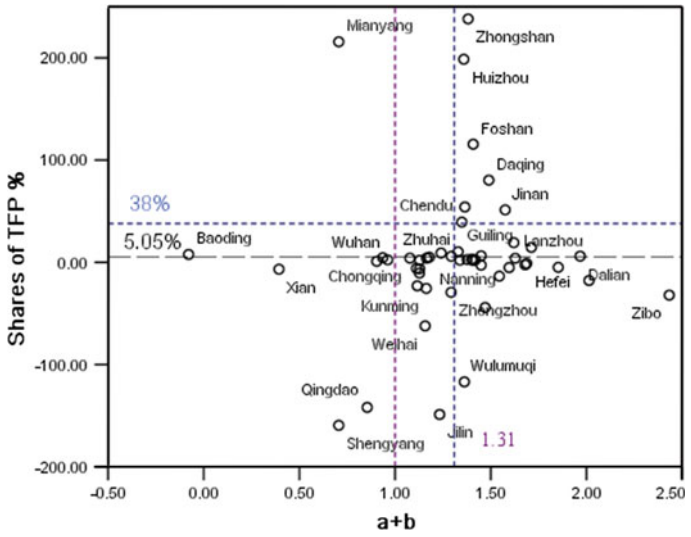


Fig. 117.2 Average aggregated output elasticity of K, L and the share of TFP in STIPs 1996–2004

River Delta Economic Zone (Zhuhai, Huizhou, Zhongshan, Guangzhou, Shengzhen, Foshan) and Yangtze River Delta Economic Zone (Nanjing, Wuxi, Hangzhou, Changzhou, Shanghai Zhangjiang, Suzhou).

It is clear that STIPs have been attracting much more investment than their mother cities. In inland China, STIPs can usually provide enterprises and entrepreneurs with more preferable business environment than their mother cities. Besides of state policies concerning STIPs, local governments often carry out preferential policies such as tax-cut, government subsidy, land rentals and so on. As the result of preferential policies, STIPs are regarded as growth pole of hi-tech industry and local economy, where the share of local invested capital (annual average total assets) had tripled to over 40 % in Beijing, Chengdu, Xi'an, and Changsha from 1996 to 2003. As to economically developed eastern coastal regions, where the reform and opening-up policy was initially implemented, the discrepancy was not significant, particularly STIPs around Shanghai and the Pearl River.

Similarly, we can get STIPs' share of labor and their time series from 1991 to 2004 by dividing labors into shares of $\varphi(t)$ and $[1-\varphi(t)]$ as formula (117.2), Where $L_{STIP}(t)$ and $L_{mother\,city}(t)$ stands for annual average employees of enterprises registered in STIP and its mother city separately. The results are given in Table 117.2. Clearly, we can see that STIPs have been experiencing more rapid inflow of labor force than their mother cities. However, STIPs only take very small fraction of employed population.

$$\varphi(t) = \frac{L_{STIP}(t)}{L_{STIP}(t) + L_{mother\,city}(t)} \tag{117.2}$$

Due to the limitation of data, this study cannot characterize the degree of geographical concentration of production in concentration ratios, such as Location quotient (*LQ*), spatial Gini coefficient (*G*), or Index of Relative Concentration (*ICR*) etc. However, the rapid increase of STIPs' share of invested capital (1996–2004) and employed people (1991–2004) demonstrated that geographical concentration of production is shaping up.

117.3 Increasing Returns to Scale of Production

An industry cluster is a group of firms, and related economic actors and institutions, that are located near one another and that draw productive advantage from their mutual proximity and connections [13]. In the emerging body of ideas given by "new economic geographers" [14], industry cluster matter for regional economic by generating agglomeration effects. Agglomeration effects, in turn, are often categorized as so-called localization economies (i.e., efficiency-boosting phenomena that come from the clustering of firms in a given sector) and as urbanization economies (i.e., efficiencies that result from the agglomeration of many different kinds of activities in a given region) which means increasing returns to scale can gain from industrial connections, inter-firm relationships, industrial organization, the mechanism of learning-by-doing and innovation.

Table 117.2 Time series of $\varphi(t)$ in STIPs

Year	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	$\Delta\varphi/\Delta t$
Guilin	NA	NA	0.020	0.034	0.027	0.040	0.055	0.062	0.046	0.046	0.069	0.073	0.120	NA	0.0068
Suzhou	NA	0.001	0.007	0.025	0.026	0.038	0.046	0.061	0.054	0.065	0.057	0.083	0.078	0.121	0.0067
Beijing	0.001	0.009	0.019	0.023	0.021	0.025	0.025	0.037	0.054	0.064	0.077	0.079	0.092	0.082	0.006
Xi'an	0.002	0.005	0.013	0.020	0.021	0.020	0.024	0.035	0.045	0.055	0.074	0.081	0.091	NA	0.0058
Changzhou	NA	NA	0.010	0.011	0.012	0.037	0.028	0.033	0.040	0.063	0.066	0.075	0.077	NA	0.0054
Changsha	0.003	0.007	0.008	0.010	0.012	0.016	0.022	0.032	0.039	0.043	0.060	0.067	0.071	NA	0.0047
Zhuzhou	0.001	0.003	0.003	0.006	0.006	0.008	0.012	0.021	0.030	0.037	0.061	0.079	0.091	NA	0.0047
Chengdu	0.002	0.004	0.006	0.010	0.014	0.015	0.011	0.018	0.027	0.031	0.052	0.056	0.063	0.068	0.0039
Qingdao	0.000	NA	0.002	0.012	0.040	0.045	0.017	0.035	0.042	0.033	0.034	0.055	0.056	0.032	0.0038
Tianjin	0.003	0.006	0.007	0.011	0.013	0.020	0.022	0.029	0.034	0.041	0.044	0.048	0.042	0.053	0.0037
Nanjing	NA	NA	0.004	0.008	0.014	0.020	0.025	0.023	0.036	0.036	0.039	0.043	0.046	NA	0.0035
Hangzhou	NA	NA	NA	NA	0.011	0.012	0.014	0.018	0.021	0.027	0.032	0.045	0.060	NA	0.0032
Nanchang	0.001	0.003	NA	0.006	0.008	0.011	0.015	0.016	0.021	0.031	0.039	0.055	0.052	0.045	0.0032
Nanning	NA	NA	0.005	0.006	0.008	0.008	0.018	0.022	0.023	0.032	0.047	0.047	0.043	NA	0.0032
Daqing	NA	0.000	NA	0.008	0.011	NA	0.022	0.021	0.018	0.023	0.046	0.047	NA	NA	0.0031
Haerbin	0.001	0.004	0.005	0.006	0.019	0.022	0.022	0.028	0.029	0.028	0.027	0.033	0.035	0.040	0.0028
Haikou	NA	NA	NA	NA	NA	0.001	0.006	0.012	0.014	0.028	0.030	0.031	0.033	NA	0.0022
Chongqing	0.001	0.003	0.006	0.009	0.009	0.005	0.005	0.011	0.016	0.016	0.017	0.020	0.019	0.018	0.0015
Shanghai	0.003	0.006	0.010	0.009	0.009	0.008	0.011	0.014	0.022	0.012	0.011	0.014	0.014	0.018	0.0014

With economic activity concentrating, much study focus on factor accumulation and its scale effects. By applying Solovian growth accounting methodology, Kremer has noted and modeled the important externalities of allocative efficiency and scale effects [16]. Elias proved increasing returns to scale in factor accumulation [17]. Easterly and Levine demonstrated that stylized facts do not support diminishing or constant returns to scale in factor accumulation [18]. However, the "residual" rather than factor accumulation accounts for most of the income and growth differences across nations.

To estimate whether there is increasing returns to scale of production in STIPs, we use an extension of the growth accounting approach long associated with Solow and Dennison. In this earlier literature, an aggregate production function is assumed of the form,

$$Y(t) = A(t)F(K(t), L(t)) \tag{117.3}$$

Among types of empirical studies of economic growth, aggregate production are often given as the standard Cobb-Douglas aggregate production function with constant return to scale [19],

$$Y(t) = A(t)K(t)^{\alpha}L(t)^{1-\alpha} \tag{117.4}$$

And with exogenous technological progress [20],

$$Y(t) = A_0 e^{gt}K(t)^{\alpha}L(t)^{1-\alpha} \tag{117.5}$$

Or with increase return to scale,

$$Y(t) = AK(t)^{\alpha}L(t)^{\beta} \tag{117.6}$$

In Eqs. (117.3)–(117.6), Y represents output, K and L represent capital and labor inputs, i.e. real GDP, real capital stock and total employment. The variable t is a time index, α is the contribution of capital to output, $1 - \alpha$ or β is the contribution of labor and the expression of A, A (t) or $A_0 e^{gt}$ is TFP. TFP means technological progress and other elements that affect the efficiency of the production process, measure the shift in production function at given levels of capital and labor.

Besides the above-mentioned equations, varieties of techniques have been used to decompose productivity growth and measure TFP in the literature (for details see Fried et al. 1993). However, the estimation of an aggregate production function confronts the researcher with various problems. The literature to date is inconclusive on the best method to decompose TFP and estimate elasticity, especially time-series analysis on panel data.

While this has been noted by Hulten [20] and Akinlo [21], the possible time-dependent agglomeration effects and endogeneity of capital and labor can be highly correlated with capital or labor in the time series, and thus the value of total factor productivity obtained might be underestimated. To provide clear evidence for these problems, we carry out further investigation into panel data of 51 Chinese STIPs as follows:

Transforming Eqs. (117.5) and (117.6) in logarithmic form:

$$\ln Y(t) - \ln L(t) = \ln A_0 + gt + \alpha_0(\ln K(t) - \ln L(t)) \tag{117.7}$$

$$\ln Y(t) = \ln A + a \ln K(t) + b \ln L(t) \tag{117.8}$$

We use annual revenue and annual average fixed assets of STIP i to express output $Y_i(t)$ and capital input $K_i(t)$, which are calculated at constant prices of 1990 in 1 million Yuan. Moreover, labor input is expressed by employment in person. Take a total derivative through Eqs. (117.7)–(117.8) and yield growth accounting equations as follows:

$$\frac{dY(t)}{Y} - \frac{dL(t)}{L} = g + \alpha_0\left(\frac{dK(t)}{K} - \frac{dL(t)}{L}\right) \tag{117.9}$$

$$\frac{dY(t)}{Y} = \frac{dA}{A} + a\frac{dK(t)}{K} + b\frac{dL(t)}{L} \tag{117.10}$$

And the shares of capital (E_K), labor (E_L) and TFP can thus be expressed as:

$$TFP = 1 - E_K - E_L \tag{117.11}$$

$$\begin{aligned} E_{0K} &= \alpha_0 \frac{K_t - K_{t-1}}{Y_t - Y_{t-1}} \frac{Y_{t-1}}{K_{t-1}} \\ E_{0L} &= (1 - a_0) \frac{L_t - L_{t-1}}{Y_t - Y_{t-1}} \frac{Y_{t-1}}{L_{t-1}} \end{aligned} \tag{117.12}$$

$$\begin{aligned} E_K &= a \frac{K_t - K_{t-1}}{Y_t - Y_{t-1}} \frac{Y_{t-1}}{K_{t-1}} \\ E_L &= b \frac{L_t - L_{t-1}}{Y_t - Y_{t-1}} \frac{Y_{t-1}}{L_{t-1}} \end{aligned} \tag{117.13}$$

The estimated average shares of TFP in STIPs' output growth from 1996 to 2004, along with aggregated output elasticity and output elasticity of K are given in Fig. 117.2.

Results in Fig. 117.2 display huge differences among STIPs both in share of TFP and aggregated output elasticity. For example, Zhongshan and Mianyang had experienced TFP shares of 238 and 215 % in output growth contrasting with −159 and −148 % that of Shenyang and Jilin. In spite of the fact that 53 STIPs account for two-thirds more of china high-tech products, 50 STIPs had experienced aggregate output elasticity of 1.31 and TFP shares of 5.05 % in average. This is much lower than the estimation of the average level in China.

We introduce an extended Cobb-Douglas production function, this form after accounting for time series (t) and STIP (i) is:

$$Y_i(t) = A_i e^{g_i t} K_i(t)^{\alpha_i(t)} L_i(t)^{\beta_i(t)} e^{\varepsilon_{it}} \tag{117.14}$$

Or by transforming Eq. (117.4) in logarithmic form:

$$\ln Y_i(t) = \ln A_i + g_i t + \alpha_i(t) \ln K_i(t) + \beta_i(t) L_i(t) + \varepsilon_{it} \tag{117.15}$$

where A_i stands for a given level of technology and management level in STIP (i), g_i is a measure of disembodied technical change in output per time.

Considering the fact that STIPs are all newly developed, where the invested capital increased eight times and the employees of enterprises increased three and a half accordingly from 1996 to 2004. We can suppose that all STIPs are developed from the same technology level (A_0) and elasticity of output with respect to capital (α_0) and labor (β_0) of 1996. Now that new investment and fresh labor will move into the regions for utility maximization, we can identify that at moment of t_j, and j ranges from 1996 to 2004, there would be:

$$\alpha_i(t_j) = \max\left[\alpha_1(t_j), \alpha_2(t_j), \ldots, \alpha_n(t_j)\right] = \alpha(t_j) \tag{117.16}$$

$$\beta_i(t_j) = \max\left[\beta_1(t_j), \beta_2(t_j), \ldots, \beta_n(t_j)\right] = \beta(t_j) \tag{117.17}$$

$$g_i = g_1 = g_2 = \ldots = g_n = g \tag{117.18}$$

And $n = 1, 2, \ldots, 53$, where $\alpha(t_j)$, $\beta(t_j)$ stands for the average elasticity of output with respect to capital and labor. Equations (117.16) and (117.17) mean that any incremental input of capital or labor will be allocated to the STIP where marginal output maximization can be achieved. Under "perfect competition" of STIPs in attracting investment and human capital, the only equilibrium state is that all STIPs have the same output elasticity of incremental input of capital and labor, depicted as $\alpha(t_j)$ and $\beta(t_j)$ separately.

Based on the analysis above, we suppose that industry clustering in STIPs as a whole can be distinguished by investigating the changing trends of $\alpha(t_j)$ and $\beta(t_j)$ as follows:

$$\frac{d\alpha(t)}{dt} + \frac{d\beta(t)}{dt} > 0 \tag{117.19}$$

If Eq. (117.19) is true, i.e., there is increasing output elasticity with respect to time series, we can prove that there is increasing returns from geographical agglomeration of production. A fundamental feature of industry clustering in STIPs as a whole thus can be verified.

To carry out regression, Eq. (117.5) can be re-written as:

$$\ln Y_i(t) = A_i^{'}(t) + \alpha_i(t) \ln K_i(t) + \beta_i(t) L_i(t) \tag{117.20}$$

Actually, $A_i^{'}(t)$ in Eq. (117.20) stands for time series of TFP (total factor productivity). Following Victor J. Elias [17], we use annual revenue and annual average fixed assets of STIP i to express output $Y_i(t)$ and capital input $K_i(t)$, which are calculated at constant prices of 1990 in 1 million Yuan. Moreover, labor input is expressed by employment in person. The resulting growth rates of variables

Table 117.3 Estimates of the output elasticity of capital and labor

Year	Coefficients			Sig.			Std. error			Model summary		
	$A'(t)$	$\alpha(t)$	$\beta(t)$	$A'(t)$	$\alpha(t)$	$\beta(t)$	$A'(t)$	$\alpha(t)$	$\beta(t)$	R^2	F	Sig. F
1997	−1.071	0.701	0.308	0.214	0.000	0.037	0.85	0.137	0.143	0.737	68.58	0.000
1998	−1.415	0.745	0.305	0.159	0.000	0.084	0.99	0.151	0.173	0.741	68.59	0.000
1999	−0.426	0.99	0.010	0.604	0.000	0.95	0.814	0.137	0.155	0.820	109.5	0.000
2000	−0.739	1.033	0.008	0.352	0.000	0.961	0.787	0.146	0.170	0.859	152.1	0.000
2001	0.051	0.989	−0.024	0.949	0.000	0.884	0.796	0.136	0.164	0.854	146.3	0.000
2002	0.141	0.952	0.007	0.873	0.000	0.969	0.878	0.141	0.173	0.831	122.8	0.000
2003	0.055	0.927	0.044	0.954	0.000	0.808	0.96	0.144	0.179	0.810	106.3	0.000
2004	−0.841	0.864	0.184	0.322	0.000	0.191	0.841	0.121	0.139	0.826	118.6	0.000

from 1997 to 2004, along with the estimated share parameters are reported in Table 117.3.

And we can get the following equations:

$$\alpha + \beta = 1.0226 - 0.0038t,\ R^2 = 0.0608 \tag{117.21}$$

$$\alpha = 0.7987 + 0.0225t,\ R^2 = 0.209 \tag{117.22}$$

$$\beta = 0.2239 - 0.0264t,\ R^2 = 0.2148 \tag{117.23}$$

$$A^{'} = -1.1441 + 0.1363t,\ R^2 = 0.3311 \tag{117.24}$$

These data indicate that there is no obvious increasing return to scale of production in STIPs as a whole. The average output elasticity is quite close to unit elasticity ($\overline{\alpha + \beta} = 1.005, R^2 = 0.00144$). A striking feature of these data is the high output elasticity of capital input and its increasing trend (rising from 0.701 to 0.864 in 2004), quite different with that of China in general. This enormous contrast well explained the geographical concentration of capital in STIPs. To get higher returns on investment, it is preferable for entrepreneurs to move into STIPs. The declining elasticity of labor input suggests the dilemma in Chinese economic growth: Inadequate labor demand due to management and technology advancement. For most developing countries, it is often marked by substantial unemployment, underemployment, and poverty, namely stalled "structural transformation" out of low-paying activities to higher value-added ones and the mounting pressure of employment [22].

117.4 Conclusion

Various types of zones in China have been served as experiment sites for transition from planned to market economy, attracting FDI and technology acquirement. In many Chinese cities, the establishment of hi-tech zones represents China's

attempt to overcome obstacles of the technology transfer form multinationals and to draw on both technology transfer and indigenous technology development in a more effective way.

Based on CD production function with external science and technology progress efficiency, the study investigated geographical concentration of production factors and increasing return to scale in High-tech Zones in China. Result shows that there was industry clustering in the zones.

However, the quality of industry clusters differs greatly among the zones, and many hi-tech zones are not successive in attracting real qualified hi-tech firms. And the analysis of the panel data indicates that the development of some state-level science and technology industrial parks in China is enterprises gathering rather than industry clustering. Moreover, much attention should be paid to the zones with poor agglomeration quality.

References

1. Walcott, S.M.: Chinese industrial and science parks: bridging the gap. Prof. Geogr. **54**(3), 349–364 (2002)
2. Wang, M.Y., Meng, X.: Building nests to attract birds China's high-tech zones and their impacts on transition from low-skill to high value added process. In: Proceedings of the 15th Annual Conference of the Association for Chinese Economics Studies Australia. RMIT Publishing (2003)
3. Furukawa, Y.: Industrial cluster study report. Industrial Cluster Study Group (2005)
4. Scott, A.J.: Regional push: the geography of development and growth in low- and middle-income countries. Third World Q. **23**, 137–161 (2002)
5. Fan, C.C., Scott, A.J.: Industrial agglomeration and development: a survey of spatial economic issue in East Asia and a statistical analysis of Chinese regions. Econ. Geogr. **79**(3), 295–319 (2003)
6. Liang, Q., Zhan, Y.-J.: Regional specialization, technology progress and industrials upgrading: proved by Yangtze River Delta. Econ. Theory Bus. Manag. **1**, 56–62 (2006)
7. Wang, J.-C.: Solving the puzzlement in the notion of cluster-talking about the problem of regional cluster development in our country. Jingji Jingwei Econ. Surv. **2**, 65–68 (2006)
8. Xia, H.-J.: The road to development of China's new & high-tech industrial park. China Citic Press, Beijing (2001)
9. Han, B.-T., Li, Q., Zhu, M.-G., Zhang, C.-B.: An evaluation of the inequality development of high-tech development zones in China based on the theory of entropy. Stud. Sci. Sci. **23**(3), 342–344 (2005)
10. Han, B.-T., Li, Q., Zhang, C.-B., Zhu, M.-G.: Evaluation of high-tech development zones in China based on the method of entropy weight. Chin. J. Manag. Sci. **13**(3), 144–148 (2005)
11. Lai, H.-C., Shyu, J.Z.: A comparison of innovateion capacity at science parks across the Taiwan Strait: the case of Zhangjiang High-Tech Park and Hsinchu Science-based Industrial Park. Technovation **25**, 805–813 (2005)
12. Macdonald, S., Deng, Y.-F.: Science parks in China: a cautionary exploration. Int. J. Technol. Intell. Plann. **1**(1), 1–14 (2004)
13. Cortright, J.: Making sense of clusters: regional competitiveness and economic development. Impresa, Inc, A discussion paper prepared for the Brookings Institution Metropolitan Policy Program (2006)
14. Henderson, J.V.: Marshall's scale economies. J. Urban Econ. **53**, 1–28 (2003)

15. Lucas, R.E.: On the mechanics of economic development. J. Monetary Econ. **22**, 3–42 (1988)
16. Kremer, M.: O-ring theory of economic development. Q. J. Econ. **108**, 551–575 (1993)
17. Elias, V.J.: Sources of growth: a study of the seven Latin American economies. ICS Press, San Francisco (1992)
18. Easterly, W., Levine, R.: It's not factor accumulation: stylized facts and growth models. World Bank Econ. Rev. **15**(15), 177–219 (2001)
19. Fajnzylber, P., Lederman, D.: Economic reforms and total factor productivity growth in Latin America and the Caribbean (1950–1995): an empirical note. World Bank policy research working papers 2114 (1999)
20. Hulten, C.: Total factor productivity: a short biography. In: Dean, E., Harper, M., Hulten, C. (eds.) New Directions in Productivity Analysis. University of Chicago Press (2001)
21. Akinlo, A.E.: Macroeconomic factors and total factor productivity in sub-Saharan African countries. Int. Res. J. Financ. Econ. **1**, 62–79 (2006)
22. Qi, J.-G.: A study of the relations between the total amount employment and technical change. J. Quant. Tech. Econ. **12**, 24–29 (2002)

Chapter 118
Evaluation of Person-Job Fit on Knowledge Workers Integrating AHP and BP Neural Network

Qing Wang and Guo Chen

Abstract In order to evaluate person-job fit on knowledge workers effectively, an evaluation method integrating analytic hierarchy process and BP neural network are proposed in this paper. The method is to get the evaluation results by calculating weighted sums of the evaluation results based on analytic hierarchy process and the evaluation results based on the trained BP neural network. A case is applied to demonstrate the method by MATLAB. The results show that the method can effectively evaluate person-job fit on knowledge workers. The method has the advantages of both analytic hierarchy process and BP neural network.

Keywords: Knowledge worker · Person-job fit · Evaluation model · Analytic hierarchy process · BP neural network

118.1 Introduction

In the knowledge-based economy, knowledge workers have become the most creative production factors of enterprises. The degree of person-job fit on knowledge workers influences the performance of enterprises. Person-job fit is a dynamic process: with the changes of employees' ability and the adjustments of jobs requirement, managers must re-allocate employees by promotion, demotion, rotation et al. As knowledge workers engaging in high-risk mental work improve their ability rapidly, it is difficult to evaluate their person-job fit. Therefore,

Q. Wang (✉) · G. Chen
School of Management, Tianjin University of Commerce, Tianjin, China
e-mail: wangqingtjcu@163.com

G. Chen
e-mail: amigo-fish@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_118,

developing a scientific method to evaluate person-job fit on knowledge workers has great significance.

To improve the reliability and validity of person-job fit evaluation, many scholars turned their attention to artificial intelligence [1–4]. These methods ignoring the subjective feature of person-job fit evaluation are all based on objective data.

Among the decision methods, analytic hierarchy process (AHP) can combine quantitative analysis with qualitative analysis. BP neural network is used to solve the problems whose rule implied in lots of disordered data. As person-job fit is affected by many implied factors, this paper introduces AHP and BP neural network into the evaluation of person-job fit on knowledge workers. Section 118.2 reviews the concept of AHP and BP neural network. Section 118.3 introduces the processes of the evaluation method of person-job fit on knowledge workers integrating AHP and BP neural network. A case is applied to demonstrate the method by MATLAB in Sect. 118.4. Finally, Sect. 118.5 provides the conclusions from this work.

118.2 AHP and BP Neural Network

The hierarchy structure of AHP is shown in Fig. 118.1 (Here is given person-job fit on R&D employees as an example). It includes target layer, criterion layer, and approach layer. By calculating the maximum eigenvalues and corresponding eigenvectors of the judgment matrices, which are made by paired comparing the importance of the elements in each layer for each element in the upper layer, we can finally obtain the weights of the elements in each layer.

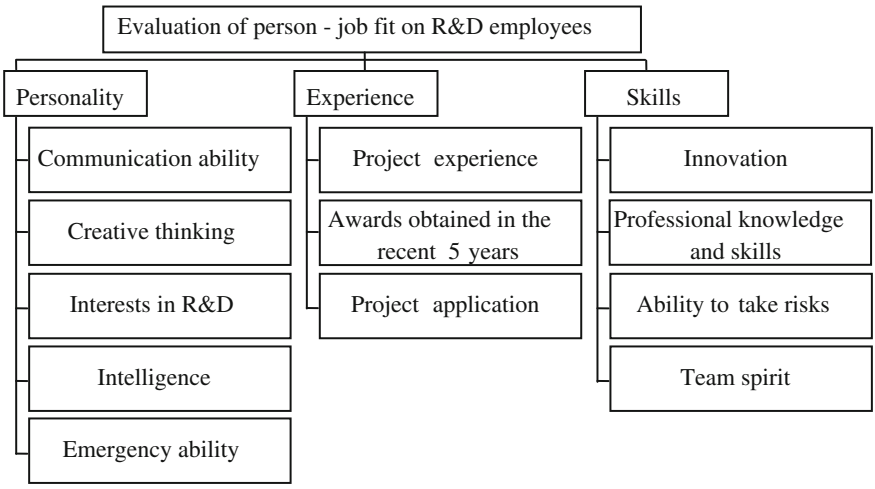


Fig. 118.1 Hierarchy structure of person-job fit evaluation on R&D employees

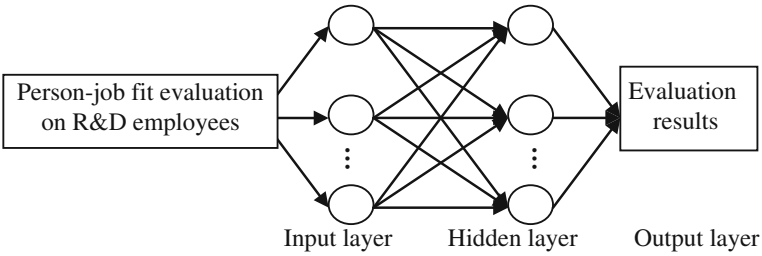


Fig. 118.2 An evaluation model of person-job fit evaluation on R&D employees based on BP neural network

Figure 118.2 shows the topological structure of BP neural network. The structure contains input layer, hidden layer, and output layer (Here is given person-job fit on R&D employees as an example). The network is to find out the nonlinear relationship among data by learning, which consists of information dissemination and error back propagation [5].

For AHP is too subjective and BP neural network is too objective, this paper provides an evaluation method of person-job fit on knowledge workers based on AHP and BP neural network, which is mainly depend on BP neural network with AHP. The final evaluation results are weighted sums of the evaluation results of AHP and the evaluation results of BP neural network.

118.3 Evaluation Method of Person-Job Fit Integrating AHP and BP Neural Network

The processes of person-job fit evaluation on knowledge workers integrating AHP and BP neural network are: (a) establish the original evaluation index system; (b) select key indexes and get the evaluation results by AHP; (c) use the samples of person-job fit on knowledge workers to train BP neural network and get the evaluation results; (d) calculate weighted sums of the evaluation results of AHP and the evaluation results of BP neural network. The specific processes for using this method are described below.

118.3.1 Evaluation of Person-Job Fit by AHP

118.3.1.1 Establish Original Evaluation Index System

The original evaluation index system of person-job fit on knowledge workers in a high-tech enterprise is established by managers and experts.

118.3.1.2 Select Key Evaluation Indexes

Establish the hierarchy structure and judgment matrices of the problem. Eigenvectors of the maximum eigenvalues for each judgment matrix are weights of the indexes. Those indexes with more weight are selected to be key indexes.

118.3.1.3 Get Evaluation Results

Evaluation result of each sample is the sum of each evaluation index score multiplied by its weight.

118.3.2 Evaluation of Person-Job Fit by BP Neural Network

118.3.2.1 Determine the Input/Output Layer Nodes

The input layer nodes are the key evaluation indexes selected by AHP. And the output layer nodes are the evaluation results.

118.3.2.2 Determine the Number of the Hidden Layer and Its Nodes

Kolmogrov theory has shown that a three-layer neural network can realize any given continuous function [6]. Thus, this paper designs a three-layer BP neural network. There are three formulas can help us decide the number of the hidden layer nodes [7]:

(a) $\sum_{i=0}^{n} C_m^i > k$, k is the number of samples, n is the number of the input layer nodes, i is a constant in [0, n], ni is the number of the hidden layer nodes.
(b) $ni = \sqrt{n+m} + a$, ni is the number of the hidden layer nodes, n is the number of the input layer nodes, m is the number of the output layer nodes, a is a constant in [1, 10].
(c) $ni = \log_2 n$, ni is the number of the hidden layer nodes, n is the number of the input layer nodes.

118.3.2.3 Determine the Output Function

The output function of each neuron is a sigmoid function:

$$I = \sum_{j=0}^{n} w_j x_j \tag{118.1}$$

$$y = f(I) = \frac{1}{1 + \exp(-I)} \tag{118.2}$$

x_j is input signal, w_j is the strength of the joint among neurons, n is the number of the input signals, j is a constant in [0, n]. I is the summing unit, which is the first processing step of the neuron. $f(I)$ is the output function, which is the second processing step of the neuron.

118.3.2.4 Pre-Process the Samples Data

Normalize the samples data and divide them into two parts, one part is the training set used to train the BP neural network, the other is the testing set.

118.3.2.5 Train and Test

After being trained, BP neural network needs to be tested to ensure whether it has got a steady network structure and weights. And the testing set should be entered into the trained BP neural network for testing.

118.3.3 Evaluation of Person-Job Fit Integrating AHP and BP Neural Network

Get the results by calculating weighted sums of the evaluation results of AHP and the evaluation results of BP neural network. And the weights are obtained by expert discussion.

118.4 Case Study

Company A is an IT enterprise. Company A has been enlarging itself these years and the problems in its person-job fit on knowledge workers have become obvious. To provide decision support for Company A, we randomly select 150 R&D knowledge workers from this company as samples and evaluate their person-job fit by the given method integrating AHP and BP neural network.

Table 118.1 Weights of person-job fit evaluation indexes on knowledge workers based on AHP

Level index	Second index	C.I.	R.I.	Weight of index			Final weight of index
				Personality	Experience	Skills	
				0.2604	0.3352	0.4045	
Personality	Communication ability	0.0170	1.1200	0.1599			0.0416
	Creative thinking			0.4185			0.1090
	Interests in R&D			0.2625			0.0684
	Intelligence			0.0973			0.0253
	Emergency ability			0.0618			0.0161
Experience	Project experience	0.0091	0.5800		0.5584		0.1872
	Awards obtained in the recent 5 years				0.3196		0.1071
	Project application				0.1220		0.0409
Skills	Innovation	0.0170	0.9000			0.2844	0.1150
	Professional knowledge and skills					0.4729	0.1913
	Ability to take risks					0.1699	0.0687
	Team spirit					0.0729	0.0295

118.4.1 Select the Evaluation Indexes by AHP

We select 12 important indexes to evaluate person-job fit on knowledge workers in Company A. And then the hierarchy structure which shown in Fig. 118.1. Judgment matrices are established. By using MATLAB, eigenvectors corresponding to the maximum eigenvalues of each judgment matrix are calculated to be the weights of the indexes (as shown in Table 118.1). The eight key evaluation indexes are professional knowledge and skills, project experience, innovation, creative thinking, awards obtained in the recent 5 years, ability to take risks, interests in R&D, and communication ability.

118.4.2 Evaluate Person-Job Fit by AHP

According to the 8 key indexes and their weights, samples can be evaluated by AHP in the method described below. Divide the eight key indexes into six grades: extremely match, very match, match, mismatch, very mismatch, and extremely mismatch. The scores of the six grades in turn are 1, 0.8, 0.6, 0.4, 0.2, and 0. Then 50 samples are randomly selected from the 150 samples as the testing set (The other 100 samples are the training set). The sum of each index score multiplied by its weight given by AHP is the evaluation result of the sample in the testing set.

118.4.3 Determine the Input, Hidden, Output Layer Nodes and the Output Function of BP Neural Network

The above eight key indexes are selected as the input layer nodes. And each node represents a key index.

The number of the output layer nodes is 1 for there is only one evaluation result for each sample. Divide the evaluation result into five levels: A, B, C, D, and E. The scores ranges of these five levels in turn are [0.8, 1], [0.6, 0.8], [0.4, 0.6], [0.2, 0.4], and [0, 0.2].

According to the Kolmogrov theory [6], we set the number of the hidden layer as 1. And the number of the hidden layer nodes is designed to be 8 to ensure the network stable.

The output function of this BP neural network is Eq. (118.2).

118.4.4 Train and Test the BP Neural Network

Take the evaluation data of the training set as learning data to train the BP neural network by MATLAB. After several test runs, we find that when the training times of the network is 50, the error is acceptable. Thus, we set the maximum training times of the network as 50. The main computer language program code is:

```
net = newff([0 1;0 1;0 1;0 1;0 1;0 1;0 1;0 1],[8 1],{'tansig' 'purelin'});
net.trainParam.epochs = 50
```

The evaluation results of the testing set can be obtained by entering the samples data into the trained BP neural network. The evaluation results on 10 partial samples in the testing set are shown in Table 118.2.

Among the 10 testing samples in Table 118.2, there are only two samples whose actual outputs are different from the target outputs. Among the all 50 testing samples, there are only 9 samples whose actual outputs are different from the target outputs. The accuracy of the designed model is about 80 % and the model can evaluate person-job fit on knowledge workers well.

118.4.5 Evaluate Person-Job Fit by the Given Method Integrating AHP and BP Neural Network

According to experts, the evaluation results based on AHP multiplied by 40 % plus the evaluation results based on BP neural network multiplied by 60 % would be more scientific for the evaluation. These new results of the 10 samples are shown in Table 118.3.

By comparing the evaluation results of 50 samples based on AHP, BP neural network, and AHP & BP neural network with the target outputs, we can find out

Table 118.2 Evaluation results of 10 partial samples based on BP neural network

Input index	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8	Sample9	Sample10
Creative thinking	1.0	0.4	0.6	0.6	0.8	0.4	0.6	0.6	0.6	0.1
Interests in R&D	0.8	0.8	0.6	0.8	0.8	0.6	0.8	0.4	0.8	0.2
Innovation	0.8	0.6	0.6	0.6	0.8	0.4	0.6	0.6	0.4	0.0
Project experience	0.6	0.6	0.4	0.4	0.6	0.2	0.2	0.6	0.2	0.1
Communication ability	0.6	0.6	0.4	0.6	1.0	0.6	0.6	1.0	0.4	0.0
Professional knowledge and skills	0.6	0.6	0.6	0.6	0.6	0.8	0.6	0.6	0.6	0.0
Awards obtained in the recent 5 years	0.6	0.6	0.4	0.2	0.2	0.2	0.4	0.2	0.2	0.2
Ability to take risks	0.8	0.6	0.8	0.8	0.8	0.8	0.8	0.6	0.8	0.2
Target outputs	0.6	0.6	0.4	0.4	0.5	0.5	0.5	0.5	0.3	0.3
Actual outputs	0.9	0.6	0.4	0.4	0.5	0.5	0.5	0.5	0.3	0.1
Level	A	B	C	C	C	C	C	C	D	E

Table 118.3 Evaluation results of 10 partial samples based on AHP, BP neural network, and AHP & BP neural network

Method	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
AHP	0.63	0.52	0.48	0.48	0.58	0.42	0.46	0.45	0.41	0.29
BP	0.90	0.60	0.40	0.40	0.50	0.50	0.50	0.50	0.30	0.10
AHP & BP	0.79	0.87	0.49	0.39	0.48	0.49	0.44	0.46	0.33	0.15
Target outputs	0.60	0.60	0.40	0.40	0.50	0.50	0.50	0.50	0.30	0.30

that: (a) the evaluation results based on AHP & BP neural network are closer to the target outputs than the other two kinds of evaluation results; (b) the evaluation results based on AHP & BP neural network are between the other two kinds of evaluation results. The method has the advantages of both analytic hierarchy process and BP neural network.

118.5 Conclusion

This paper gives a new method combining AHP with BP neural network to effectively evaluate person-job fit on knowledge workers. The results of the case show that: (a) the evaluation accuracy of the designed BP neural network is about 80 %; (b) the evaluation results based on AHP&BP neural network are between the results based on AHP and the results based on BP neural network, and the former are closer to the target outputs. The method has the advantages of both analytic hierarchy process and BP neural network. There are still some deficiencies in this method: (a) the science and the convergence speed control of the BP neural network need to be improved; (b) Based on AHP and BP neural network, which are only relied on expert experience, the respective weights of the evaluation results need to be improved.

References

1. Li, J., Gao, B.: Study of person-job fit based on grey system theory. Chin. Bus. **3**, 103–104 (2008)
2. Güngör, Z., Serhadlıoğlu, G., Kesen, S.E.: A fuzzy AHP approach to personnel selection problem. Appl. Soft Comput. **9**, 641–646 (2009)
3. Şen, C.G., Çınar, G.: Evaluation and pre-allocation of operators with multiple skills: a combined fuzzy AHP and max-min approach. Expert Syst. Appl. **37**, 2043–2053 (2010)
4. Zhang, Z., Lv, M., Li, C.: Study on the evaluation model of person-job fit based on BP nerural network. Trans. Tianjin Univ. **12**, 390–395 (2010)
5. Ge, Z., Sun, Z.: Neural Network Principle and its Realization by MATLABR2007. Electronic Industry Press, Beijing (2008)
6. Jiao, L.: Theory of Neural Network System. Xian Electronic Science & Technology University Press, Xian (1995)
7. Sheng, H., Wang, Z., Gao, C., et al.: Determine the number of the hidden layer node of BP neural network. J. Tianjin Univ. Sci. Technol. **24**, 13–14 (2008)

Chapter 119
Discussion of ITM-Based Human Resources Planning and Management: An Example of W Corporation

Hong-ming Chen and Ya-nan Kang

Abstract The paper aims to improve the chaotic situation of human resource management within Chinese enterprises, especially some family business and state-owned enterprises, and to optimize their internal human resource structures to maximize the staff value. The paper introduces the ITM human resource management theory, for an integrated and systematic human resource management research of Chinese enterprises. Based on the ITM human resources information needs analysis, the writer specifically introduces the concrete application of ITM in Chinese enterprises taking example of W company, and reveals that ITM will improve the structure of human resources and improve the resource efficiency to achieve the maximized human capital value of the whole enterprises.

Keywords ITM · Human resource management · Planning and management · Case study

119.1 Foreword

In order to adapt to the changing work environment, the practices and technologies which related to talent management are continuous developing, improving, adjusting and adapting over time. While the most effective measure is concrete analysis of concrete problems, that is, make individual managing measure and corresponding adjustment to different enterprise.

H. Chen (✉) · Y. Kang
School of Economic and Management, Changsha University of Science and Technology, Changsha, China
e-mail: chmdsh@163.com

Y. Kang
e-mail: hnkyn@126.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_119,

The Integrated talent management (ITM), which means integrated human resource management, is a kind of integrated human resources management theory that newly sprung in Chinese enterprises in the past two years. Namely, the ITM theory aims to success an integrated management from staff selection, recruitment, reward and retaining top talent by adjusting or improving the processes and technologies, which enable employees to adapt to the strategic objectives and business value of their corporate. And then it will enhance personal power as well as organizational through the integrated talent management system.

119.2 The Information Needs Analysis of ITM

The study entitled "Integrated Talent Management: Improving Business Results through Visibility and Alignment"(which was conducted by the Aberdeen Group, a Harte-Hanks Company (NYSE: HHS) [1], with support from Talent-Scope, Inc., the only fully integrated provider of talent management solutions.) shows that, to the context of shaky job market and divergent views of the huge turnover predict, once the economics recovered, compared to those with traditional talent management processes and individual employee data systems, the enterprises with ITM-based personnel management processes will not only achieved higher profitability, but also better to adjusting themselves to the position to achieve long-term interests and development [2]. The study confirmed:

Firstly, ITM enable higher profit and other benefits: compared to the leading enterprises without ITM model, those who implemented ITM-based talent management process have higher annual profit by 7 % and their employees also increased 19 %. Moreover, almost more than 50 % of the positions in these enterprises have identified successors—that is a critical factor to effectively manage attrition that due to resignation and retirement.

Secondly, there are high correlation between the integration and the standardization of talent management process: in fact, in the companies with excellent performance (including staff performance, manager satisfaction and employee engagement), 85 % had integrated at least part of the talent management processes and workflow; while 63 % of the bad performance enterprises with no comprehensive talent management process at all.

Thirdly, ITM views that explore potential staff value is important: what successful companies need to do is much more than just managing personnel, but the more important thing is try to train and explore staff potential value. Especially in the current economic situation, only be correctly identified and retain senior talents can survive the enterprises. Thus the most urgent thing for the businesses is mastering the real human resources situation and making effective talent management planning.

In addition, combination interests: combined the traditional human resources with other functional "talent" (such as training, development, compensation, etc.)

can not only helps reduce workflow bottlenecks, but also has a supporting role on measuring the key indicators about talent, such as the recruitment quality and the strength of key personnel.

In the study, 76 % of senior managers declared that they will increase investment in human resources management. Obviously, the ITM will commercially benefit the organization and has a high degree of market information needs [3].

119.3 Take M's to Analysis the ITM-based Human Resources Information Planning

119.3.1 W Company Profile

Founded in 1996, W is a private limited company engaged in production and sales of specialty products in Henan Province whose predecessor is a small hand workshop, and developed into a minor celebrity food processing and raw material suppliers in 5 years. It mainly engaged in processing and trading grinding sesame oil, black fungus, mushrooms, mushrooms and other edible fungus in distribution and wholesale mode, main products are edible fungi products and authentic grinding sesame oil Featured of Henan, with exports of 50–100 million and annual turnover of 300–500 million.

119.3.2 The Current Human Resource Management Situation of W

There are procurement department (including warehouse management), production department, sales department, finance department, the office (contained human resources management) in W company at present, in total of more than 200 employees. On the impact of family hand Workshop predecessor, the company is still in the family business management models and concepts with big problems in W's HR management.

Sources of indicators: To analysis W's current HR management issues, the writer made some questionnaire surveys towards more than 200 employees and got 120 valid questionnaires with the help of the W's talent manager and some employees after sorting and analysis we get the following results.

119.3.2.1 Personnel Recruitment

The company didn't make any relevant recruitment plan, even without special human resource department but part-worked by the office personnel. In W's, the

Table 119.1 Satisfactory or not to the company's recruitment methods

	Strongly	Satisfactory	Not much	Unsatisfactory
Number	2	31	49	38
Proportion (%)	1.7	25.8	40.8	31.7

recruitment is mainly by staff recommendation, relatives and local labor intermediary introduction, and the interview are mainly charged by the chairman of the board who is also the general manager. No paper test, no personality test, people can be employed as long as the chairman agreed. Despite that the staff do not agree (as Table 119.1), but no one put forward objection, most of the employees regard that "whose enterprise, listen to whom".

119.3.2.2 Staff Execution

The most critical factor affecting employee execution is whether one's ability matched its post or not. Only the employee competence matched its post requirement can it effectively play staff execution, which is also helpful to employees' satisfaction and their improved performance. But the questionnaire to company W is not ideal. See Table 119.2.

Simultaneously, playing staff executive closely related to the intensity and effectiveness of staff training [4]. On respect of training, although W had some degree of training measures, however, due to lack of specialized personnel management department, most of them become a mere formality when down to implementation as the company's requirements.

119.3.2.3 Development

After sorting and classification those questionnaires, the paper started from three aspects—training, career development satisfaction and selecting system to reflect W's development. As Tables 119.3, 119.4 and 119.5 show, what the W's did in respects of staff training and Incubation Management are still unsatisfactory, which cannot make the staff generally feel good and be full of enthusiasm for work.

Table 119.2 How much the company personnel matched the post

	Number	Proportion (%)
Can fully display personal talents	11	9.2
Can display on Some degree	31	25.8
Can basically play	54	45
Cannot play	24	20

Table 119.3 Satisfaction of the company training effect

	Number	Proportion (%)
Adapt to the job requirement, be very useful	13	10.8
Have certain effect, be useful	30	25
Little help to the work	33	27.5
Mere formality	44	36.7

Table 119.4 Staff career development satisfaction

	Number	Proportion (%)
The own work be of promising	33	27.5
Be possible of promising	39	32.5
Unclear, have no idea	29	24.2
No promising	19	15.8

Table 119.5 Selection system satisfaction

	Number	Proportion (%)
Fair, the capable be on	25	20.8
Relatively fair	31	25.8
Basic fair	45	37.5
Unfair	19	15.8

119.3.2.4 Whether to Retain Key Talent

The deletion in aspects of incentives and positions promotion (Tables 119.4 and 119.5) makes a sense of accomplishment and loss of future about personal development to part of the staff, which leading to staff turnover. In Tables 119.6 and 119.7, that is reflected from the relationship between work performances and the rewords or positions promotion, and the resigning intention (see also Table 119.8).

119.3.3 ITM-Based Human Resource Management Information Planning for W

119.3.3.1 Broken Down the Organizational Structure by Function

The ITM theory pays attention to the influence of each department to the whole human resources' performance. While the organizational structure division in W is too rough, especially the fuzzy partition of the Office and HR Department, so the company should add marketing department, logistics department and HR Department according to developing needs.

Table 119.6 Do your salary and incentive have relationship with your work performance?

	Great relationship	Certain relationship	Little relationship	No relationship
Number	15	31	41	33
Proportion (%)	12.5	25.8	34.2	27.5

Table 119.7 Does your job promotion have relationship with your work performance?

	Great relationship	Certain relationship	Little relationship	No relationship
Number	15	36	40	29
Proportion (%)	12.5	30	33.3	24.2

Table 119.8 Resigning Intentions survey

	Would not departure	Up to the opportunity	To consider	Sure to leave
Number	30	37	34	19
Proportion (%)	25	31	28	16

After functional subdivision, the organizational structure will be more concentrate, much easier to extend and be better for responsibility division and performance appraisal [5]. In view of company W is expanding import and export to explore foreign markets currently, the marketing department can recruit some experienced market analysts to open the foreign markets stably; and the HR Department should carry out specific plans to form a comprehensive talent management processes containing recruitment, training, development and retention key talent.

119.3.3.2 Innovative Recruitment Processes and Methods Based on ITM

In view of the status of nepotism in W, the personalized recruitment site with the ITM platform can be helpful. The recruitment site can refuse those "family relationship" effectively when screening resume through a "threshold". Based on resume integration from various channels, the ITM system will screen candidates fast and accurately, match and recommend automatically according to the talent evaluation results, and transfer information more efficiently among the interviewer and the candidate which may achieve a smooth recruitment process.

119.3.3.3 Innovative Development Design

Effective implementation of the human resources development plays a direct promoting role to improve organizational effectiveness and profitability, which also picked up about 15.4 % of the total shareholder return. However, W's seriously neglected its human resources development.

Based on the ITM system, W's can screening personnel rapidly according to post matching analysis, recognize people's quality, characteristic or potential exactly, and discover the individual strengths and weaknesses. Then "Prescribe the right medicine"-design and implement effective strategy for talent training, which may form a talent pool and enable the manager to control talent reserve, plan talent development, make continuous tracking evaluation, and identify suitable candidates finally [6].

119.3.3.4 Enhance the Staff Execution

As is well-known, looking for experienced, skilled employees is not an easy thing, which makes the importance of staff execution even more prominent. The most safe execution management mode is supposing that whether for its current role or preparation to the future role, all employees own space to growth and development in the organization.

The ITM-based human resource management can achieve effective connection of each the functional departments by design and implementation. The integration and planning of each module enable the competency-based conclusion of other steps can be invoked in any link so that policymakers can make more clear and more persuasive decisions.

119.3.3.5 Performance Assessment

Firstly, design the ITM overall index system that close to the actual can complete performance appraisal process in the shortest, resolve strategic target layers after layers, grasp core operations, interact and communicate in real-time and adjust timely as the response and feedback [6, 7].

Secondly, designing associated performance evaluation system be able to review each other, of which each method should be comprehensive systematic.

Meanwhile, combined with perfect designed ITM website system, the W's may achieve evaluate the human resources management or even management of all the work in each section in 360 °. And then automatically implement evaluation, questionnaire design and statistical analysis with safety answer mechanism and reliable results, and conclude professional feedback report finally.

119.4 Conclusion

From the case analysis of W's, we can get that benefited from its strategic human resource planning, complete HR processes and effective loss management, ITM strives to build an integrated talent management ambiance with inter- convergence function, mutual restraint system and links affecting performance appraisal.

Apparently, for many Chinese family enterprises like W's and part of the under-system state-owned enterprises, the ITM management idea brings them methods and energy to dredge talent management. Thus, the ITM theory has great potential and corporate recognition in Chinese market, whether for the companies with inefficient talent management or enterprises with considerable scale human resources. The ITM will bring them a talent management revolution.

Every management has its profound sociality. Therefore, managements are obliged to adapt to their native culture and tradition. Thus in the process of talent management, companies not only need to learn from others advanced experience, but also to adhere to its national conditions, to constantly conclude experience in practice. Then they can make innovation on the basis of reference and absorption to realize humanistic and scientific talent management gradually.

References

1. Martin, K.: Integrated Talent Management: Improving Business Results Through Visibility and Alignment. Aberdeen Group, Benchmark (2009)
2. Yang, S., Nan, Li.: Review and Prospect of International High-Performance Human Resource Practices. Management Review. **23**(10), 83-90 (2011)
3. http://cn.reuters.com. New research confirms business benefits of integrated talent management
4. Shuming, Z., Suying, G., Chunjie, G.: The relationship between strategic international human resource management and firm performance: based on the empirical date of MNCs in China. Nankai Business Review. **14**(1), 28-35 (2011)
5. Xindi, W., Qingjie, Z.:Developing the Framework of Integrated Performance Management System for State-Owned Enterprises in China. IEEE International Conference on e-Business Engineering, pp. 456-460 (2006)
6. Stamm, A.: The Strategic Development of Core HR Systems: Helping Leaders Go Beyond Administrivia and Compliance. Aberdeen Group (2007)
7. Ready., D., Conger, J.:How to fill the talent gap. Wall Street Journal. September 15-16. R4. (2007)

Chapter 120
Is Inequality or Deficiency the Real Trouble?—The Influencing Factors of the Income Satisfaction of Chinese Sci-tech Personnel

Yi Yang

Abstract With an increasing gap among current incomes, the Chinese sci-tech personnel's income satisfaction does not just depend on the income itself. By controlling some demographic variants such as gender, age, work seniority, etc., this study conducted a thorough analysis of the influence of some factors on the income satisfaction via optimal scaling regression and some other regression methods. The influencing factors include the income itself, the income gap, and the allocation system and so on. It was found that the inter-industry income gap, to which the income satisfaction is majorly attributed, can more easily cause discontent among the sci-tech personnel than the income itself. The allocation system can exert stronger influence on the income satisfaction among the sci-tech personnel than the intra-regional income gap, yet not as strong as that caused by inter-industry income disparity.

Keywords Income · Satisfaction · Income gap

120.1 Introduction

Confucius says that it is the inequality rather than deficiency is the cause of trouble and it is the instability rather than poverty is the root of disaster. This archaism implies that compared to low incomes, the disparity among people's gaining is more liable to give rise to discontent, wrath and some other negative effects among people. When it comes to social stability, large income gap and uneven allocation

Y. Yang (✉)
Nanjing University of Information Science and Technology, Nanjing, Jiangsu, China
e-mail: Chinayy1999@126.com

Y. Yang
Southeast University, Nanjing, Jiangsu, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_120,

rather than low income are what should be worried about. This old wisdom saying, in plain words but with profound meaning, is worth of pondering particularly when the current China is undergoing a daily increasing of income gap. The study in this paper is to thoroughly discuss the income satisfaction from three major aspects. First, it will compare the income itself and the income disparity to see which factor can more easily cause discontent. Secondly, in terms of income disparity, it will explore which factor can exert stronger influence on the income satisfaction, the inter-industry income disparity or the intra-regional income disparity? Thirdly, it will try to make clear which one can affect the income satisfaction more between the income disparity and the allocation system.

120.2 Literature Review

Income satisfaction is the link connecting income and individual behavior, which is triggered by income and some relevant factors and directly influences individual working behaviors and working attitude. Since 1960s, foreign researchers have conducted systematic study of income satisfaction, among which the study of the influencing factors over income satisfaction is the research hot spot. The research conducted by American researcher Adams discovers that the absolute value of the payment of the members in an organization does not have direct or consequent relations with the members' working initiative, yet the fairness of allocation can produce huge impact on employees' job satisfaction and working initiative [1]. Lawler also finds that if the employee believes there is gap between the actual gaining and his or her expected payment, the income satisfaction can be impaired, thereby influencing their working behaviors and working attitude [2]. Heneman and Schwab conduct more detailed research on income satisfaction and point out that income level, income increase, welfare, income structure, income management and some other influencing factors can all affect people's satisfaction towards income [3].

Meanwhile, domestic related researches are not rare to be found and have made quite achievement in this field. Ye Qin, Dai Dashuang and Wang Haibo analyzed the questionnaire results by fractional multiple regression method, which aims to investigating the income satisfaction of the employees in a domestic mobile communication operator, proving that there is significant correlation between environmental factors and income satisfaction. Also they analyzed the differentiation effect on income satisfaction exerted by environmental factors and the difference of the income satisfaction among employees at different ranks [4]. Zhang Junqin, Li Lianshui and Zang Zhipeng find that non-economic compensation can exert stronger influence upon income satisfaction than economic compensation and there is significant correlation between economic compensation and non-economic compensation after they established hypotheses and studied the relations between the two kinds of compensation and income satisfaction [5]. Chen Tao and Li Lianshui conducted a questionnaire survey investigating 12,000 sci-tech

personnel in 12 cities in Jiangsu Province, China, discovering that there is significant difference in incentive compensation satisfaction factoring in age, gender, position and nature of unit [6].

Despite that researchers home and abroad have done quite an exploration in income satisfaction and its influencing factors, there are three prominent problems in the research. First, most researches focus on the income and allocation system within one company or industry, ignoring the external comparison. China is undergoing a daily increasing of income gap among different industries; therefore, external comparison of income is with both Chinese characteristics and epic meaning [7–9]. Secondly, in terms of method, many researches adopt factor analysis to probe into the influencing factors of income satisfaction from its internal dimension. Yet factor analysis is easily affected by the observation parameters set by the researcher; moreover, it is difficult to explore the interaction effect among the variants via factor analysis. Thirdly, the researches tend to ignore the influence caused by common method biases which can produce severe interference to research results.

Therefore, the study presented in this paper will lay emphasis on the influence brought by income, income gap, etc. on the income satisfaction of the sci-tech personnel in China while strictly controlling the extraneous variants such as relevant demographic variants, common method biases, etc. [10].

120.3 Research Method

120.3.1 Subject

The study group chose the sci-tech personnel in a certain industry in China as the subject. With the generous support from the relevant departments in this industry, totally 18,880 original questionnaires were recovered via e-investigation. To ensure the accuracy and validity of the data, the study group screened the original questionnaires grounded on three points, the time spent on answering the questionnaire, the key questions in the questionnaire being answered, and the consistency of answers to similar questions. After the screening 15,165 valid questionnaires were left, among which males are 9,281 and females are 5,769, accounting for 61.7 and 38.3 % of the total samples respectively. This is proper in terms of gender structure. As far as age is concerned, there are 3,942 investigated employees under 30, occupying 26.0 % of the total samples; 4,419 from 31 to 40, taking up 29.2 % of the total samples; 4,491 from 41 to 49, accounting for 29.7 % of the total amount; 2,293 over 50, holding 15.1 % of the total samples. It generally fits the normal distribution. When it comes to title, those without titles account for 9.2 %, those with junior title 33.3 %, those with intermediate title 45.0 %, and those with senior title 12.5 %. Geographically, employees investigated in the eastern areas are 4,257, taking up 28.1 % of the total samples; those in the central regions 3,700, holding 24.4 % of the total samples; those from the

western areas 6,381, accounting for 42.1 % of the total samples. Generally speaking, the data gathered is reasonably distributed in gender, age, title, region, etc. and is close to the original samples in proportion.

120.3.2 Research Instruments

120.3.2.1 Questionnaire

Many researches employ measurement scale to measure the income satisfaction [10]. In order to accurately measure the influencing factors of income satisfaction and the interaction effect among them, the income satisfaction in this study refers to the overall income satisfaction of the employees, which is measured by the question "*Are you satisfied with your current payment*", the answers to which are scaled into five levels from "*Very dissatisfied*" to "*Very satisfied*" [11]. In terms of absolute income, two indexes are adopted. The first index is average monthly earnings. Based on the reality in the investigated industry, incomes in this industry are graded into five levels from those who earn <1600 RMB to those who earn more than 8001 RMB. The second index is income pressure. Owing to the significant disparity of income and price level among different regions in China, this study particularly introduced the concept of income pressure to measure the earnings of the investigated employees. Moreover, this study refined the income comparison, which includes internal comparison and external comparison. The former is measured by two questions, "*In your company/working unit, can you get well paid*" and "*Is the allocation system in your company/working unit fair to you*". The latter is also measured by two questions. The first question is "*What level do you think your income is locally at"*, which is to measure the local income gap. The second question is "*Compared to similar positions in government and public institutions, what level do you think your income is at"*, which is to measure the inter-industry income gap. All the questions are scored by a 5-Likert scale [12].

120.3.2.2 Statistical Analysis Method

This study adopts SPSS16.0 to process the data.

120.4 Research Results

120.4.1 Descriptive Statistics of the Income Satisfaction

In general, the income satisfaction in the investigated industry is relatively low, the average value of which is 2.26, close to the level of "*Not so satisfied*". The income

satisfaction is significantly different factoring in gender ($t = 3.746$, $P = 0.000$) that females are more satisfied with their incomes than males. There is also significant difference among different age groups ($F = 11.167$, $P = 0.000$) that the income satisfaction is growing with age. The income satisfaction of the sci-tech personnel in different regions is distinct as well. Those in the western regions have the lowest income satisfaction, close to "*Not so satisfied*", slightly lower than the income satisfaction of those from the central regions. The highest income satisfaction comes from the eastern regions [13].

120.4.2 Influencing Factor Analysis

To further detect the relevant factors that might exert influence on the income satisfaction of the sci-tech personnel and the influencing degree of these factors, SPSS is adopted to conduct Chi square test and optimal scale regression analysis. The results are presented as below.

120.4.2.1 Chi Square Test

Eight variants are involved in the Chi square test to explore their influence upon the income satisfaction of the sci-tech personnel and to see if the influence is statistically significant. See Table 120.1.

It can tell from Table 120.2 that difference in the income satisfaction of the sci-tech personnel is statistically significant factoring in region, gender, age, education, title, position, working rank and working hours. This indicates that all of these demographic variants can influence the income satisfaction of the sci-tech personnel.

120.4.2.2 Optimal Scale Regression Analysis

The above analysis focuses on the influence that sample characteristics exert on the income satisfaction of sci-tech personnel. Besides, there are many other factors which might affect the income satisfaction of sci-tech personnel. In this particular study, the influence exerted by income gap, allocation system and the income itself gets most attention. Because questionnaires were adopted in this study, the most obtained is ordinal categorical data, hence the result will be insufficient if linear regression is performed. Yet using logistic regression could cost the useful information in such kind of data. Therefore, this study decides to use optimal scale regression method. By using some non-linear transform methods, the four aspects, i.e., the sample characteristics, the absolute income gap, the psychological income gap and living burden, can be iterated thus to find out an optimal equation.

[1] The dependent variables in Equation 1, 2, 3, 4 and 5 are all income satisfaction.

Table 120.1 Chi square test

Variants	Region	Gender	Age	Education	Title	Position	Working rank	Working hours
Chi square value	930.527	32.287	39.787	24.578	47.410	242.83	230.253	56.264
P value	0.000	0.000	0.000	0.017	0.000	0.000	0.000	0.000

It can be told from the logistic regression analysis that gender, age, education, title, working sonority can influence the income satisfaction of the sci-tech personnel. Therefore, it is these five variants that are under control in the optimal scale regression analysis. Via stepwise regression, the influence of absolute income gap, psychological income gap and living burden on the income satisfaction of the sci-tech personnel is analyzed.

(1) Sample characteristics. In Equation 1, via analyzing and comparing the influence that the five variants (gender, age, education, title and work seniority) exert on the income satisfaction of the sci-tech personnel, it is found that the regulated R^2 is only 0.013. Therefore, the model established based on Equation 1 is not statistically meaningful. Given this Equation 1 will not entail much analysis.
(2) The income. The average individual monthly income is plug into Equation 2 after eliminating the interference caused by the sample characteristics such as gender, age, education, title, work seniority, etc. From Equation 2 it can be told that the income itself can affect the income satisfaction of the sci-tech personnel to a great extent with the coefficient of estimate reaching 0.36 and the coefficient of determination reaching 0.123.
(3) The income gap. Equation 3 takes into an extra consideration of the factors such as the relative income and allocation system to probe into their influence on the income gap among the sci-tech personnel. The results indicate that the industry income gap can exert the most influence with its coefficient of determination being 0.320, which means that the most primary norm for the sci-tech personnel to compare their incomes is the earnings of similar positions in similar industries. If the gap is large, it is most probable to give rise to income dissatisfaction among the sci-tech personnel. Moreover, the sci-tech personnel care a lot about the allocation system. If they consider it is the allocation system that causes large income gap, they can be quite irritated. Last, intra-regional income gap also affects the income satisfaction among the sci-tech personnel. The coefficient of determination of the intra-regional income gap is 0.249, which is way larger than the income itself, i.e., the coefficient of determination of the average monthly income, 0.106. After plugging the relative income and allocation system into the equation, the coefficient of determination of the whole equation reaches 0.533, which means that the three factors above can well contribute to deciphering the income satisfaction of the sci-tech personnel.

Table 120.2 Optimal scale regression of income satisfaction to independent variables

Variants		Equation1[1]	Equation 2	Equation 3	Equation 4	Equation 5
Sample characteristics	Gender	0.045^{***}	0.058^{***}	0.033^{***}	0.050^{***}	0.030^{***}
	Age	0.143^{***}	0.056^{***}	0.011^{***}	0.035^{***}	-0.010^{***}
	Education	-0.044^{***}	-0.124^{***}	-0.058^{***}	-0.069^{***}	-0.064^{***}
	Title	-0.048^{***}	-0.059^{***}	-0.045^{***}	-0.049^{***}	-0.048^{***}
	Work sonority	-0.147^{***}	-0.149^{***}	-0.051^{***}	-0.112^{***}	-0.046^{***}
Absolute income gap	Monthly earnings		0.360^{***}	0.106^{***}		0.077^{***}
Relative income gap	Regional income level			0.249^{***}		0.204^{***}
	Industrial income level			0.320^{***}		0.298^{***}
	Allocation system			0.305^{***}		0.289^{***}
Living burden	Family income proportion				0.065^{***}	-0.066^{***}
	Income and living				0.510^{***}	0.154^{***}
F value		16.268	119.397	630.471	293.122	568.089
P value		0.000	0.000	0.000	0.000	0.000
Regulated R^2		0.013	0.123	0.553	0.262	0.569

Note: *** indicates significance on the level of 0.01

(4) Living burden. Equation 4 probes into the influence that living burden exerts on the income satisfaction among the sci-tech personnel, which includes the proportion that individual income accounts for of the whole family earnings and the income and expenses. For example, the family might have a high income living an affluent life, or the family can only manage to meet the ends. It can be told from Equation 4 that the proportion that individual income accounts for of the whole family earnings does not influence the income satisfaction much. Its estimated coefficient is only 0.065 while that of the latter (the income and expenses) is as large as 0.51. This indicates that income and expenses can relatively affect the income satisfaction of the sci-tech personnel more.
(5) Equation 5 brings into all the independent variants after it gets rid of the interference from sample characteristics. The results show that compared to Equation 3, the coefficient of determination is increased from 0. 533 to 0.569, which indicates that living burden can impact upon the income satisfaction of the sci-tech personnel yet to a slight extent. Equation 5 again shows that income gap, especially the industry income gap, is the major influencing factor of income satisfaction of the sci-tech personnel.

120.5 Conclusion

Via logistic regression method, optimal scale regression method and some other methods, this study probed into factors that can influence the income satisfaction of sci-tech personnel. It was discovered that compared with income itself, income gap is more liable to give rise to dissatisfaction among the sci-tech personnel. Among all the inequality of allocation, inter-industry income gap can cause the most discontent, followed in order by the allocation system of the working unit, intra-regional income gap, and living burden. The last influential factor is actually the income itself. Currently, it is urgent to reform the social allocation system. The results in this study imply that it might be meaningless to re-invigorate the allocation system just by increase the income. It is more important to lessen the inequality in allocation, improve the allocation system in the working unit, and reasonably regulate the income level based on the actual situation and consumption level in different regions, thus practically easing the life burden on the sci-tech personnel.

References

1. Stacey, A.J.: Inequity in social exchange. In: Berkowitz, L. (ed.) Advances in Experimental Social Psychology, vol. 2, pp. 267–299. Academic Press, New York (1965)
2. Lawler, E.E., Hackman, J.R.: Corporate profits and employee satisfaction: Must they be in conflict? Calif. Manage. Rev. **14**, 46–55 (1992)

3. Heneman, H.G., Schwab, D.P.: An evaluation of research on expectancy theory predictions of employee performance. Psychol. Bull. **78**(1), 1–9 (2003)
4. Ye, Q., Dai, D., Wang, H.: A study of the influence of environmental factors on income satisfaction-based on an empirical study on a Chinese mobile communications operator. Sci. Technol. Manage. Res. **15**, 123–130 (2007)
5. Zhang, J., Li, L., Zang, Z.: An empirical study of the influencing factors of income satisfaction of sci-tech personnel based on decision trees. Forum on Science and Technology in China, 10, unpaged (2007)
6. Chen, T., Li, L.: A difference analysis of the incentive income satisfaction of sci-tech personnel. Science of Science and Management of S. & T, 05, unpaged (2008)
7. Zhang, M., Wu, M.: Measures to intensify the allocation system reform. Development Forum **4**(49) (2008)
8. Chen, R., Jing, R.: Examining task-oriented diversity: A multilevel analysis of the computing industry. AISS: Adv. Inf. Sci. Serv. Sci. **4**(22), 738–744 (2012)
9. Xie, X., Xue, S.: An empirical study of the income satisfaction of the enterprise human resource management. Sci. Manage. Res. **9**, 318–321 (2009)
10. He, W., Long, L.: A review of the dimensions of income satisfaction and their functions. Ruan Kexue (Soft Sci.) **11**, 87–91 (2010)
11. Yife, G.: A study of approaches to increase employees' income. Market Modernization **476**, 288–299 (2011)
12. Liu, B., Wang, H., Yang, W.: A study of the measurement of income satisfaction and its functional mechanism—take the government employees as the subjects. Psychol. Sci. **3**, 717–721 (2011)
13. Zhang, J., Lai, P.: A study of the income satisfaction of the sci-tech personnel. J. Hohai Univ. (Philos. Soc. Sci.) **10**(4), 451-454 (2008)

Chapter 121
Understanding Knowledge Sharing Willingness in Virtual Academic Community: A Survey Study on College Students

Rui Liu and Xiao Shao

Abstract Virtual Academic Community, as a type of Learning Community, features on the exchange of academic-related ideas, experience, documents, comments and feedbacks. Although there is a growing interest in Learning Community and Virtual Communities of Practice, few studies have examined the influencing factors of knowledge sharing in Virtual Academic Community from an integrated viewpoint including personal and cultural perspectives. This paper conducts a survey study on the influencing factors of college students' knowledge sharing willingness in Virtual Academic Community. The results indicate that self-efficacy, personal outcome expectation, guanxi and gaining face have positive effects on knowledge sharing willingness. Additionally, self-efficacy has a positive effect on community-related outcome expectation. Finally, research limitations are drawn and discussed.

Keywords Virtual academic community · Knowledge sharing · College students

121.1 Introduction

Virtual Community of Practice (VCoP) centered upon the communications and interactions of participants to generate specific domain knowledge that enables the participants to learn from, contribute to, and collectively build upon that knowledge [1]. The terms Learning Community and Virtual Community of Practice are often used interchangeably, as both relate to the process of learning and the socialization that serves to facilitate learning [2]. Learning Communities have been linked to online education for promoting critical thinking skills and

R. Liu (✉) · X. Shao
School of Information Management, Central China Normal University, Wuhan, China
e-mail: liuruiccnu@hotmail.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_121,

facilitating the achievement of learning outcomes [3, 4]. Virtual Academic Community, as a type of Learning Community, featured on the exchange of academic-related ideas, experience, documents, comments and feedbacks. Unfortunately, knowledge sharing among such communities has not lived up to expectation.

The question of why individuals share their knowledge in Virtual Academic Community has been examined largely in the context of learning communities. The researches have proved that students' heightened sense of belonging has positive relationship with higher-level thinking and problem solving [5, 6], and sense of community functions as one critical factor that influences knowledge sharing [7]. Some researches on academic blog communities reveal that trust and motivation have positive relationships with knowledge sharing [8, 9]. Related studies on VCoP have applied Social Cognitive Theory, Social Capital Theory, Social Exchange Theory, TAM [10–12] to illustrate or test knowledge sharing and most scholars focus on either personal factors [13–16] or contextual factors [17–19]. Huang et al. tested how face and guanxi affect knowledge sharing in general VCoP [20]. However, few studies have empirically examined cultural factors on knowledge sharing and influence of college students' knowledge sharing in Virtual Academic Community remains a critical area on which few studies have been performed.

121.2 Theoretical Model and Research Hypotheses

In this study, an integrated framework was proposed to develop a more comprehensive perspective of the relationships between college students' knowledge sharing willingness and its personal as well as cultural influencing factors in Virtual Academic Community.

Social Cognitive Theory is a widely accepted model for validating individual behavior [11]. This theory states that an individual will take an action that has personal cognition in a social environment. Furthermore, individual's cognition to act in a certain way has two basic determinants: self-efficacy and outcome expectation. Therefore, Social Cognitive Theory was used to examine how self-efficacy, personal outcome expectations and community-related outcome expectations affect knowledge sharing willingness.

In Chinese cultural context, guanxi defines one's place in the social structure and provides security, trust and a prescribed role [20]. Guanxi largely reduces the opportunistic behavior and encourages the commitment from the Chinese counterpart and at the same time enhances the channel performance as well. Closely related to guanxi is the idea of Face. The claim to face may rest on the basis of status, whether ascribed or achieved; it may also vary according to the group with which a person is interacting [21]. Since Virtual Academic Community is rooted in specific culture and serves as a communication channel involving human interaction, cultural factors as guanxi and face may exert influence on knowledge

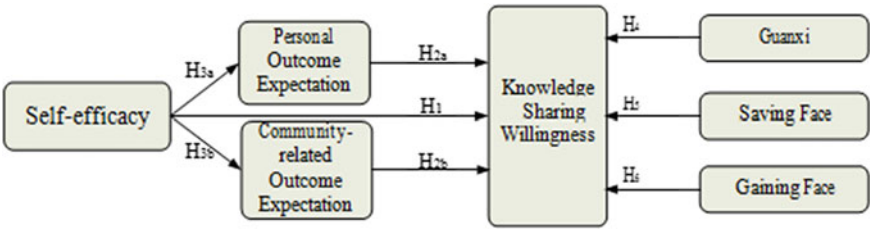


Fig. 121.1 Theoretical model

sharing. Therefore, Guanxi-Face theory was applied to examine how Guanxi, Saving Face and Gaining Face affect Knowledge Sharing Willingness. The theoretical model is proposed in Fig. 121.1.

Bock et al. examined that self-efficacy influenced knowledge sharing willingness [17]. Kankanhalli et al. considered self-efficacy as an intrinsic incentive and tested its positive effect in Electronic Knowledge Repositories [10], and this study included web-based self-efficacy (WBSE) and knowledge creation self-efficacy (KCSE) for studying self-efficacy. WBSE refers to a learner's beliefs about his/her capabilities in using the functions of the website. KCSE refers to a learner's beliefs about his/her capabilities in articulating the ideas and experiences, synthesizing knowledge, and learning from others by embodying explicit knowledge into tacit knowledge. In our study, knowledge sharing willingness is defined as people's intention to share knowledge [17]. Hence, we propose,

H1: Self-efficacy has a positive effect on knowledge sharing willingness.

Following Compeau and Higgins [11], we identified personal outcome expectations and community-related outcome expectations. Studies on Social Cognitive Theory suggested that individuals would share knowledge with the expectations of enriching knowledge, seeking support and making friends [22]. Individuals would also share knowledge with the expectation of helping the VCoP to accumulate knowledge, continue its operation and grow [16]. Hence, we propose,

H2a: Personal outcome expectation has a positive effect on knowledge sharing willingness.

H2b: Community-related outcome expectation has a positive effect on knowledge sharing willingness.

Marakas' research found that self-efficacy has positive effect on outcome expectation via individual action performance [23]. Hsu et al. also proved knowledge sharing self-efficacy positively effects both personal outcome expectation and community-related outcome expectation [14]. Hence, we propose,

H3a: Self-efficacy has a positive effect on personal outcome expectation.

H3b: Self-efficacy has a positive effect on community-related outcome expectation.

The Chinese word "guanxi" refers to draw on connections in order to secure favors in personal relations. It contains implicit mutual obligations, assurances, understanding and governs Chinese attitudes toward long-term social and business

relationships. Kotlarsky and Oshri's research indicated that guanxi promotes knowledge transfer among companies [24]. Guanxi was also proven to positively influence trust and knowledge transfer among individuals [25]. Hence, we propose,

H4: Guanxi has a positive effect on knowledge sharing willingness.

Face has two tendencies namely gaining face and saving face. Gaining face means the establishment of positive self-image; while saving face is avoid of losing face. The exposure of personal knowledge and mistake can be embarrassing and disrespectable. In that case, people are less likely to contribute knowledge and communicate with other in the fear of losing face [21]. Hence, we propose,

H5: Saving face has a negative effect on knowledge sharing willingness.

Chu suggested that gaining face is acquired by self-representation [21]. When one's advantage and capacity apply with others' expectation, recognition and respect is gained. Hence, we propose,

H6: Gaining face has a positive effect on knowledge sharing willingness.

121.3 Data Analysis and Research Results

In this study, measurement items were adapted from the literature wherever possible. A pretest of questionnaire was performed by 3 experts in knowledge management field to assess its logical consistencies, ease of understanding, sequence of items and contextual relevance. The final questionnaire consisted of 28 items was conducted after a pilot study, and a five-point Likert scale was adopted with anchors ranging from strongly disagree (1) to strongly agree (5). The questionnaire was designed to target on college students' knowledge willingness on emuch. Emuch is a Virtual Academic Community in China which gathers 300000 members and its subjects cover a large variety of academic disciplines. After data collection, 230 questionnaires were returned, 11 of which were deemed invalid resulting in a total of 219 complete and effective responses.

121.3.1 Analysis of Measurement Model

Following the recommended two-stage analytical procedures, we used LISREL 8.70 to first assess the measurement model and then examine the structural model.

Firstly, LISREL8.70 was used to apply confirmatory factor analysis (CFA). For convergent validity testing, the results in Table 121.1 showed that AVE (0.4472) and CR (0.6107) of self-efficacy were below the benchmarks, AVE of personal outcome expectation (0.4560) and AVE of community-related outcome expectation (0.3275) were also below the accepted thresholds. After theoretical analysis, X3, X6 and X10 were deleted. Then AVE, factor loadings and CR were above the thresholds.

Table 121.1 Convergent validity testing of self-efficacy, personal outcome expectation and community-related outcome expectation

Latent variable	Measure	Loading	AVE	CR	Cronbach's α
Self-efficacy	X1	0.73	0.55	0.71	0.70
	X2	0.71			
Personal outcome expectation	X4	0.83	0.64	0.88	0.82
	X5	0.76			
	X7	0.82			
	X8	0.74			
Community-related outcome expectation	X9	0.78	0.59	0.81	0.80
	X11	0.73			
	X12	0.82			

Table 121.2 Discriminant validity testing of self-efficacy, personal outcome expectation and community-related outcome expectation

Sqrt AVE	Guanxi	Saving face	Gaining face
Self-efficacy	**0.7416**		
Personal outcome expectation	0.2336	**0.8000**	
Community-related outcome expectation	0.4562	0.6783	**0.7681**

Table 121.3 Convergent validity testing of guanxi, saving face and gaining face

Latent variable	Measure	Loading	AVE	CR	Cronbach's α
Guanxi	X14	0.61	0.53	0.64	0.73
	X15	0.57			
	X16	0.64			
	X17	0.69			
Saving face	X18	0.72	0.67	0.89	0.76
	X19	0.81			
	X20	0.82			
	X21	0.73			
Gaining face	X22	0.95	0.62	0.83	0.82
	X23	0.81			
	X24	0.63			

As can be seen in Table 121.2, test of discriminant validity was acceptable.

For convergent validity testing of guanxi, saving face and gaining face, CFA results showed that loading of X13 was 0.17, AVE of Guanxi was 0.3026. After deleting X13, AVE of three latent variables exceeded the generally accepted value Table 121.3.

The Discriminant Validity test results were in Table 121.4.

The CFA results of knowledge sharing willingness can be seen in Table 121.5.

Table 121.4 Discriminant validity testing of guanxi, saving face and gaining face

Sqrt AVE	Guanxi	Saving face	Gaining face
Guanxi	**0.7280**		
Saving face	0.3409	**0.8185**	
Gaining face	0.1623	0.7058	**0.7874**

Table 121.5 CFA results of knowledge sharing willingness

Latent variable	Measure	Loading	Cronbach's α
Knowledge sharing	Y1	0.64	0.73
	Y2	0.89	
	Y3	0.82	
	Y4	0.69	

121.3.2 Assessment of Structural Model

The structural model reflecting the assumed linear, causal relationships among the latent variables was tested. The model fit indices showed that $\chi 2$ to degrees of freedom ratio of 2.24 ($\chi 2 = 533.12$; df = 238), CFI = 0.96, NFI = 0.93 and RMSEA = 0.086, suggesting that the model fit the data well.

As expected, the path coefficient of self-efficacy was 0.30 (T = 2.91) indicating self-efficacy's positive relationship with knowledge sharing willingness, thus H1 was supported. Personal outcome expectation had a significant positive effect on knowledge sharing willingness (path coefficient = 0.39, T = 4.48), while community-related outcome expectation (path coefficient = −0.24, T < 0) had no significant influence on knowledge sharing willingness. Consequently, hypotheses H2a was supported while H2b was not. Contrary to hypothesis H3a, the results showed an insignificant path between self-efficacy (path coefficient = −0.07, T < 0) and personal outcome expectation, while self-efficacy (path coefficient = 0.27, T = 2.94) positively affected community-related outcome expectation, supporting H3b. The results revealed guanxi (path coefficient = 0.32,T = 3.08) and gaining face (path coefficient = 0.24,T = 2.13) were positively related to knowledge sharing willingness, providing support for H4 and H6. The result also showed that saving face had no significant influence on knowledge sharing willingness (path coefficient = −0.36, T < 0), thus H5 was not supported.

According to data analysis, self-efficacy and personal outcome expectation showed significant and positive effects on knowledge sharing willingness. The findings reveal that college students execute knowledge sharing to be a capable, superior means of achieving personal objectives, express a high willingness to share their knowledge in Virtual Academic Community. This is in consistency with Bock and Kim's finding [16]. Contrary to our expectations, community-related outcome expectation did not have a significant impact on knowledge sharing willingness. According to Purvis et al. [19], the lack of interior regulations

and incentives in VCoP may hinder members' willingness to sharing knowledge. One plausible reason is that, in Virtual Academic Community like emuch, knowledge sharing activity is not that intense, so members don't hold the belief that one should contribute knowledge for community development. This study provides support that self-efficacy had positively effect on community-related outcome expectation, but self-efficacy had insignificant influence on personal outcome expectation. One possible explanation is that web-based self-efficacy was excluded in our study for good measurement model design, leaving only knowledge-creation self-efficacy for testing its relationship to community-related outcome expectation. The results indicate that supported hypotheses for web-based self-efficacy may not apply to knowledge-creation self-efficacy. Study also indicates that guanxi and gaining face had positive effects on knowledge sharing willingness, while saving face didn't exhibit a significant negative effect on knowledge sharing willingness. We infer that the Virtual Academic Community provides a platform for people to express themselves freely and take off the heavy burden of maintaining the established face in real world. This is consistent with the nature of web for easy communication.

121.4 Conclusion

This survey study was targeted on much. Authors tested college students' knowledge sharing willingness in Virtual Academic Community and found that self-efficacy, personal outcome expectation, guanxi and gaining face positively affected knowledge sharing willingness. Also, self-efficacy had a positive effect on community-related outcome expectation.

Although the findings are encouraging, the present study has certain limitations. This study examined only one aspect of self-efficacy. Authors did not include web-based self-efficacy in the measurement model, thus the results could have been impacted. Further study should include better measures to examine how self-efficacy influences outcome expectation in Virtual Academic Community.

Acknowledgments Supported by the Fundamental Research Funds for the Central Universities, Central China Normal University Dangui Project 2010 (120002040304).

References

1. Lee, F.S., et al.: Virtual community informatics: A review and research agenda. J. Inf. Technol. Theor. Appl. **5**(1), 47–61 (2003)
2. Reigeluth, C.: Instructional-Design Theories and Models Volume II: A New Paradigm of Instructional Theory, pp. 269–292. Lawrence Erlbaum Associates, New Jersey (1999)
3. Tinto, V.: Learning communities: Building gateways to student success, Annual Meeting of the American College Personnel Association (2004)

4. Gibbs, P., et al.: Preliminary thoughts on a praxis of higher education teaching. Teach. High. Educ. **9**(2), 183–194 (2004)
5. Uribe, D., et al.: The effect of computer-mediated collaborative learning on solving ill-defined problems. Educ. Tech. Res. Dev. **51**(1), 5–19 (2003)
6. Rovai, A.P.: Building a sense of community at a distance. Int. Rev. Res Open Distance Learn. **3**(1), 1–16 (2002)
7. Richard, C.O., et al.: A comparison of student satisfaction and value of academic community between blended and online sections of a university-level educational foundations course. Internet High. Educ. **14**(3), 164–174 (2011)
8. Yu, T.K., et al.: Exploring factors that influence knowledge sharing via weblogs. Comput. Hum. Behav. **26**(1), 32–41 (2010)
9. Chai, S., Kim, M.: What makes bloggers share knowledge? An investigation on the role of trust. Int. J. Inf. Manage. **30**(5), 408–415 (2010)
10. Kankanhalli et al.: Contributing knowledge to electronic knowledge repositories: An empirical investigation. MIS Quarterly, **29**(1), 113–143 (2005)
11. Compeau, D.R., Higgins, C.A.: Social cognitive theory and individual reactions to computing technology: A longitudinal study. MIS Quarterly **23**(2), 145–158 (1999)
12. Nahapiet, J., Ghoshal, S.: Social capital, intellectual capital, and the organizational advantage. Acad. Manag. Rev. **23**(2), 242–266 (1998)
13. Lin, M.J., et al.: Fostering the determinants of knowledge sharing in professional virtual communities. Comput. Hum. Behav. **25**(4), 929–939 (2009)
14. Hsu et al.: Knowledge sharing behavior in virtual communities: the relationship between trust, self-efficacy, and outcome expectations. Hum. Comput. Stud. **65**(2), 153–169 (2007)
15. Chui et al.: Understanding knowledge sharing in Virtual Communities: An integration of social capital and social cognitive theories. Decis. Support Syst. **2**(42), 1872–1888 (2006)
16. Bock, G.W., Kim, Y.G.: Breaking the myths of rewards: An exploratory study of attitudes about knowledge sharing. Inf. Resour. Manage. J. **15**(2), 145–149 (2002)
17. Wasko, M.M., Faraj, S.: Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. MIS Quarterly **29**(1), 35–57 (2005)
18. Bock, G.W., et al.: Behavioral intention formation in knowledge sharing: Examine the roles of extrinsic motivators, social-psychological forces, and organizational climate. MIS Quarterly **29**(1), 87–111 (2005)
19. Purvis, R.L., et al.: The assimilation of knowledge platforms in organizations: An empirical investigation. Organ. Sci. **12**(2), 117–135 (2001)
20. Huang, Q., et al.: Impact of personal and cultural factors on knowledge sharing in China. Asia Pacific J. Manage. **25**(3), 451–471 (2008)
21. Chu, R.L.: Social interactions among the Chinese: On the issue of face. Chin. Soc. Psychol. Rev. **15**(3), 79–106 (2006)
22. Zhang, Y., Hiltz, S.R.: Factors that influence online relationship development in a knowledge sharing community, Ninth American conference on information systems (2003)
23. Marakas, G., et al.: The multilevel and multifaceted character of computer self-efficacy: Toward clarification of the construct and an integrative framework for research. Inf. Syst. Res. **92**, 126–163 (1998)
24. Kotlarsky, J., Oshri, L.: Social ties, knowledge sharing and successful collaboration in globally distributed system development projects. Eur. J. Inf. Syst. **14**(1), 37–48 (2005)
25. Ramasamy, B., et al.: Is guanxi (relationship) a bridge to knowledge transfer? J. Bus Res. **59**(1), 130–139 (2006)

Chapter 122
An Application of Entrepreneurship Score Model in College Student Entrepreneurship Education

Guanxin Yao, Jing Xu and Jian Xu

Abstract For avoiding the waste of entrepreneurship resources and resolving the problem of lacking pertinence in entrepreneurship education, as well as to realize the value of the limited resources to the greatest degree, fifty students who have established their own enterprises are surveyed with questionnaires to find the eighteen main factors influencing the entrepreneurship. Taking the correlation between these factors into account, the principal component analysis method was applied and five main components representing the different indicators characteristics of entrepreneurs were filtered out. Finally the score model in college student entrepreneurship can be constructed. Through this model, entrepreneurial aspirations of college students and the groups with higher scores can be obtained, which can improve the pertinence of entrepreneurship education. Therefore, this model provides a reasonable basis for the effective delivery of entrepreneurship education resources.

Keywords Entrepreneurship education · Entrepreneurship competence · Principal component analysis · Scoring model

122.1 Introduction

In recent years, government and society have paid much attention on social employment, and China has been beset by this problem for many years. Due to the gloomy employment market, more graduated students determine to become

J. Xu (✉) · J. Xu
School of Management, Jiangsu University, Zhenjiang, China
e-mail: xujing1990mail@126.com

G. Yao
Yancheng Institute of Technology, Yancheng, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_122,

self-employed, which is not only in accordance with economic trend, but also a best way for the realization of students' human capital value [1], besides it is also beneficial for national development, social stability and the improvement of each student's life quality. It is quite necessary to research the entrepreneurial phenomenon of students, and increasing efforts to support students in starting their own business particularly in the weak market after the financial crisis. In recent years, the government gradually increases investment in college students' entrepreneurship, and schools have also invested more in entrepreneurial education, however, the ratio of students who built their own enterprises to all graduate is just 2 %, negligible in compare with the 28 % of United States. The fundamental reason for this phenomenon is the waste and inappropriate distribution of entrepreneurship assistance and educational resources, which also directly bring down the success rate of entrepreneurship. In domestic entrepreneurship education, [2] and [3] separately conducted theoretical analysis of college students entrepreneurship from the perspective of college students' entrepreneurship consciousness, capability cultivation and entrepreneurship education innovation. Guoxing [4] focuses his attention on the evaluation of the students' entrepreneurial capability, but limited to the students in the entrepreneurship competition. Kun [5] first proposed the concept of entrepreneurship endowment, but did not give a specific calculation method. So designing a selecting method which is suitable for the college students' entrepreneurship will not only improve the college students' entrepreneurial capability, but also promote employment, and it can provide theoretical basis for reform of talents education in higher-education institutions at the same time.

122.2 The Team Heterogeneity Theory

In foreign country, the research and practice of college students' entrepreneurial ability originate from the development of entrepreneurship education. Although entrepreneurship has already drawn much attention, it still lacks a clear definition until today [6]. Meanwhile many scholars began to focus their attention on entrepreneurship heterogeneity theory. The study mainly go through three stages: the first stage emphasizes the effects of external heterogeneity(age, sex, degree of education, etc.) on entrepreneurship performance; on the basis of the first stage, the second stage regards enterprise development process as an adjustment variable to reflect the complexity of the entrepreneurial process requirements; the third phase of the study cares the impact of external heterogeneity as well as internal heterogeneity (values, cognitive style, experience, background, etc.) on entrepreneurial performance, and instead of researching separately it considers the interaction between the two factors [7]. The team heterogeneity theory has become mature in the continuous development process, the essence of the theory is that different entrepreneur possesses different entrepreneurial qualities, and their capability of

resolving the difficulties encountered in the venture are quite different, thus affecting their entrepreneurial performance.

Recent studies mainly focus on inspiration, development and evolution analysis [8] of entrepreneurship, while few concern the psychology of college entrepreneur. There are some studies comparing student who has become self-employed and those who hasn't, however most of these studies lack in depth research. Bige Askun in his studies on entrepreneurship education in Turkish public universities points out that entrepreneurship courses in public universities in Turkey are not sufficient to provide skills or mindsets that are required for creating entrepreneurs that can contribute to economic growth and employment for students [9]. In conclusion, it is useful and urgent to establish an entrepreneurs' score calculation model.

122.3 Calculating Model for Scores of College Students Entrepreneurship

122.3.1 Sample Selection

Given that it is hard to obtain relevant resources, fifty college entrepreneurs are surveyed with questionnaires from about ten higher-education institutions in Jiangsu Province. Sample Characteristics are as follows in Table 122.1.

122.3.2 Analysis of Factors Affecting College Student Entrepreneurship

College students are not born with the ideas of entrepreneurship. It is influenced by inner interest and value as well as social environment and school education. Lots of studies indicate that individual background affects the entrepreneurship aspiration greatly. Brockhaus in his study argues age, sex, degree of education and parents' roles are all relevant factors; Krueger and Carsud also consider that the offspring may have a strong desire to be entrepreneurs if their parents have an enterprise. Some studies in this field prove educational background is also a key

Table 122.1 Sample characteristics

No.	Male	Female	Total
Sophomore	3	2	5
Junior	13	7	20
Senior	14	6	20
Master	3	2	5
Total	33	17	50

Table 122.2 Potential factors affecting college students' entrepreneurial willingness

	Measurable variable ($X_1 \sim X_{18}$)
Factors affecting the willingness of college students venture	Confidence of entrepreneur Perseverance of entrepreneur Negative mood management Ability of risk-taking Communication Teamwork Honesty and Faith Self-control Opportunity recognition Available resource Ability of practice Basic knowledge of business Basic knowledge of entrepreneurship background Professional knowledge Knowledge of law and policy Frequency of extramural practice or concurrent post Sense of internship Level of independence

factor as well. For example, students in business administration have a stronger entrepreneurship consciousness than those of engineering subjects. Feng Lei, on the basis of document analysis, as well as visiting and interviewing, defines factors which have been generally accepted as entrepreneurship competence, such as innovation, sociability, risk bearing, teamwork, decision-making, commitment, self-control and knowledge acquisition [10]. And he also believes these factors conversely influence entrepreneurship willingness. Besides, students' entrepreneurship psychology are also greatly influenced by their personal characteristics, including innovation, opportunity, perseverance, leadership, teamwork, desire for power, self- realization and spirit of adventure. Considering that personal background are not controllable, therefore, in the formation process of the college students' entrepreneurial awareness, cultivating their entrepreneurship thoughts can only be achieved through the indirect method of changing their living environment and education model. Accordingly, factors that reflect inner psychology, entrepreneurship competence and knowledge are selected as main factors; Table 122.2 gives specific information of these factors.

122.3.3 Principal Component Analysis of Each Factor

A superficial glance at the correlation coefficient matrix of various factors shows a strong correlation between the 18 variables. Considering those variable overlaps with each other, for better exploring those factors' influence on college students'

Table 122.3 Total variance explained

No.	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.16	39.869	39.869	7.176	39.869	39.869
2	1.71	9.511	49.38	1.712	9.511	49.38
3	1.5	8.334	57.714	1.5	8.334	57.714
4	1.24	6.935	64.649	1.248	6.935	64.649
5	1.09	5.607	70.256	1.009	5.607	70.256
6	0.90	5.009	75.265			
7	0.75	4.179	79.445			
8	0.70	3.912	83.357			
9	0.56	3.123	86.48			
10	0.49	2.733	89.213			
11	0.45	2.5	91.713			
12	0.35	1.928	93.64			
13	0.31	1.74	95.38			
14	0.26	1.44	96.82			
15	0.20	1.088	97.908			
16	0.17	0.928	98.835			
17	0.14	0.773	99.608			
18	0.07	0.392	100			

entrepreneurial willingness, principal component analysis method is chosen to select main factors and their contribution. Table 122.3 shows variance contribution and cumulative contribution rate of each variable, and the latent roots of previous five are more than 1, so only these five are selected by SPSS. The first principal component variance accounts for 39.87 % of the total, while the previous five account for 70.26 %.

Since the five principal components are found, according to factor loading matrix (not listed due to space limitations), the function of principal component can be determined, which is:

$$\begin{cases} F_1 = 0.865ZX_1 + 0.849ZX_2 + 0.749ZX_3 + 0.474ZX_4 + \cdots + 0.629ZX_{18} \\ F_2 = 0.107ZX_1 - 0.004ZX_2 + 0.219ZX_3 - 0.454ZX_4 + \cdots + 1.131ZX_{18} \\ \cdots\cdots \\ F_5 = 0.110ZX_1 - 0.097ZX_2 - 0.240ZX_3 + 0.040ZX_4 + \cdots + 0.149ZX_{18} \end{cases} \tag{122.1}$$

And then standard function of principal component can be drawn directly according to the component score coefficient matrix (Table 122.4):

$$\begin{cases} ZF_1 = 0.120ZX_1 + 0.118ZX_2 + 0.104ZX_3 + \cdots + 0.088ZX_{18} \\ ZF_2 = 0.063ZX_1 - 0.002ZX_2 + 0.128ZX_3 + \cdots + 0.077ZX_{18} \\ \cdots\cdots \\ ZF_5 = 0.109ZX_1 - 0.096ZX_2 - 0.238ZX_3 + \cdots + 0.148ZX_{18} \end{cases} \tag{122.2}$$

Table 122.4 Component score coefficient matrix

Component	1	2	3	4	5
Confidence of entrepreneur	0.12	0.063	0.156	0.137	0.109
Perseverance of entrepreneur	0.118	−0.002	0.142	0.197	−0.096
Negative mood management	0.104	0.128	−0.002	−0.139	−0.238
Ability of risk-taking	0.066	−0.265	0.064	0.254	0.039
Communication	0.102	0.187	0.053	−0.12	0.116
Teamwork	0.117	0.082	−0.084	−0.073	−0.152
Honesty and Faith	0.065	0.371	0.002	−0.011	0.397
Self-control	−0.014	0.187	0.039	0.611	0.055
Opportunity recognition	0.102	−0.128	0.207	−0.2	−0.08
Available resource	0.1	−0.06	0.129	0.097	−0.229
Ability of practice	0.075	0.299	−0.112	−0.065	−0.175
Basic knowledge of business	0.083	−0.269	0.059	−0.122	0.258
Rudiments on entrepreneurship	0.069	−0.148	−0.096	−0.269	−0.13
Background					
Professional knowledge	0.053	−0.129	−0.336	0.016	0.604
Knowledge on law and policy	0.087	−0.011	−0.201	0.175	−0.225
Practice or concurrent post	−0.023	0.128	0.521	−0.199	0.291
Sense of responsibility	0.104	−0.191	0.131	0.207	0.142
Level of independence	0.088	0.077	−0.283	−0.109	0.148

Table 122.4 shows that different factor has different contribution to each principal component. In the first principal component, the coefficients of X_1 (Confidence of entrepreneur), X_2 (Perseverance of entrepreneur), X_6 (teamwork), X_3 (Negative mood management), X_4 (sense of responsibility) and X_5 (communication) are larger, so F_1 can be regarded as a composite indicator which reflect inner quality of the entrepreneur. The coefficient of X_{11} (ability of practice) is much bigger than others in F_2 and it can be considered as an indicator to reflect the entrepreneur's ability of transferring ideas into practice. In F_3 the coefficients of X_{16} (frequency of internship) and X_9 (Opportunity recognition) are larger, so F_3 reflects the ability of seizing commercial chances. In the same way, we can summarize that F_4 reflects entrepreneur's capability of self-control, and F_5 reflects entrepreneur's knowledge level. Finally, according to formula (3), the score calculation formula can be drawn as:

$$F = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \cdots + \lambda_5} F_1 + \frac{\lambda_2}{\lambda_1 + \lambda_2 + \cdots + \lambda_5} F_2 + \cdots + \frac{\lambda_5}{\lambda_1 + \lambda_2 + \cdots + \lambda_5} F_5 \tag{122.3}$$

In the formula (122.3),

$$\lambda_1 = 7.176, \lambda_2 = 1.712, \lambda_3 = 1.5, \lambda_4 = 1.248, \lambda_5 = 1.009$$

122.4 Conclusion

The ultimate purpose of this study is discovering a series of measures, so as to improve the success rate of entrepreneurship among college students. Numerous studies have shown that the entrepreneurial capability can be enhanced through education. Complicate as it is, college students' entrepreneurship has attracted much attention abroad in the theoretical and empirical study. Foreign scholars incline to improve the educational system, method and technique of entrepreneurship through case study, role-play, outward bound and visiting, etc., and the result is quite fruitful in recent years. Domestic study in this field stems from 2000 and some typical entrepreneurship educational type are developed, such as on-class lecturing of Renmin University of China, entrepreneurship in practice of Beijing University of Aeronautics and Astronautics and entrepreneurship education system of Shanghai Jiao Tong University.

Although the government, society and universities are paying more attention and investing more than before on entrepreneurship, but most of the entrepreneurship education, particularly domestic education, are formalistic and lack of pertinence These studies only encourage students to participate in business training which will benefit in short term, but with low efficiency in the long run, and not conform with the ideas of intensification, detailed management and sustainable development of modern society.

Resources are limited, students entrepreneurship education unable and unnecessary to cover each student. Universities can evaluate students with entrepreneurship score model, so as to discover potential entrepreneurs. Universities can also analyze and improve entrepreneurship skills for students who have strong willingness but score low. For the students who lack entrepreneurship willingness, universities can judge whether the potential entrepreneur can develop into a real entrepreneur or not by analyzing their score. Those who achieve high scores should be encouraged with certain measures. Universities may also improve the competence of students with little willingness in entrepreneurship and scoring low, thus alleviating the problem of employment. In general, by precisely investing the scarce resources to the true potential entrepreneurs, the entrepreneurship score model may greatly improve the effectiveness of educational resource investment. It will increase the success rate of entrepreneurship, and at the same time, it can boost the student employment level and maximize the benefit.

Acknowledgments This research was supported by the Philosophy and Social Sciences Key research project- Students entrepreneurial ability structure and training system research (No. 2011ZDIXM039) of Jiangsu Higher Education Institutions in China, the Graduate Innovation Project of Jiangsu Province (CXLX12_0681).

References

1. Jie, Y.: Training for the identification of opportunities In Students' entrepreneurial process. J. Hubei Correspondence Univ. **24**(3), 16–17 (2011)
2. Weigang, H., Xin, Z.: Some considerations about enhancing college students pioneering consciousness and ability training. J. Tianjin Normal Univ. (Soc. Sci.) **3**, 69–71 (2011)
3. Xuanli, Z.: Entrepreneurial educational alliance: Conception of breaking through bottlenecks of entrepreneurial education in colleges. J. Hunan Univ. Sci Technol. (Soc. Sci. Ed.), **14**(5), 169–171(2011)
4. Guoxing, R., Ying, H.: The study of comprehensive evaluation model of entrepreneurial team entrepreneurial capacity in student business plan competition. J. Northeast Normal Univ. (Philos. Soc. Sci.) **6**, 224–227 (2009)
5. Kun, W., Yuntao, L.: Research on the relation between entrepreneurship education and entrepreneurial willingness mediating by entrepreneurial endowment. J. Wuhan Univ. Technol. (Soc. Sci. Ed.), **25**(2), 189–193 (2012)
6. Yalou, H.: Research model and framework for reconstruction of the entrepreneurial process. Enterp. Econ. **10**, 12–15 (2009)
7. Yufang, H., Mingqing, Q.: Review on the heterogeneity of entrepreneurial team. Sci. Technol. Manage. Res. **16**, 142–145 (2010)
8. Shane, S., Stuart, T.: Organizational endowments and the performance of university start-ups. Manage. Sci. **48** (2000)
9. Askun, B., Yildirim, N.: Insights on entrepreneurship education in public universities. Procedia Soc. Behav. Sci. **24**, 663–676 (2011)
10. Lei, F.: Entrepreneurial skills education on the willingness of college students ventures. Enterp. Econ. **3**, 75–80 (2011)

Chapter 123
The Research on Teaching Methods of Object-Oriented Approach in Management Information Systems Curriculum

Xianhong Liu

Abstract It is difficult for students who major in economics and management to learn object-oriented approach based on Unified Modeling Language (UML) in Management Information Systems (MIS) curriculum. To solve this problem, four methods about how to teach this approach are offered to teachers as follows. Summarizing consistency rules helps students to understand logical relationships between UML diagrams. Defining the core of UML helps students to grasp emphases. Designing creative experiments helps students to understand concepts and notations of UML deeply. With the help of specific program codes, abstract UML diagrams can be explained for students more clearly. These methods are proved to be effective to teach this approach in MIS curriculum.

Keywords Object-oriented approach · UML · Management information systems · Teaching methods

123.1 Introduction

There are two basic development approaches of Management Information Systems (MIS): traditional structured approach and newer object-oriented approach [1]. It is a growing trend that more and more teachers begin to teach the object-oriented approach in MIS curriculum. The object-oriented approach represented in this paper is based on Unified Modelling Language (UML) from the Object Management Group (OMG). These are some similar researches on how to teach this approach. Some researchers argue that teachers should combine UML theory with MIS development practice [2, 3]. Some researchers argue that teachers should

X. Liu (✉)
School of Management, Henan University of Science and Technology, Luoyang, China
e-mail: Lxh2072@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_123,

apply the method of case teaching [4, 5]. Other researchers argue some specific teaching methods such as project teaching method, goal driven teaching method [6, 7]. Unfortunately, almost all of these researches are given to students who major in computer science. Compared to these students, students who major in economics and management have some obvious disadvantages such as: lack of knowledge of software engineering, inadequate experience of software development, unsatisfactory abilities to think logically. Therefore, these researches aren't always to be applicative to students who major in economics and management. I have taught object-oriented approach in MIS curriculum for over 5 years. I am glad to share experience with other teachers.

123.2 Teaching Methods of Object-Oriented Approach in MIS Curriculum

As mentioned above, students who major in economics and management are different from students who major in computer science. We should apply appropriate teaching methods according with their features. Four teaching methods are represented as follows.

123.2.1 Summarize Consistency Rules to Help Students to Understand Logical Relationships between UML Diagrams

OMG defines 13 types of diagrams in UML 2.x which support developers to model information systems from different angles and levels. This kind of multi-view modelling way is useful to reduce complexity of models. Unfortunately, it leads to a problem, the logical relationships between these diagrams puzzle students. Even if they grasp the concepts and symbols of each UML diagram, they often make inconsistencies between UML diagrams. How can we judge UML diagrams are consistent or not? Rules are needed. Therefore, it is necessary to summarize consistency rules between UML Diagrams. Quite frankly, it is not easy to summarize consistency rules between UML diagrams completely. So far, I gain about 20 consistency rules between UML diagrams. Some of them are listed below.

Rule 01. An object in Sequence Diagrams must be an instance of a normal (not abstract) class in Class Diagrams.

Rule 02. When the name of a class is modified in Class Diagrams, the name of the corresponding class must be updated synchronously in Sequence Diagrams.

Rule 03. If an object sends a message to another object in Sequence Diagrams, there must be a dependency relationship between the two classes that the two objects belong to respectively. Contrariwise, if there is a dependency relationship

between two classes, there must be at least one message interaction between the two objects.

Rule 04. A message of Sequence Diagrams must correspond to an operation of the receiver (an object), and the operation is visible to the sender (an object).

Rule 05. If a class is deleted in Class Diagrams, the corresponding objects and messages (corresponding to operations) of the class should be deleted synchronously in Sequence Diagrams.

Rule 06. The same object in Sequence Diagrams and Communication Diagrams must belong to the same class in Class Diagrams.

Rule 07. An object that State Machine Diagrams describes must correspond to an instance of a normal class in Class Diagrams.

Rule 08. If a class is deleted in Class Diagram completely (not hid only), all the corresponding State Machine Diagrams should be deleted.

Rule 09. A state in State Machine Diagrams must be a legitimate value of one or more attribute of a class in Class Diagrams.

Rule 10. If an action in State Machine Diagrams is to call an operation of a class, the operation should keep consistent with the definition of the operation in Class Diagrams, including the name, parameters, etc.

Rule 11. If an activity in State Machine Diagrams is to call an operation, the operation should be a message in Sequence Diagrams.

Rule 12. A use case in Use Case Diagrams must be assigned to operations of classes in Class Diagrams.

Rule 13. If an Activity Diagram is used to describe a use case, activities and swimlanes in the Activity Diagram correspond to operations and classes in Class Diagrams respectively.

Students are asked to debate these rules with their group to understand them. Afterward, students are asked to abide by these rules when they model an information system. These rules, just as a guide, are very useful for students to grasp the relationships between UML diagrams. Another paper of mine, identification and check of inconsistencies between UML diagrams, discusses why these rules are correct in detail.

123.2.2 Define the Core of UML to Help Students to Grasp Emphases

UML evolved from the work of many renowned computer scientists, corporation and other organizations. UML absorbed their respective object-oriented concepts and notations, fused them into a single, standardized model [8]. As a result, UML is very complex and enormous. Users often complain that UML is too difficult to learn and use. Students encounter the same problem. So we should define the core of UML, tell student what is the key content that they should learn and grasp.

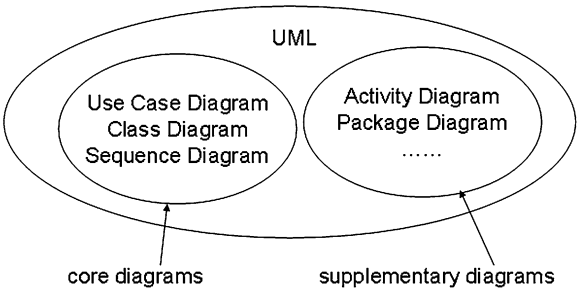


Fig. 123.1 The core of UML

I divide UML diagrams into two parts: core diagrams and supplementary diagrams, as shown in Fig. 123.1. The core diagrams include Use Case Diagrams, Class Diagrams and Sequence Diagrams. The supplementary diagrams include Activity Diagram, Package Diagram and other 8 types of diagrams. According to this result, students should concentrate on Use Case Diagrams, Class Diagrams and Sequence Diagrams. Why? Although UML provides wide support on all kinds of users from different fields, our students are these who are learning MIS curriculum. The goal of our students is to grasp the process and technology of information systems analysis and design. So they should learn the most frequent UML diagrams in this domain. Use Case Diagrams are an important requirement model which helps developers to capture and represent behaviors of information systems. So Use Case Diagrams are indispensable. It is accepted that Class Diagrams are the heart of object-oriented models. Sequence Diagrams are a useful tool to find out the operations of classes. It can also provide developers with direct bases of coding programs. So Sequence Diagrams is also a necessary.

Therefore, Use Case Diagrams, Class Diagrams and Sequence Diagrams are the core of UML in the domain of information systems analysis and design. They deservedly become the emphases of student in MIS curriculum. So I concentrate on teaching these diagrams, and ask students do a great of corresponding exercises.

123.2.3 Design Creative Experiments to Help Students to Understand Concepts and Notations of UML Deeply

Teachers provide students with detailed guides to an experiment, and then students complete the experiment step by step according to the guides. Most experiments are completed in this way. In the process above, students are passive. They only do what teachers ask them to do. They don't think and do anything farther. In order to solve this problem, I design some creative experiments. Creative experiments define that only an experiment title is given to students, and they are asked to work out an experiment scheme by themselves. No guides or directions are provided by teacher. Students themselves do all the things needed in experiments such as

collecting information, installing software, analyzing data and so on. Some creative experiments are listed as follows.

Experiment 1. Secondary development of UML modelling tools based on open source software.

Experiment 2. Represent three-tier architecture using UML and code corresponding C# programs .

Experiment 3. Represent design patterns using UML.

Experiment 4. Test the functions of checking inconsistencies between UML diagrams of UML modelling tools.

Experiment 5. Test compatibility of UML modelling tools with windows 8.

Experiment 6. Test cross-platform features of UML modelling tools.

Experiment 7. Analyze performance of UML modelling tools based on web explorer.

Experiment 8. Test and compare domestic UML modelling tools.

Experiment 9. Install and use UML plug-in modules.

Students are asked to complete these experiments in their spare time. Let's take experiment 1 as an example. I only provide students with the experiment title: Secondary development of UML modelling tools based on open source software. I don't tell them how to complete this experiment. They may first open http://sourceforge.net or http://www.oschina.net, then download source codes of software such as StarUML, Smartuml, C# Uml Designer, ArgoUML, afterward modify these codes using Pascal, C#, Java or other program languages. Eventually, they submit reports to me.

Compared to traditional experiments, creative experiments can motivate creativity and autonomy of students. Although it's full of difficulties and frustrations, students gain deep cognizance in the process of overcoming them.

123.2.4 Explain Abstract UML Diagrams Using Specific Program Codes

Generally speaking, MIS curriculum is more abstract than programming curriculum. Because of this, most students who major in economics and management find it is hard to learn UML in the beginning. Although modeling and programming are belonging to two different phases in MIS development, it can't be denied that there is a close relation between them. In the process of MIS development, they can't be completely separated. So it is helpful to explain the concepts and models of UML using program codes.

For example, an Include Relationship and an Extend Relationship are two important parts of Use Case Diagrams. An Include Relationship means that a use case includes the behavior defined in another use case. An Extend Relationship defines that a relationship from an extending use case to an extended use case that specifies how and when extending use case can be inserted into the behavior

defined in the extended use case [9]. They are very similar. Many students confuse them. It is easy to explain the differences between them using program code. I take two types of promotion information as an example: 20 % discount for all goods, 20 % discount for goods priced above 200 RMB. I regard "pay" and "discount" as two use cases, and code two programs as follows to simulate them. In the left program, "pay" calls "discount" unconditionally, which represents the Include Relationship. In the right program, "pay" calls "discount" conditionally (price ≥ 200), which represents the Extend Relationship.

A program to simulate the Include Relationship:

```
public class coat
{
   public string name;
   public double price;

   public coat(string x, double y)
   {
      name = x;
      price = y;
   }

   public string pay()
   {
      discount();
      return "payment:" + price;
   }

   public void discount()
   {
      price = price * 0.8;
   }
}
```

A program to simulate the Extend Relationship:

```
public class coat
{
   public string name;
   public double price;

   public coat(string x, double y)
   {
      name = x;
      price = y;
   }

   public string pay()
   {
      if (price > = 200)
         discount(0.8);
      return "payment:" + price;
   }

   public void discount(double x)
   {
      price = price * x;
      //x: discount rate
   }
}
```

By running the program and comparing the results, students can understand the similarities and differences between them. It is helpful for students to understand them. Other UML diagrams can also be explained using program code.

123.3 Conclusion

The object-oriented approach based on UML is an important content of MIS curriculum. This paper summarizes teaching methods of this content in MIS curriculum. These methods are proved to be effective and efficient. I hope they are

helpful to other teachers who teach this content to students who major in economics and management. Teachers should pay attention to that students must gain enough object-oriented knowledge before MIS curriculum. Thus, curriculum structure may be optimized.

Acknowledgments This work was supported by the Foundation of Henan University of Science and Technology Education Reformation (No. 2012Y-024).

References

1. Satzinger, J.W.: Systems analysis and design. In: A Changing World 6th Edition. pp. 98–99. Course Technology, Boston (2011)
2. Sabine, M., Jean-Paul, R.: Teaching object-oriented modelling and UML to various audiences. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 40–54. Springer Verlag, Heidelberg (2010)
3. Jin, Z.: Research on UML teaching. China Educ. Innov. Herald **14**(31), 172–173 (2008)
4. Lethbridge, T.C., Mussbacher, G., Forward, A., Badreddin, O.: Teaching UML using umple: Applying model-oriented programming in the classroom. 24th IEEE-CS conference on software engineering education and training, CSEE and T 2011—Proceedings, Piscataway: IEEE Computer Society, pp. 421–428 (2011)
5. Yu, Z., Wei, T.: Think and research on case teaching of UML. China Electr. Power Educ. **10**(12), 86–88 (2008)
6. Lei, J.: Application of behaviour-oriented project teaching method in UML teaching. Sci. Technol. Inf. **14**(26), 190–192 (2009)
7. Dai, C.: Application of goal driven teaching method in UML and use of UML modelling tools course. Sci. Technol. Inf. **24**(1), 193–194 (2008)
8. Weizhong, S., Fuqing, Y.: Object-oriented Systems analysis, pp. 197–199. Tsinghua University Press, Beijing (2007)
9. Booch, G., Rumbaugh, J., Jacobson, I.: The Unified Modelling Language User Guide, 2nd edn. pp. 14–15. Addison-Wesley Professional, Indianapolis (2005)

Chapter 124
Engineering Material Management Platform for Nuclear Power Plant

Zhifeng Tan, Zheng Zhang, Liqing Hu, Shan Chen and Zhijun Wang

Abstract The materials are the basis of the nuclear power plant construction. Effective nuclear power engineering material management can accelerate the construction and reduce project cost of the nuclear power plant. In order to put the design of the source data into engineering material management, regulate the process of material procurement management and provide material procurement data and design drawings data in real-time for site management, this paper proposed the solution of the engineering material EPC (engineering, procurement, construction) based on VPRM system and SAP system and finally realized it. This solution also solved the main problems existing in engineering material management business of the nuclear power plant construction, and meets the meticulous management concept at the same time.

Keywords Nuclear power plant · EPC · Material management · VPRM · SAP

124.1 Introduction

The materials throughout the whole process of the nuclear power plant life cycle are the basis for nuclear power plant construction. Effective nuclear power engineering material management can accelerate the construction, shorten the duration, reduce project cost and increase the profit of the nuclear power plant. The process of the nuclear power plant construction is complex. In addition, the nuclear power plant building is a 'tri-ongoing' (three things at the same time) project, so the process of engineering material management is much more complex [1].

Z. Tan (✉) · Z. Zhang · L. Hu · S. Chen · Z. Wang
Information Technology Center China Nuclear Power Technology Research Institute, China Guangdong Nuclear Power Holding Co., Ltd, Shenzhen, China
e-mail: tanzhifeng@cgnpc.com.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_124,

The process of the nuclear power plant building is as following: since design, procurement, construction have the characteristic of long cycle and high cost [2], it is not suitable if the pattern is designed and then purchased, purchased and then constructed. It must be the executive of design, procurement and construction in parallel, while the biggest problem of the pattern is easy to cause independent work of EPC, lack of linkage, and bring the following issues: procurement data is not entirely comes from the design; construction materials cannot be purchased timely, the alteration of design and procurement and construction cannot communicate and exchange in time; the Lack of the uniform material code and code system through the whole process of the engineering material [3]. This paper focused on the building of the engineering material management platform based on the VPRM and SAP system. The paper is organized as follows: Sect. 124.1: Brief introduction. Section 124.2: Introduce engineering material code and the process of material management. Section 124.3: Propose a solution of the engineering material EPC based on VPRM and SAP system. Section 124.4: The deployment and implement of platform. Sect. 124.5: Some related works. Concluding remarks are added.

124.2 Nuclear Power Plant Material Management

The whole process of EPC in nuclear power plant engineering material management included two parts: engineering material code and the process of material management.

124.2.1 The Engineering Material Code

In the nuclear power plant, the engineering material is mainly divided into three categories: the material of equipment items, the material of tags items and the material of bulk type [4].

The Material of equipment items: the equipment items which are all distinguishable and uniquely identified by a unique label (called Process Tag). The Process Tag contains information on functional aspects of the equipment for per unit number of the nuclear power plant (number of unit, type of equipment, system it belongs to, sequence number) which give indications on what the equipment is and what its purpose is, for example: 1RCP1001PO. This type and its related management process are generally applied to main and not numerous equipment items.

The Material of tags items: The material of tags is a tagged type, applied to numerous and identical ("identical" meaning of the same type or model) items. For example, tags items are identified by a unique tag at the design stage. However, as they might be very numerous, the tags are ignored during the purchasing

process where the type or model is preferred to identify them. The purchasing process can manage them as bulk material, and the quantity of items needed for one kind is defined as the total number of items of this kind (example: N tagged valves of a given kind will be supplied as one item with the quantity "N"). By consequence of this definition, the tags of items are unique when they are tagged but multiple when they are treated as bulk and their units of measure are also the piece. The type is also called "Bulk Equipment" or "Itemized Equipment Type".

The Material of bulk type: This type material called "Bulk Material" is applied to all undistinguishable untagged items. Namely pipes, fittings, plates, cable trays or cables are bulk material, as any piece of it is undistinguishable from another one (of the same type or model). The management of bulk material requires the definition of a catalogue, which is simply the list of all types of a material, plus some technical attributes defining each type. The number and the nature of these attributes vary in the kind of material described. The bulk items are then simply defined through the identifier of the type within the catalogue, plus their variable dimension if any (ex: length for a pipe or a cable), or just their number when there are pieces (ex: tees, elbows). Their units of measure are the physical unit of their variable dimension ("m", "m^2", "kg" …) or "piece", which is also called "Bulk Material Type".

124.2.2 The Process of Engineering Material Management

The process of engineering material management is the whole process tracing management for the engineering material from making the BOM (bill of material) of equipment and material procurement according to the design of the nuclear power engineering to the delivery and fix of site material. The main process have produced procurement technical specification and BOM, divided into procurement package, registered project, selected supplier, bidding, contract, purchase order, expediting, equipment Inspection, packing, dispatch, traffic, site material receipts, stores, issued material, tracking management of site material application and so on [5, 6].

Upwards just display the process of project material management and the actual business of project material management process is very complex. It not only needs to monitor supplier equipment production and delivery, but also to understand the actual site material demand, adjust material delivery plan in time and avoid inventory overstocking.

After investigation and research, materials cost account for more than 60 % of the project cost, plus transport charges. It may be more than 70 %, therefore, to ensure the smooth progress, reducing cost and improving economic efficiency for engineering material management plays an important role [7].

124.3 The Whole Process of Project Material Management EPC Technology Solutions

Material management of the nuclear power engineering related to the design, procurement, construction, had the characteristic of high technical, long duration and complicated process. To feed back the process information and results to the upstream and downstream in the process of engineering material management, it requires the use of information systems. The following would provide the whole process of project material management EPC technology solutions based on VPRM system and SAP system.

124.3.1 VPRM System

Vantage Project Resources Management (VPRM) contains seventeen functional modules. According to the nuclear power plant engineering material classification features, the following modules would be frequently-used. VPRM modules covering design, procurement and site phase, see the Table 124.1.

In addition, during the design phase, VPRM realizes the professional engineering material extraction and summarization by the data interface.

VPRM, VPE and PDMS relationship diagram can be seen in Fig 124.1.

AVEVA provided date interface between VPRM and other design platforms, such as VPRM and Vantage Project Engineering (VPE), VPRM and Plant Design Management System (PDMS). These interfaces would import data automatically from VPE database that included equipment, valves and instrumentation with EDF Code System (ECS) encoded items, and also realized the piping material automatic data import in PDMS pipe model.

Table 124.1 The function modules in VPRM

Phases	Modules		Functions
Design phase	ADMIN	Administration	Administration and equipment management
	MCAT	Material catalogue	Material catalogue
	SPEC	Specification	Piping specification
	MTO	Material take-off	Engineering material extraction, including PMTO and GMTO
	REQ	Requisition	Engineering material requisition
Procurement phase	VDB	Vender database	Supplier information
	PROC	Procurement	Inquiry, purchase orders, inspection, packing
Site phase	SITE	Site	Material site management, including transport, cargo, warehousing, material issue

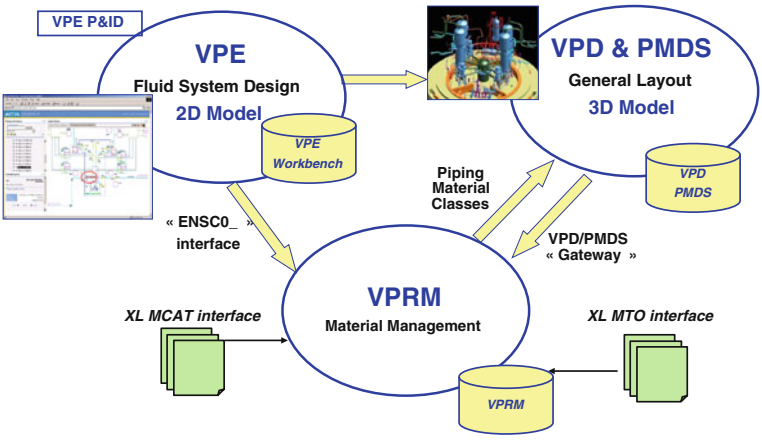


Fig. 124.1 The relationship diagram among VPRM and design platform

124.3.2 SAP System

Material management (MM) is the function modules of SAP system related to the material management, which has the following functions:

- Material master data management. All of the enterprise material data integrated in a single material database, eliminated data redundancy problem and realized the data-sharing among the departments;
- Vendor master data management. It integrated all of the supplier information that supplied for the enterprise, and also realized the data-sharing;
- Procurement master data management. It included all kinds of master data which it was related to the procurement activities, for example: transaction information recording, source list, quota agreement, framework agreement, contract, order and delivery schedule, price terms, etc.
- Procurement management. It used to be the purchase of material and service, also included the function of confirming the supply of goods, purchase order and the control of account payable.
- Stock management. It focused on the material receipt, transfer, issue, the material quantity, the amount of money management and inventory, etc.

124.3.3 The Design of the Platform Framework

After investigating the situation of company informatization, understanding the function of VPRM and SAP, analyzing the problems existed in the business, the project had proposed building the engineering material management system which was the whole process management of EPC [8, 9]. Figure 124.2 shows the framework of the engineering material management platform.

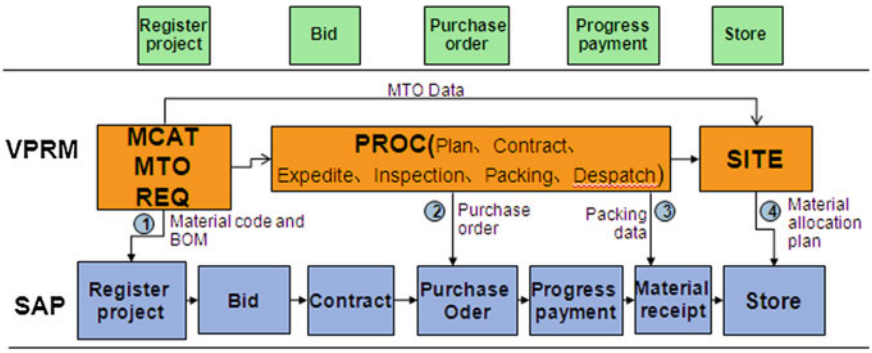


Fig. 124.2 The framework of the engineering material management platform

It could be seen from the Fig. 124.2, the platform is composed of VPRM and SAP system, through the interface development to realize the data transmission between the two systems. The platform made full use of the VPRM system function, used the MCAT module unified the engineering material code, created a database of material; then it used the MTO module to take off the material data from engineering design tools, for example PDMS and VPE, and provided the data support for the REQ module to create the BOM; it used the REQ module to produce the material requisition and used the PROC to manage the process of purchase order, expediting, inspection, packing, dispatch, and so on; At last, it used the SITE module to receive the MTO data to make the plan of material issued.

SAP system received the material code from VPRM MCAT module by the interface, received the BOM and completed the project registry and approval; In the SAP system carried out the management of bid and contract, according to the data of purchase order and packing realized the data synchronization of procurement management between VPRM and SAP which completed the management of progress payment and material receipts in SAP system.

The platform made full use of the advantages of VPRM system and SAP system, achieved the whole process management of engineering material in EPC, provided the individuation function and filled up the lack of VPRM by SAP. The platform promoted the realization of uniform code system, integrated the design outcomes of the design department into the engineering material supply chain, and provided the data support to procurement and construction, which is convenient for them to deal with the relative work depending on the design data.

124.4 The Implement of Engineering Material Management Platform

The implement of the platform includes three parts: first is the requirements development; second is the interface development; the third one is the configuration of the operation parameter.

The platform realized the EPC process of the material management. It related to the tools of design, the system and function of procurement and construction material management [10]. VPD and VPE are the tools of design. The platform used VPRM system and SAP system to realize the procurement and site material management. The workflows of engineering material management platform see Fig. 124.3.

The details of the workflows are stated as follows: VPRM system utilizes the MCAT module to create the uniform codes for the bulk material, and transmits the uniform material codes to the SAP system via the interface between VPRM and SAP. And then the SAP system generates the material master data on the basis of the uniforms material codes from the VPRM system. In the process of the VPRM system, SPEC module extracted the data from MCAT module according to the requirement of design and transmits it to the design tools such as PDMS. The design engineer selected the related components based on the material data and completed 3D design in PDMS, transmitted the sheet of material that generating in PDMS to VPRM MTO module, transmitted the equipment data that generating in VPE to VPRM ADMIN module. In the VPRM system, REQ module created the MR (material requisition) depending on the data of ADMIN and MTO. The SAP system created the register project after receiving the MR data from VPRM REQ module, accomplished the bidding and contract. The VPRM system created the purchase order and packing, and transmitted the data to SAP MM module. The VPRM SITE module received the data from PROC and MTO, created the scheduling and allocation of construction and transmitted it to the SAP store management, and carried out the management of store value.

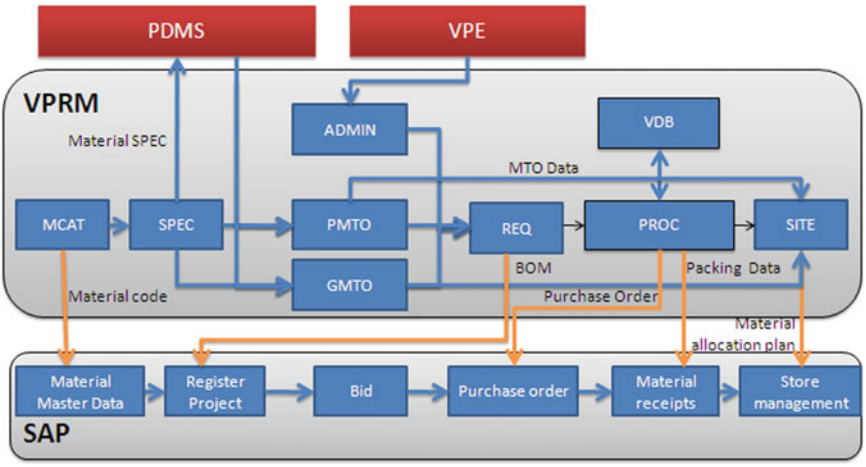


Fig. 124.3 The workflow of engineering material management platform

124.5 Conclusion

The building of the engineering material management put the design of the source data into engineering materials management, established and realized the uniform material code system.

The platform involved many departments and business and it is also complex. It does not show the platform had been built after the platform development has been finished. It also needs to guide the users to use the platform to complete related work. When the system process and the actual business process do not match completely, users need to be guided to change and regulate the related business process or change the corresponding system process according to the requirement of actual business. Thus the platform played a greater role in the engineering business and would made greater contribution for the nuclear power plant engineering construction.

References

1. Lewis, H.T., Livesey, C.A.: Materials management in the airframe industry. Harv. Bus. Rev. **22**(4), 477–494 (2009)
2. Sillak, G.C.: Project code of accounts—as applied in engineering, procurement and construction for the process industries. Cost Eng. **44**(7), 42–53 (2002)
3. (Kees) Berends, T.C.: Cooperative contracting on major engineering and construction projects. Eng. Econ. **51**(1), 35–51 (2006)
4. Wu, T., Xu J., Tan Z.: Engineering material management pattern research and implementation for generation III nuclear power plant. In: Innovations bio-inspired Computing and Applications (IBICA), pp. 290–293 (2011)
5. Johansson, E.: Information management for materials supply systems design. Int. J. Prod. Res. **47**(8), 2217–2229 (2009)
6. Kasim, N.B., Anumba, C.J., Dainty, A.R.J.: Improving materials management practices on fast-track construction projects. In: Khosrowshahi, F (ed.), 21st Annual ARCOM Conference, 7–9 Sept 2005. SOAS. University of London, Association of Researchers in Construction Management, vol. 2, pp. 793–802. (2005)
7. Perera, S., Karunasena, G.: Best value IT procurement for construction organizations. In: AACE International Transactions, pp. 1–10 (2004)
8. Evbuomwan, N.F.O., Anumba, C.J.: An integrated framework for concurrent lifecycle design and construction. Adv. Eng. Softw. **2**(7–9), 587–597 (1998)
9. Perng, Y.-H., Chang, C.-L.: Data mining for government construction procurement. Build. Res. Inf. **32**(4), 329–338 (2004)
10. Said, H., El-Rayes, K.: Optimizing material procurement and storage on construction sites. J. Const. Eng. Manage. **137**(6), 421–431 (2011)

Chapter 125
An Analysis of IT Customer Service Quality System

Yongrui An

Abstract As today's society is experiencing a rapid progress, it is far more important for customers to pursue better service than other factors like price, taste, etc. at the same premise. So it is obvious that the establishment of customer service system is a top priority for any company. And enhancing the quality of service for each customer is a crucial segment because it will directly affect the finance situation of any company. Thus, it is quite necessary to set up an effective customer service system for enterprises under the help of continuous improvement and development of social and economic system. The focus of this paper is how to improve the quality of customer service on the base of working experience in the Large Enterprise Group and the current trend of modern society.

Keywords Quality of service · IT customer service quality

125.1 Introduction

Nowadays, the fierce competition is pushing every company to enhance customer's satisfaction so as to improve its own competitiveness. In a word, a company has to constantly improve its measures to enhance customer satisfaction in this modern market economy system which is also an indispensably inevitable product under the modern markets economy system. This not only reflects the macro control of modern economic system, but also reflects today's economy development up to an unprecedented height [1].

With the development of science and technology, modern city is entering a highly electronic and informational age, electronic will replace everything [2].

Y. An (✉)
Information Technology Center, China Nuclear Power Technology Research Institute, China Guangdong Nuclear Power Holding Co., Ltd., Shenzhen, China
e-mail: pcitayr@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_125,

And improving the customer satisfaction is playing a particularly important role in clean energy technology which is characterized by high-quality, high precision and advance. All control system and operating system are under electronic control which will bring about a higher request for the quality of IT customer service. As clean energy industry is characterized by high risk and high quality, IT customer service will not allow any error which will bring IT customer service for clean energy industry to a much higher level, not only to meet the daily needs, but also to improve the ability to handle the emergency. Thus clean energy power industry will be more reliable on the IT customer service. Moreover, IT customer service will set up a new role as backbone in clean energy industry. In other words, IT customer service is playing a more important role in perfecting the quality of IT customer service.

125.2 Significance

To build up an effective IT customer service system is much needed; good service quality does not mean searching the highest service level. So pursuit of qualified service must be clearly understood by most companies.

The service can be undertaken after the user's demand was settled and examining of that was finished. And the most important aspect is to meet the continual requests of users which will inevitably bring constant changes to service content. Also the needs of customer are stratified, so this requires us to adapt to the changing trend step by step and ensure the service continuity which is one of the basic requirements of service quality. It emphasizes that any organization should provides high quality service at any time and any place. And for group enterprise, there is a most tough question to handle which is the result of different professional technology level of each division. And this will definitely mislead decision-making of top leaders. So it is quite necessary to provide training courses and evaluating them at a regular interval to avoid such situation to satisfy different kind of requests from customers so that to ensure the continuity of service. And this continuity service must be functional, economic, safe, and timeless and comfort which imply that the organization should response to the demands of customer.

125.3 Overview on IT Customer Service Quality

What exactly service quality is? And how to have a full understanding of this major economic reform direction and put it into implementation? The answers are below. First of all, the definition of service quality must be clearly understood; secondly, the company should fully understand the significance of setting up the customer service quality system. Last but not least, definite the elements individual character of customers that interfere the quality of service.

For now, the Large Enterprise Power Group has used a lot of effective methods and introduced many successful IT service support process to improve the IT customer service quality. And this indeed plays an important role in IT customer service.

1. Introduction of a functional desk

A functional desk is introduced to connect the operating system and business at operating system level. As a result, the functional desk becomes the primary source of commutations on terminal service, operation and infrastructure. The mainly purpose of introduction a functional desk is to solve problems, coordinate normal operation of the repair service thus to minimize the impact on business.

2. Event management

Event management, mainly use necessary means to rehabilitate the user service which can reduce the influence of business events when interruption occurred in service [3]. These necessary means include log definition, recording and solving process to make business go back to normal and to reduce the negative effect of the business operation, thereby maintain the highest service quality usability.

3. Problem management

Problem management with characteristic of forward looking, and the focus of which is more on problem prevention than on emergency treatment. Problem management use problem prevention and fast repair to support event management, not just for solving events.

4. Configuration management

Configuration management is used to the IT department have a good understanding about how the infrastructure equipment, components link with each other to support the business processing and functioning smoothly [4]. And it is also in charge of organizing all IT assets (framework) that it serves including maintain record and verify configuration items. And all these will build up a solid foundation for emergency management, problem management, change management and release management.

5. Change management

Change management is used to reduce IT service interruption in hardware, network, system software and application software etc. [5]. caused by unauthorized and not coordinating changes. The aim of change control is to improve infrastructure and services, therefore to minimum the service interruption.

6. Release management

Release management, mainly is committed to combine the service development with service operation to ensure that all the authorized technology can be managed through ways of an appropriate check list or sighed by a relevant member [6] (Fig 125.1).

125.3.1 Definition of Service Quality

The service quality refers to the service characterized by satisfying all the potential demands of customers, and satisfies the demand of customer to better degree. It is

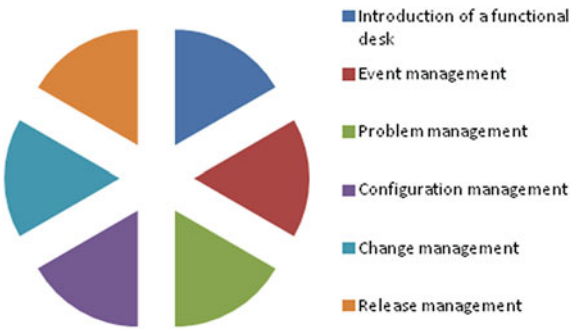


Fig. 125.1 The introduction of IT service process technology

also the lowest service level that a company can provide for a customer and consistency degree that a company should maintain.

125.3.2 Factors Infect Service Quality

In the modern service market, if a company wants to succeed, not only it has to meet the needs of customers but also has to pay more attention to the methods in the process of serving. So establishing an effective IT customer service system is a necessity. And the following factors and requirements must be fully concerned about.

1. Reliability. This means the engineer has to provide customer a reliable and accurate promise according to one's own characteristics and ability. This is the so-called think before you say, do it after you say it and write it down after you make it. And this is what the customer expecting for.
2. Response. The service department has to response to the users who propose the demands and solve the problem at the first moment. On the other hand, it will reflect whether the company has paid enough attention on this service or not. And this is one of the most important ways for a customer to find out whether the service is satisfying or not.
3. Professional skill. This requires the service personnel should have certain technical skill and ability to solve problems. Also the service personnel have to be polite and confident to let the users believe that the personnel have ability to tackle with problems. And all of above have service consciousness and good communication skills.
4. Ability to think from customers' situation. The customers will express their dissatisfaction when their rush demands are not immediately met. So the service personnel have to arm with good service consciousness and the ability to think from the customers' situation. The service personnel should try best to be nice to the customer and avoid the sensitive problem so that the personnel and

customers can better cooperate with each other to tackle the problem until the customers are satisfied.
5. Thoughtfulness. This is what the customers expecting for. So it is very important to improve the quality service level.

125.4 Promotion of Service Quality

125.4.1 Ways to Promote Service Quality

The world most famous management guru Peter F. Drucker has put forward an important theory that affects the development of world economy.

By which the importance of service in the modern economy system is pointed out: enterprise separated from society can not make profit, society can not develop without the support economy, and economic market can not normally function without high quality service. Providing a good service is the top priority of a competitive company [7]. But how to improve service is a hard nut to crack. Following are summarization of effective ways proved by facts.

1. Set up service system to improve customer satisfaction. As a big company owning emerging nuclear source, the Large Enterprise Group is struggling to build perfect IT customer service quality system, how to improve the customer satisfaction and how to serve customers effectively all of which are the top priority of a group enterprise. Generally speaking, if a company has a variety of projects, plenty of production bases and multiple customers can make a specific strategic plan according to its own situation to satisfy customer. And then a company must combine an organic integration of service concepts and service standards from different zones to collect all kinds of demand information so as to fulfill the contentment of customers and to help bolster the business. And by means of which the company can gain an edge in the project and economy market and develop a personalized operation system.
2. Aware the market dynamics and assess the demands of customers. A company full of energy must update and stretch advantage its advantages and take an objective view of development towards all changes in market and grow an insightful opinion about the market so as to reply to the changing and harsh needs of customers. Only fully aware that the demands and direction of market can a company makes a most timely and effective feedback to customers, at the same the company has ability to lead the market trend and does better to optimize the IT customers' service quality system to fulfill the overall needs of customers.
3. Think highly of customer in a correct attitude. A decrease in service quality and without a reasonable advice from customers will put a company into a danger of elimination by economy market. So think highly of customers in correct attitude is one of the most important things that guide our group.

4. Strengthen supervision and inspection. The behavior and emotion of service personnel will be varied from length of service time and the number of customers. In facing with the dynamic change of customer service on the spot, the service personnel have to be flexible to strengthen the supervision and inspection on every factor. Also the service personnel are required to have the ability to please the customers and the ability to adapt to any uncertainty while providing service.
5. Pay more attention to details. Details are always vital in the modern service concept. Every company and individual will accomplish a qualitative leap if they pay enough attention to details. And a successful company would not neglect any dullest detail so as to ensure every process go smoothly so that every customer will get the most economical and comfortable service.
6. Organize training and upgrade service skills. In the world of digitalization and e-management, the sound operation of company increasingly relies on the construction of IT support team and the upgrading of service skills. In the fierce commercial competition, the most important factors affecting the position of a company are whether the service system is advanced or not. And this will thanks to the training causes and skills promotion in all-round to meet the needs of customers fast and quickly.
7. Introduction is a perfect assessment system. A perfect assessment system can be introduced to evaluate all the service spots reward the good and punish the bad through of which a better internal competitive atmosphere can be formed. Encouraged by which every department can nurture a unique working style with a purpose to meet the need of customer and perfect the customer service quality through a way of constant innovation and flexible adoption for the requests raised by both customer and market.
8. Implementation of flexible management. The supervisor should administrate the entire service member in a flexible manner all the way with an aim to update the on spot management and try best to make sure that if a request emerges, there will be a person guided by a leader to solve it; make sure that the most positive and effective treatment will be brought out at the first moment when there is a problem.

125.4.2 Support by the System

ITIL can provide an object, rigorous, amortizable standard and norm. IT department and its final users can choose the services at different degrees by themselves in terms of their capability and needs, and formulating the basic frame and the service management reference to ITIL can ensure a better support to the business operation of the company by the IT service management. For the company, the most significance is the combination of the IT service and the business operation, thus the return of IT investment is maximized.

Table 125.1 The chronology of the IT operation and maintenance platform of the large enterprise group

Years	Name of the IT operation and maintenance platform	Description	Purpose
2002	System called computer on line	Abbreviation of computer on line-one stop service system	For a customer to register a repair list, service list; for a information center to process and trace work list; to record and manage IT assets
2004, 2009	Help system, Service system	Online support center	Accept information consultancy, distribute COL lists, coordinate and handle complaints and pay a return visit. Ensure the desk of the dispatch center, the command center of the role
2006, 2009	OVSD system OVSM system	HP products	Connector of computer on Line used for distributing work list. Increased problems with management, configuration management, change management, knowledge management module
2011	COL second system	Reformed COL system which is running now	User-oriented, the original function is re-developed and knowledge base is introduced to handle system failure fast, and convenient for a customer to track

It is already 10 years since the ITIL theory was introduced into the Large Enterprise Group in 2001. In this decade, the Large Enterprise Group has formed the ITIL theory with features and realized to improve the service. The following is the chronology of the IT operation and maintenance platform of the Large Enterprise Group (Table 125.1).

125.5 Conclusion

The current IT maintenance service related to clean energy industry faces steep challenge of service quality for the more requirements, fussy procedures, pressing time, and more and more difficult tasks. Nationwide information-based situation makes it significant to enhance technique and improve the service quality.

The Large Enterprise Group is oriented at co-existence of many projects and many plants, and at the marketing economic trend. All the staff is constantly to improve its IT customer service quality, by which to lay a solid foundation for clean energy at the process of social economic mechanism evolvement. Enhancing its service quality level is always the soul of company survival. It is a key point cannot be neglected.

References

1. Krajewski, L.J., Ritzman, L.P.: Operations management processes and value chains, Prentice Hall, pp.454–461 (2006)
2. Fitzsimmons, J.A.: Service management: operation and strategy and information technology. Second Version (2002)
3. Selig, G.J.: Implementation of IT Management, Beijing, pp.25–29 (2010)
4. http://www.questforum.org/qf_resources/brochures/TL_9000_09.pdf (2009)
5. Welch, S., Prabhu, K.: TL 9000 Quality system rquirements, New York, pp.18–29 (2009)
6. http://www.qmi.com/information_center/literature/tl_9000_english_us.pdf[M](2009)
7. Parsons, B.: Quality management system, pp.08–09 (2005)

Chapter 126
The Research and Application of IT Room Monitoring System in Nuclear Power Plant

Li-Xuan Ye and Yang Jiao

Abstract With the purpose of solving the lack of real-time monitoring of nuclear power business problem and controlling the resources operational state, four layer model of monitoring system is designed. In accordance with the computer room, system equipment, safety equipment, network equipment, application system of specialization, it realizes the centralized monitoring and integrated display of the dependent elements of nuclear power station service system. The design and application of the monitoring system can provide a lot of equipment monitoring and operation monitoring index. Aiming to greatly improve the fault timely discovery rate and running trend and to predict the hazard events, this paper includes focus on the improvement of reliability and serviceability of monitoring center and provides an all-around centralized control mode for the future of monitoring center of electric power companies.

Keywords Computer room · Monitoring · Monitoring systems

126.1 Introduction

With the enterprise information network infrastructure building and expansion step by step, the network of information system has extended to various production departments of enterprises. As the bearer of enterprise core system and the important foundation of network equipment, the comprehensive monitoring system of computer room plays an indispensable role in the information and network management. Power supply and distribution, UPS, air-conditioning, fire equipment, security systems, are essential important equipment in the room [1].

L.-X. Ye (✉) · Y. Jiao
Center of Information Technology, China Nuclear Power Technology Research Institute, China Guangdong Nuclear Power Holding Co., Ltd, Shenzhen, China
e-mail: yelixuan@cgnpc.com.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_126,

These devices are in a high degree of correlation, and provide the necessary reliable safeguard for the normal operation of the computer systems. With higher mutual coordination of equipment, a set of data center computer room monitoring systems must be established, which are used to monitor closely the operational aspect of the equipment and to deal with various faults alarm situation and possible warning situation in time to guarantee of the data center computer room normal operation.

126.2 Technical Characteristics of the Room Monitoring System

The monitoring system is the product of the development of the computer room. With the development of system technology, the monitor and control has entered the unattended operation phase. It's necessary to use reliable control systems to monitor the overall situation of the computer room [2].

The computer room monitor system has achieved the following five objectives:

(1). Provide reliable monitoring information resources of the stable operation of the various systems and equipment in the computer room;
(2). Reduce the management costs and human resources in the computer room operation;
(3). Ensure to improve the efficiency of the computer room work and to provide a safe and comfortable room environment;
(4). Monitor the operation of the computer room equipment comprehensively and generate early warning in time;
(5). Adapt the development needs, provide a scalable room environment and also provide a secure computer room environment.

126.3 The Present Sitution of Monitoring System of the Nuclear Power Plant

The data center computer room of the nuclear power is deployed in the Nuclear Power Plant at present. According to different functions it's divided into supply and distribution area, the network area, server area, communication area, safety equipment area, the duty area.

Supply and distribution areas provide energy support for the entire room, made of main distribution cabinet, UPS, etc. Network area provides the access to the entire core network, made of core switches, etc. Server area is the key area of the center room, all services provided by the production environment servers are

placed in this area. Communications areas are the stored-program control exchange and other conventional communications equipment storage area. The firewall, Internet traffic monitoring, Trojan detection, intrusion prevention and other safety equipment are placed in safety equipment area. Duty area is the work office space for the staff on duty. At present plant control system has solved the problem of the environment monitoring, and provided a stable, reliable, comfortable and safe room environment. It's the development trend of nuclear power plant control to integrate the current respective independent monitoring system into a comprehensive set of computer room monitoring system, to deal with failure alarms and early warning. Meanwhile, it's the necessary to prevent risk effectively, and establish the computer room early warning mechanism.

126.4 Design Principle of Monitoring System of The Nuclear Power Plant

The monitoring system software adopts four layer model, divided into a display layer, application layer, communication and data collection layer for the convenience of the engine room equipment for efficient unified management [3].

The data acquisition layer is the lowest level and most basic of the entire system module, including the server data acquisition software, the temperature and humidity sensor module, temperature sensor, fire system monitoring module, leakage detection systems, distribution systems and other related equipment [4].

Communication layer mainly execute the change of the data from the layer collection, and sending these data to the application layer.

The application layer is the key processing layer of the entire system. It recognizes the abnormal data and classifies the related data based on pre-set parameters, and sends fault information and early warning information to the presentation layer through analyzing and processing the data from the communication layer.

The presentation layer consists mainly of management views, performance report and fault tips, and provides management view, print failure report, maintenance personnel knowledge-based systems.

In Fig. 126.1, the entire plant control system is divided into server system monitoring, network monitoring, security devices, and monitoring computer room environment.

Integration of the four computer room monitoring software, you can use Web services to achieve the four monitoring software access granted to different users with different access control permissions. Users can access the internal network of any machine login the management console, real-time monitoring of the computer room situation.

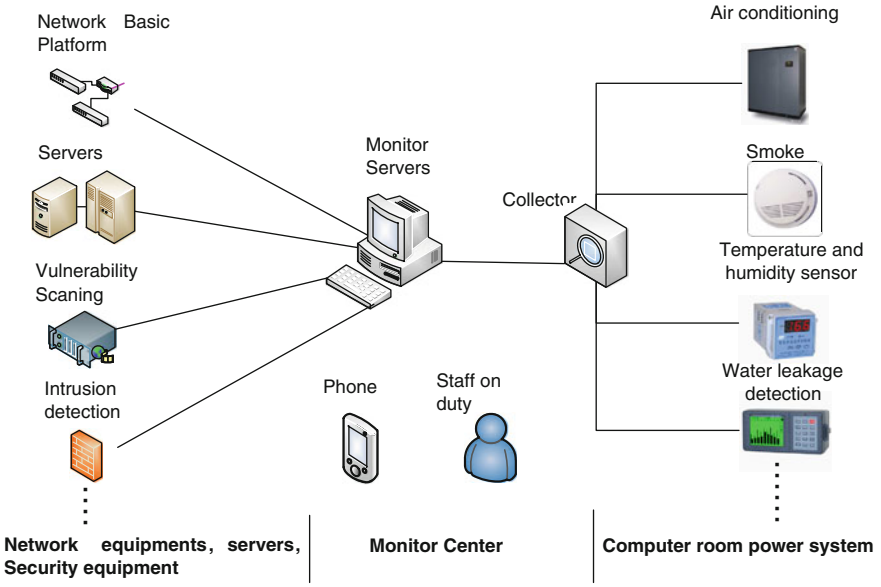


Fig. 126.1 Architecture of room monitoring system

126.5 Application and Realization of Monitoring System of the Nuclear Power Plant

The computer room monitoring system can help us to keep abreast of the operational status of the computer room equipment, and detect of anomalies. For the example of nuclear power, the center room is located in the first floor of the building, divided into distribution area, network equipment area, servers area, security devices area, communications areas and duty area. The monitor function can be divided into four parts: environment monitoring, server monitoring, the safety equipment monitoring and network monitoring according to existing equipment.

126.5.1 Environment Monitor System

The computer room environment is the basis of ensure the normal operation of the center room. In environmental monitoring, detailed monitoring of the power supply and distribution, temperature and humidity, precision air conditioning, fire and other equipment of the computer room, are shown in Fig. 126.2.

At the bottom of monitoring software-the data acquisition layer, data modules are deployed to monitor different environments parameters, which obtain real-time parameters and alarm information from the communication card.

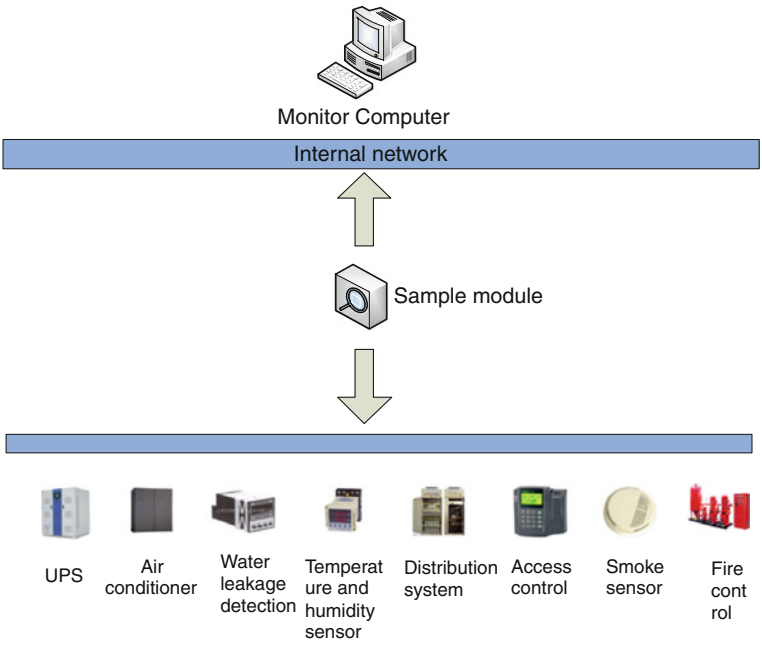


Fig. 126.2 Environmental monitoring system

126.5.2 Server Monitoring System

The server is the core part of the center room, it is divided into two parts:

The hardware parameter monitoring, we use ITIMS software on the server to monitor. CPU utilization of each server, memory, disk space data are in monitor.

The software monitoring, a different server service are monitored, focused on the database and some special software services. For other software services, we focused on the software usability, session situation monitor to ensure the normal operation of the server.

126.5.3 Network Monitor System

Network system has a direct impact on the operation of the whole system as an important part of the data center room [5]. Monitoring of network systems is divided into three platforms, the basic platform, monitor platform and process platform. Probes of the data have been deployed on the base platform to read the real time data from network equipment and communication lines; The column charts about number of information events, warning events, fault events will be seen in comprehensive event; network performance can be a complete display of

the relevant parameters of the network equipment port, such as state protocol, bandwidth, traffic, the IP address of; the dynamic network topology shows the current composition of the whole network system equipment interrelated [6].

126.5.4 Safety Equipment Monitoring System

Safety equipment, as the interface of the entire information systems to the outside world, is both the first line of defense against external invasion, and the first pass to control user access [7]. The monitoring of the safety equipment can be divided into:

IDS/IPS security Monitor: availability, extranet illegal scan, and so on;

Firewall security monitor: availability, unauthorized access, spoofing attacks;

Anti-virus security monitor: availability, virus outbreaks, virus inhibition failure;

Anti-Trojan security monitor: availability, Trojan activity state, killing.

126.5.5 Other Features of The Computer Room Monitor System

Room monitoring system not only provides real-time data monitoring for the current situation of the center room, also provides historical data query and generate relevant reports in accordance with the needs of users, and export Excel file. In operation of computer room monitoring system alarm management, the monitoring system can be linked to a number of different monitoring devices or subsystems, especially at some serious events [8].

126.6 Development Tendency of Monitoring System of the Nuclear Power Plant

126.6.1 Business Monitor

The data center of nuclear power plant is to provide the information service, as well as the operation service. Therefore, the solution of future monitoring room require much more from the business and service level and take closely relationship between a separated physical device and a real business, which finally will form business equipment views and make the availability of each device to reflect that of business.

126.6.2 Virtual Cloud Control

The trend of center future data room is virtualization. It is difficult for the monitoring system to distinguish that the monitoring sever is a physical machine or a virtual machine, and also difficult to know hardware system whether plays potential impacts on the server availability or not. The availability of virtualization platform affects directly the availability of virtual server which is running over it. Therefore, the high availability solution of virtual cloud environment will be one of the future trends [9].

126.6.3 Dependence Analysis Model

The basis of business service monitoring and virtual environment can establish a clear management relationship between each different device and then form networks between device and device. All of that need to build CMDB (Configuration Management Database), which describe the attribute information of each device and the relationships between the devices. The Fig. 126.3 is the effect of modeling carried on a nuclear power plant operation system:

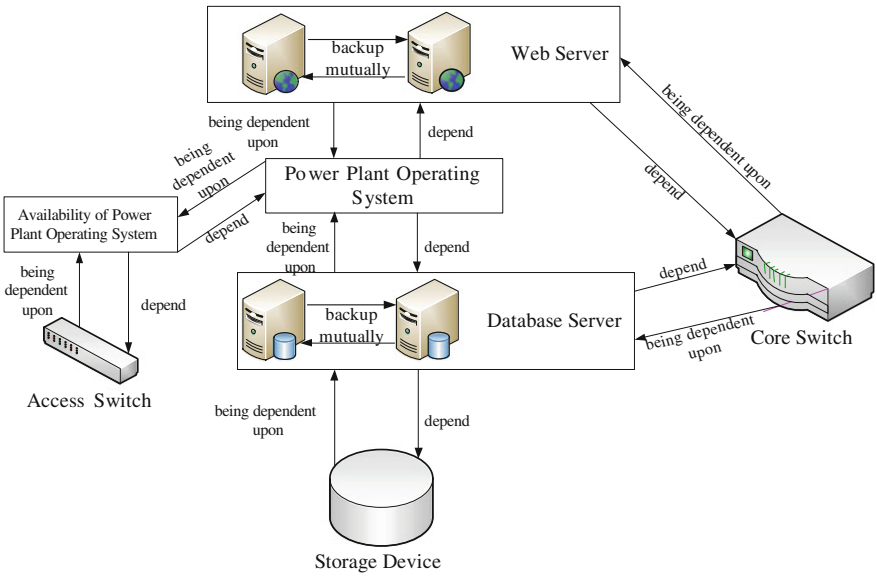


Fig. 126.3 Analysis model of dependence of power plant operating system

126.7 Conclusion

From the current service condition of nuclear data center room, the use of computer room monitoring system helps IT operation and maintenance departments solve a lot of monitor cases. Automated and intelligent management of equipment in the computer room are executed with the help of the room monitoring system. Manpower and material resources are economized effectively. Stable and reliable, efficient, convenient, safe and comfortable room environments are provided. Room monitoring system is providing the reliable and first-class technical support means for the normal operation of the data center room.

References

1. Cullen, G.L.: Server Room Climate and Power Monitoring: How to Protect Computer Equipment Against Damage & Downtime Using Low-Cost, Web-Based Devices. It Watchdogs Inc, Austin, pp 22–25, (2006)
2. Xiao-rong, Y.: Application and analysis of computer lab monitoring system in LAN. Comput. Knowl. Technol. **7**(16), 3846–3847 (2011)
3. Ying, X.: Design and implement of the monitoring system of enterprise-wide information center's computer room. Comput. Knowl. Technol. **7**(13), 3197–3199 (2011)
4. John, P.: Steve Mackay CPEng BSc (ElecEng) BSc (Hons) MBA. Practical Data Acquisition for Instrumentation and Control Systems. Newnes, San Diego, pp 3–10 (2003)
5. Nagios, W.B.: System and Network Monitoring. No Starch Press, San Francisco, pp. 78–88 (2008)
6. David, T., Perkins, R.: Remote Monitoring of SNMP-Managed LANs. Prentice Hall, Upper Saddle River, pp. 157–169 (1998)
7. Fry, C., Nystrom, M.: Security Monitoring: Proven Methods for Incident Detection on Enterprise Networks. O'Reilly Media, Sebastopol, pp. 143–146 (2009)
8. Michael, L.J.A.: The Operating Room for the 21st Century. Amer Assn of Neurological Surgeons, Park Ridge, pp. 10–15 (2003)
9. Nelson, R., Danielle R.: Virtualization, A Beginner's Guide. McGraw-Hill Osborne Media, New York, pp. 315–319 (2009)

Chapter 127
Communication Resource Management Technology of Nuclear Power Plant

Yanliang Zhou and Tianjian Li

Abstract In the information environment, resources of communication equipment become increasingly varied and complex. Except for common information management systems, equipment resources can be organized in connection. Equipment resources can form a work order. The work order has at least one object and can be broken down into smaller ones. So the work order can be separated into large numbers of subtasks in order to form hierarchical structure. The Management System is designed to break down these tasks and assign its task nodes. As a result of decomposition, all equipment resources can be managed by this system.

Keywords Nuclear power plant Communication resource management · Interface layer · Network · Software architecture

127.1 Introduction

At present, there is a great deal of systems in the Nuclear Power Plant. These systems are independent and not connected to each other. In the information environment, the Nuclear Power Plant requires a more powerful and efficient management system. This article discusses a communication resource management system which manages tens of thousands of terminal equipment of which data are shared.

The system capacity is increasing while the Nuclear Power Plant is expanding. The distributed management model is not fit for the operating requirements and management requirements. In order to remedy this deficiency, an integrated management system need to be established, and a unified and intelligent resource information management platform are needed to be created. It can produce platform optimization and efficiency improvement.

Y. Zhou (✉) · T. Li
Information Technology Center China Nuclear Power Technology Research Institute, China Guangdong Nuclear Power Holding Co. Ltd, Shenzhen, China
e-mail: zhouyanliang@cgnpc.com.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_127,

127.2 Communication Resource Management Description

127.2.1 Definition of Communication Resources

What is communication resource? Communication resources can provide strong operational capacity and dispatch combined equipment. These relationships are linked by the physical resource. Communication resources change their cycle in the process, and the equipment resources can be accessed and maintained.

127.2.2 Communication Resources of the Nuclear Power Plant

The Nuclear Power Plant communication equipment can be divided into three categories: the private branch exchange (PBX) communication resources; the emergency communications resources; and the electricity network communication resources.

The communication resources systems include messaging platforms, billing systems, mobile terminals, carrier system, broadcast system, monitor system, sound alarms system and microwave system. There are a huge number of complex relationships and line resources among these devices.

127.3 Communication Resource Management System Design

127.3.1 Software Architecture

The system architecture of communication resource management employs three-tier software architecture. The level of software design is of good scalability [1]. The system is divided into three layers: user interface presentation layer; operation logic processing layer; data access processing layer.

The following is the introduction of each layer:
The following is the introduction of each layer:

(a) User Interface Presentation Layer

The main function of the layer is representing data. In the second layer of operation logic, the user interface layer requires the core service to display the results.

(b) Operation Logic Processing Layer

The operation logic is managed by the middleware of the system. The layer manages customer services and customers' requirements. The operation logic layer

can directly have access to the data from the data access layer. In order to use a variety of logical services, the results can return to the customers.

(c) Data Access Processing Layer

The data access processing layer constitutes the third layer of the model. The layer mainly uses the relational database. It is responsible for managing the data resources and completing the data manipulation. The relational database is easy to maintain. The integrity of the database is the entity integrity and the referential integrity and the integrity of the user-defined. As a result of the model, the system greatly reduces the data redundancy and inconsistency [2].

127.3.2 System Hardware Architecture

The database server employs hot standby, and uses cluster architecture for its application. The storage server adopts the disk array (RAID 5). The hardware structure is shown in Fig. 127.1:

127.3.3 System Interface Design

The resource management system is not an isolated system, which need integrating existing management system and unified management. The interface figure and interface agreement are shown in Fig. 127.2:

(a) Telephone Yellow Pages System (TYS) Interface

The communication resource management system is one-way synchronization the TYS Yellow system.

(b) IT Server Management (ITSM) System Interface

In the communication operation process, the resource management system interface adopts the ITSM system for real-time transfer by TCP \ IP protocol. The results are transmitted back to the ITSM [3, 4].

(c) IT One-stop Service Platform Interface

The resource management System, which is merged into the IT one-stop service platform, provides the required operating platform for the IT one-stop service platform.

(d) Message Platform Interface

The message records are passed to the resource management system. In order to achieve unified billing management, the message platform interface uses TCP\IP protocol.

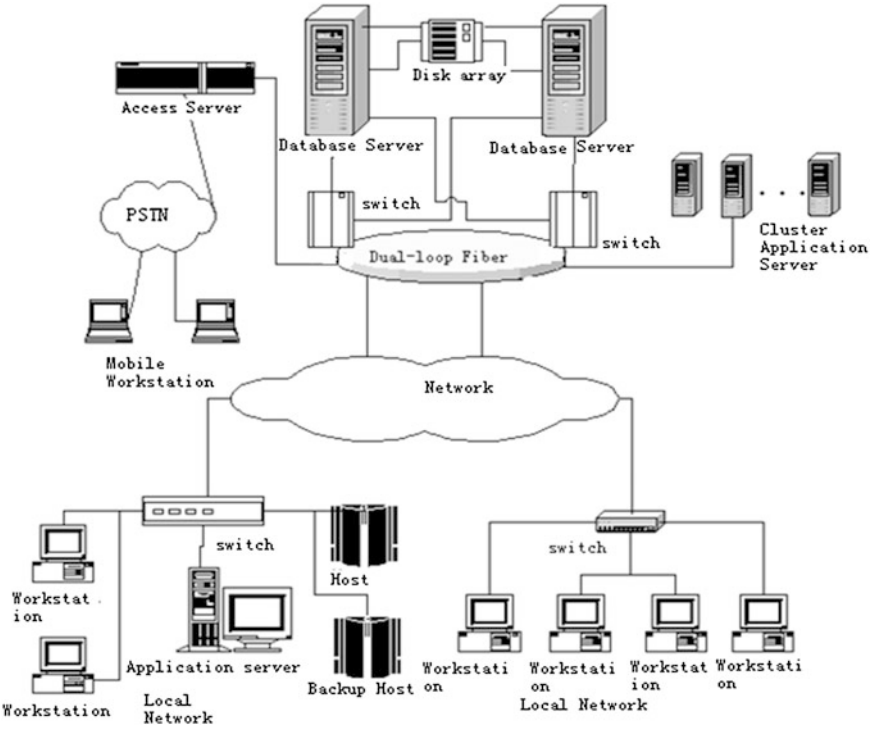


Fig. 127.1 System hardware architecture

127.4 Functional Design

127.4.1 Communication Resource Management

There are a large number of types of communication resources in the Nuclear Power Plant. The communication resources are divided into four categories: line resources, equipment resources, equipment connected relationship, equipment spare.

(a) Line Resource Management

The line resource management includes the PBX port, cable, transfer boxes, connecting cables, junction box, extension installed location and extension information point [5]. The line resource management deals with this information. With this information, users can check each usage and query system information. The system information is the total number and the using of the number and the idle quantity and the number of faults.

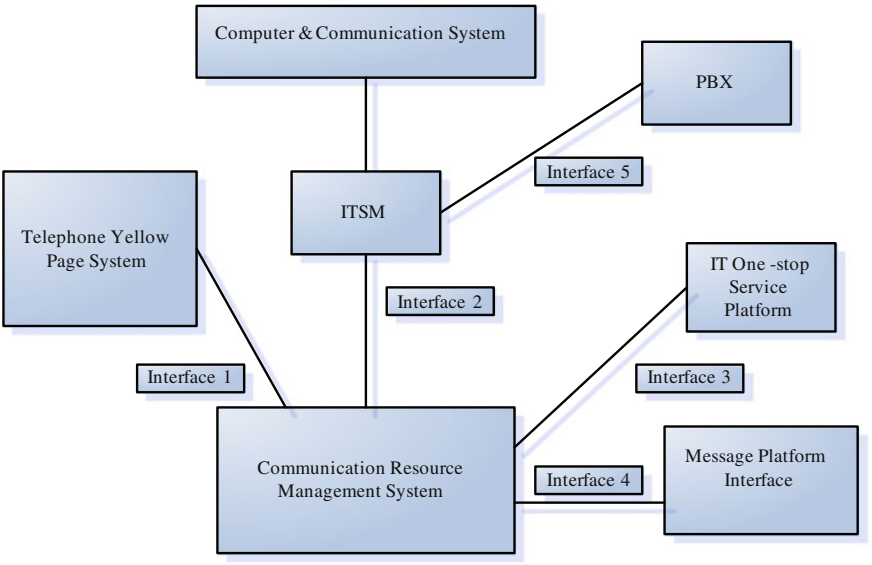


Fig. 127.2 External interfaces

(b) Equipment Manager

The main equipment information includes the equipment name, the equipment ID, the model, the specifications and the installation site. The equipment can be easily collected statistics through the information. The equipment resources include equipment management, equipment maintenance and document management.

(c) Equipment Connection

Communication equipment uses a complicated communication system to link all equipment in the communication resource management system. The system can use Visio system to display relationship among the equipment, through which graphical data and list data can be transformed.

(d) Spare Parts Management

There is a range of communications systems in the Nuclear Power Plant. The module mainly complete the database management and timely process the spare parts.

127.4.2 Billing Management

This system uses intelligent caching system. It connects Voice switches and reads the calls data of the voice switch. The billing data processing computer receives the processing data. The system can achieve real-time billing. Billing management includes billing parameter management, billing records management, billing

inquiries, billing data table, billing statistics and other functions. The billing system adopts calls data of different departments and different permissions of the user.

127.4.3 Work Order Management

At first, the computer communication system (CCS) receives the request from the user. The request, which is installed, dismantled and moved by telephone. The event orders are repaired and the fax repaired by the telephone. The communication resource management system class work orders through ITSM communications. Then, these work orders convert into the resource management system. When the work orders are processed, they will automatically update to the ITSM [6, 7]. The structure of work order management is shown in Fig. 127.3:

127.4.4 Decision-Making Management

Through the communication resource management system, it will provide historical data. The historical data can be analyzed accurately. Users can grasp the equipment

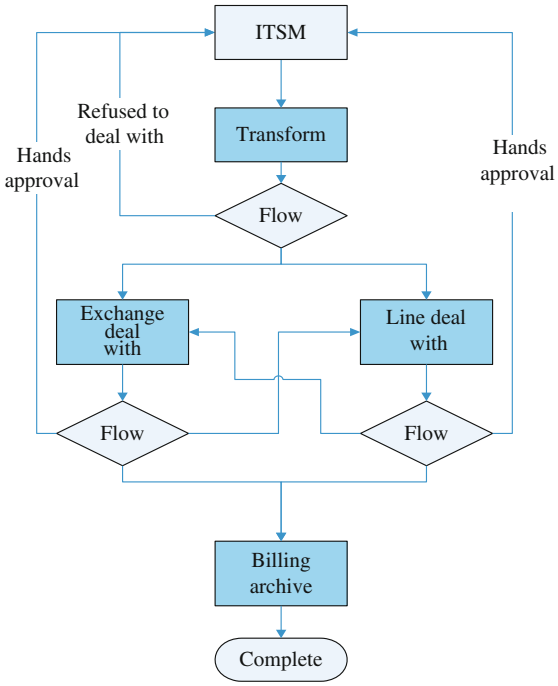


Fig. 127.3 Work Order Management

resource information. The information includes the total system capacity, free numbers and damaged numbers. When users know the equipment resource information, they can protect the normal operation of the system, and provide the spare parts for installation and adjustment.

127.5 Key Technologies and Innovation

127.5.1 Operation Processes

In the process design of the system, each work order is independent. The process uses the state machine design pattern in software design. It can achieve high flexibility and high scalability for the system. All work orders are cycling through the corresponding "state" in the system. The system will automatically track work orders state to place in a new work order status, until the work order is completed. This design pattern can ensure the correctness and greatly improve the scalability of the system. When the work order processes are changed, users can achieve the change of the process [8, 9].

127.5.2 C/S and B/S Two Modes

The Client Server(C/S) and Web Client Server (B/S) are used for different users. In general, the system administrator uses C/S mode. Users install specific client to login in. When users login in, they need to go through a stringent certification process and use a graphical interface to complete all operations. For an ordinary user, the system provides the B/S model for management. The user does not need to install the client, who can use a native browser to login in [10, 11].

127.6 Conclusion

The communication resource management system uses computer technology, network technology and database technology. The system creates an intelligent information management platform and achieves centralized management equipment for the Nuclear Power Plant. The system architecture of communication resource management employs hierarchical structure. It can create platform optimization and high efficiency. The system implements these functions such as equipment information input, query, modification, and user management. It improves the management level of the equipment and relieves the manager's burden.

References

1. Wen, Y.: Software architecture design. Posts and Telecom Press, Beijing (2008)
2. Zheng, Yu Shan, Zheng, Zheng Hong: Database applications. Comp. Eng. Des. **29**(22), 5722–5728 (2008)
3. Yang, Dong Qing, Tang, Shi Wei: Introduction to database systems. Machinery Industry Press, (2000)
4. Zhou, ZhuLi, Feng, Jia Hua, Meng, Xiao Feng: SQL Server database principle. Tsinghua University Press, (2004)
5. Ye, M.: Program-controlled digital exchange and exchange networks. Beijing University of Posts and Telecommunications Press, Beijing (2004)
6. Bjarne, S.: The C++ programming language. Higher Education Press Pearson Education, Beijing (2003)
7. Wen, Y.C., Song, H.Z.: The C++ programming language tutorial. In: Xi, A. (ed.) University of Electronic Science and Technology press (2004)
8. Sun, Chang Ai, Jin, Mao Zhong, Liu, Chao: Software architecture research. J. Softw. **13**(7), 1228–1237 (2002)
9. Gong, W.H., Wang, C.G., Yu H.J.: Based on the MVC pattern and. NET internal management information system design. Comp. Eng. Des. **28**(5), 2142–2144 (2007)
10. Ming, Y.: Visual C++ programming. Ocean Press, Beijing (2001)
11. Ren, J.Z..: Software engineering. Tsinghua University Press, Beijing (1999)

Chapter 128
Institutional Factors Analysis of Listed Company's Equity Financing Preference: Based on the Latest Data of Listed Company in Manufactory Industry

Jianqiang Guo, Hang Zhang and Hongna Wang

Abstract In order to analyze the influence of institutional factors on listed company's equity financing preference, this paper collects the data of manufacturing companies during 2003–2010 in China's A-share market, which has begun to take shape after several years of rapid development. The paper mainly analyzes the role of control properties, equity structure and price earnings ratio in the company's financing preference. The statistical regression results show that different control property will result in different financing preferences of the listed companies. And only equity structure and price earnings ratio are combined with control property. They would play a role in company's financial structure. It argues that this is mainly because of the main body of the listed company is restructured state-owned enterprises, which are controlled by state capital. So the choice of financing mode reflects the will of the managers. This paper provides important reference to the whole stock market and capital market.

Keywords Financing preference · Control property · Equity structure

128.1 Introduction

The financing structure of listed companies in developed countries generally follows the pecking-order theory of Myers and Majluf (1984). A firm's financing order is retained earnings, debt and equity. Based on the study of Chinese listed companies' financing preference, scholars found that Chinese listed companies prefer equity financing rather than debt financing. The formation of the equity financing preference of listed companies in China is strongly linked to institutional

J. Guo (✉) · H. Zhang · H. Wang
School of Business, Shandong University, Weihai, China
e-mail: gjq939@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_128,

constraints in our country including the development of securities market and listed companies' characteristics and so on. Therefore, this paper mainly analyses the reasons of equity financing preference of listed companies in China from the point of institutional factors.

Under the study of papers on financing behavior of the listed companies and capital structure, the author found that some scholars had analyzed the formation mechanism of listed companies' financing behavior and done some empirical analysis from the institutional factors. But the previous research has ignored the systematic analysis of institutional factors. This paper will attempt to overcome this defect. From the view of institutional factors on China's special background, the research and its conclusions have directive meanings to the whole stock market and the efficiency of resources allocation of capital market.

128.2 Literature Review

Some foreign scholars pointed out that the main factor which influences the choice of corporate finance is the enterprise features including firm's size, growth, value of asset-backed, profitability, non-debt tax shields and others [1]. While through the study of financing behaviors in developing countries, other scholars paid attention to the institutional factors gradually. Rejan and Zingales pointed out that institutional differences lead to differences between countries companies' financing structure [2]. Sing (1992–1998) stated briefly that the pecking order of the listed companies in developing countries is significantly different from the developed countries due to the particularity of the degree of market and financial system in developing countries [3]. In short, the order of financing is equity financing, debt financing and internal financing in developing countries.

Domestic scholars who do research on financing preferences of Chinese listed companies mainly focus on internal characteristics, the cost of financing, equity structure, corporate governance structure and other factors according to Chinese unique conditions. Based on the empirical analysis of Lu Zhengfei and Xin Yu, the corporate capital structure and corporate profitability have negative correlation, and the corporate capital structure is not related to the firm size, the value of asset-backed and growth factors [4]. However, Hong Xixi and Yi Feng got the opposite conclusion: asset-liability ratio has a significant positive correlation with size and profitability of company [5]. Lv Changjiang and Han Huibo concluded that corporate debt ratio was negatively correlated with the profitability, current ratio, and fixed assets ratio, and was positively correlated with the size and growth of the company [6].

Before 2005, our stock market was split with outstanding shares and non-tradable shares. The company of the corporate shares was more prefer to equity financing. Wu Jiang and Ruan Tong pointed out that the company was more inclined to equity financing as the dual equity structure made the cost of equity financing less than the cost of debt financing and the non-tradable shareholders can

obtain financing return from the equity financing [7]. Xiao Zuoping and Hong Zheng explicitly pointed out that the equity structure was an important factor in the company's capital structure [8].

Huang Shaoan and Zhang Gang pointed out that there were internal control problem in the listed companies [9]. Managers will choose equity financing in order to obtain control benefits and avoid the risk of bankruptcy. And because of the proportion of managerial ownership is very low or even zero, managers choose equity financing would not affect their control, and would protect them from the "hard constraints" of liabilities [10].

Recently, scholars are interested in the institutional factors more and more. This paper will provide a systematic analysis of the equity financing preference from the view of institutional factors on China's special background.

128.3 Research Design

128.3.1 Variables Set

This paper selects debt-to-long capital ratio as financing preference proxy variable, or explained variable (represented by Debt). In addition, we apply institutional factors proxy variable as explanatory variable. The following variables are mainly included: (1) PE ratio (represented by Pe). Despite the low PE ratio is worth to be invested for people, however, from the shareholders and managers' perspectives, high PE ratio is more suited to their interests. Therefore, this paper forecasts PE ratio and debt-to-long capital exert negative correlation. (2) Ownership concentration and level of managerial share ownership. Governance structure of our listed companies is not perfect, mainly reflected by serious phenomenon that one shareholder or a few shareholders control stakes. This paper forecasts the first large proportion of shareholding and debt-to-long capital exerts negative correlation. From the view of benefit maximization, low level managerial share ownership (represented by Manage) leads company manager to prefer equity financing. So this paper forecasts the coefficient is positive. (3) Actual controller property (represented by Contrl). This paper introduces the dummy variables of actual controller property. 1 stands for state-owned capital holding, while 0 stands for other capital holdings. This paper cannot forecast symbol, only data can illustrate the problem. Finally, according to the research conclusions of domestic and foreign scholars, we select five control variables to regulate the impact of other factors on equity financing preference in empirical research, they are company size (represented by Size), profit rate to net worth (represented by ROA), tangible asset rate (represented by Tang), increase rate of business revenue (represented by Gr_oi), actual income tax rate (represented by Tax). The paper forecasts there is positive correlation between debt-to-long capital ratio and actual income tax rate, debt-to-long capital ratio and tangible asset rate. As for other variables, only data can illustrate the problem.

128.3.2 Model Design and Research Hypothesis

(1) Model Design. The purpose of this study is to investigate whether there is any influence exists between actual controller properties of listed company and finance preference. So we constructed multiple regression model, to facilitate comparison, we introduce two cross variables, they are actual controller properties and equity structure, actual controller properties and PE ratio. (model shown below)

$$\text{Debt} = a + \beta 1\text{Pe} + \beta 2\text{Manage} + \beta 3\text{Centrl} + \beta 4\text{Size} + \beta 5\text{ROA} + \beta 6\text{Tang} + \beta 7\text{Tax} + \beta 8\text{Contrl} + \beta 9\text{Gr\_o} + \beta 10\text{Pe*Contr} + \beta 11\text{Manage*Contrl} + \beta 12\text{Centrl*Contrl} + u$$

In the model, a represents a constant, u is a random variable.

(2) Research hypothesis. We propose the following null hypothesis and alternative hypothesis.

H1a Financing preference of listed companies controlled by different nature capital don't exist significant difference

H1b Financing preference of listed companies controlled by different nature capital exist significant difference

H2a Equity structure of listed companies controlled by different nature capital don't exist significant difference

H2b Equity structure of listed companies controlled by different nature capital exist significant difference

H3a Companies' PE ratio of listed companies controlled by different nature capital don't exist significant difference

H3b Companies' PE ratio of listed companies controlled by different nature capital exist significant difference

H4a Financing preference of listed companies and actual controller properties, equity structure PE ratio have no correlativity

H4b Financing preference of listed companies and actual controller properties, equity structure PE ratio have correlativity.

128.3.3 Data Selection

We select listed company in manufactory industry as the initial sample, 2003–2010, a total of 8 years of the company's data for research, which is mainly from the GTA database.

Sample selection is guided by the following principles: (1) the company must be listed before 2003 and continuing operations until the end of 2010, (2) Exclude ST and PT class companies and incomplete data.

Table 128.1 Variable descriptive statistical of different controller

	Variables	Mean		Standard deviation	
		1	0	1	0
Explained variable	Debt	0.130215	0.109101	0.150054	0.131468
Explanatory variables	Pe	53.46059	52.01222	188.3963	164.254
	Manage	0.087698	0.000426	2.171436	0.001928
	Contrl	44.9547	34.93484	15.47918	14.01239
Control variables	Size	21.86222	21.82674	1.330735	0.977027
	ROA	0.053211	0.059306	0.060447	0.048345
	Tang	0.970623	0.957912	0.039234	0.040834
	Tax	−0.11684	0.999909	6.089479	13.57966
	Go_oi	0.29344	0.21714	1.071168	0.414061

128.4 Empirical Research and Analysis

128.4.1 Variable Descriptive Statistical Analysis

This paper use EVIEWS 5.0 software to do empirical analysis. From Table 128.1, we use the results of variable descriptive statistical under different controller to illustrate the above hypothesis, including mean and standard deviation, it can be concluded:

From the numerical comparison, debt-to-long capital ratio of listed companies controlled by state-owned capital is higher than listed companies controlled by non-state-owned capital. Financing preference of listed companies controlled by different nature capital exist significant difference supporting the hypothesis 1b.

From the number of the mean and standard deviation, equity concentration of listed companies controlled by state-owned capital is higher than listed companies controlled by non-state-owned capital. Managerial ownership is higher than the latter. So equity structure of listed companies controlled by different nature capitals exist significant difference. Support the hypothesis 2b.

Meanwhile, PE ratio of listed companies controlled by state-owned capital is higher than listed companies controlled by non-state-owned capital. Companies' PE ratio of listed companies controlled by different nature capital exist significant difference supporting the hypothesis 3b. As for other variables, results of multiple linear regression can give better answers.

128.4.2 Multiple Linear Regression Results and Analysis

Through validating, the selected data in this paper does not have heteroscedasticity and self-correlation.

Table 128.2 The results of least square estimation

Variables	Coefficiency	Standard deviation	T-value
Pe	3.85E-05	4.13E-05	0.932295
Manage	−2.291235	3.512028	−0.6524
Centrl	0.000225	0.000484	0.465396
Size	0.057566	0.004271	13.47693
ROA	−0.835113	0.088755	−9.4092
Tang	−0.163824	0.118507	−1.3824
Tax	0.000257	0.000429	0.597635
Contrl	0.004881	0.024675	0.197794
Gr_oi	0.027199	0.010274	2.647469
Contrl*Pe	−2.84E-05	4.94E-05	−0.57615
Contrl*Manage	18.18025	9.113277	1.994919

Table 128.2 shows the model of multiple linear regression results, debt-to-long capital ratio of listed companies and actual controller properties, equity structure PE ratio have correlativity. So support the hypothesis 4b.

According to the results, we can analyze the influencing mechanism between explained and explanatory variables. The coefficient of actual controller properties is positive, which means relying on the relationship between government and bank; listed companies controlled by state-owned capital are easy to achieve long-term loans from the bank. The coefficient of level of managerial share ownership is negative; the conclusion seems to be explained by the study of Jensen and Meckling (1976). However, the explanation does not apply to listed companies in China, the conclusion that based on perfect corporate governance. Furthermore, the coefficient of level of managerial share ownership is negative, which contrary to the paper forecasts. But this does not mean the research is wrong, because in the analysis, the paper has made it clear to the premise that listed companies controlled by state-owned capital. The coefficient of cross variable between actual controller properties and equity structure is positive in the model, proving the analysis of paper. Listed companies controlled by state-owned capital, the lower the level of managerial share ownership, the lower of debt-to-long capital ratio and the listed companies preference for equity financing.

The coefficient of the first large proportion of shareholding is positive, which contrary to the paper's analysis and forecast. The reason is as same as the level of managerial share ownership does not match the forecast. The coefficient of cross variable between actual controller properties and ownership concentration is positive, which is contrary to the paper forecasts. It can also uses the above reasons to explain it. The coefficient of PE ratio is positive, which is contrary to the paper forecasts. This may be because the higher the PE ratio, the company's share price is overvalued, which can cause the ingredients of bubbles. The value of investment decreases compared to the companies with low PE ratio. So this will affect the company's equity financing effect. The coefficient of cross variable between actual controller properties and PE ratio is negative, which contrary to the paper forecasts.

In conclusion, if the actual controller property were different in the listed company, neither would be their financing preference. The influence of equity structure and PE ratio on the listed company's preference financing should be combined with the actual controller's nature. For the control variables in the model, there is a significant positive correlation between debt-to-long capital ratio and actual income tax rate, increasing rate of business revenue and company size. However, there is a negative correlation between debt-to-long capital ratio and tangible asset rate, profit rate to net worth.

128.5 Conclusion

To sum up, the conclusions of this paper have important theoretical and practical significance. Firstly, financing behavior of listed companies in China has become an important part of the domestic capital market, so their financing preferences will have obvious and far-reaching impacts on the whole stock market and the efficiency of resources allocation of capital market. Secondly, the paper analyses financing preference's mechanism and influencing factors of listed companies in China, which has important theoretical and practical guidance that is significant to improve the efficiency of China's capital market financing.

References

1. Titman, Sheridan, Wessels, Roberto: The determination of capital structure choice. J. Finance **43**(3), 17–40 (1988)
2. Rajan, Zingales: What do we know about capital structure? some evidence from international data. J. Finance **6**(6), 65–76 (1995)
3. Sing, A., Hamid J.: Corporate financial structures in developing countries. Finance Corp. Tech. **1**, 4–7 (1992)
4. Lu, Z., Yu, X.: Empirical analysis of the determinants of capital structure. Account Res **8**, 131–138 (1998)
5. Hong, X., Shen, Y.F.: Empirical analysis of the impact of the capital structure of listed companies in China. J. Xiamen Univ. **3**(2), 172–197 (2000)
6. Lv, C., Wang, K.: Empirical analysis of the dividend policy of listed companies. Econ. Res. J. **12**(8), 121–153 (1999)
7. Jiang, Wu, Tong, Ruan: Equity trading structure and financing behavior of listed companies in China. J. Finan. Res. **6**, 35–41 (2004)
8. Xiao, Z., Zou, H.: The determinants of capital structure and equity financing preference in listed Chinese companies. Econ. Res. J. **6**, 110–114 (2008)
9. Huang, S., Zhang, G.: Equity financing preference of listed companies in China. Econ. Res. J. **11**, 89–96 (2001)
10. Bin, Liu, Weiming, Yi: Capital structure and financing preferences of China's listed companies in the post-crisis Era. Enterp. Econ. **7**(2), 48–60 (2011)

Chapter 129
Empirical Analysis of Positive Feedback Trading in Chinese Stock Market

Jianqiang Guo, Qian Yang and Qi Li

Abstract The paper aims to research positive feedback trading behaviour and its effect on the market fluctuations in Chinese stock market, and the relation between the level of market fluctuations and autocorrelation coefficient of market yields. It uses the data of daily close price of Shenzhen Component from August 22, 1996 to May 16, 2012 and finds severe boom and slump in Shenzhen stock market, notable non-normal distribution and ARCH effect. By combining the positive feedback trading model and EGARCH Model, the ARCH Regression result suggests: (1) there exists positive feedback traders in Chinese stock market who makes market returns fluctuate severely; (2) the relation is negative between the market fluctuation and autocorrelation coefficient of market yields; (3) the asymmetry and leverage effect will make Chinese stock market have greater fluctuation during descendant periods than rising periods.

Keywords Positive feedback trading · Chinese stock market · Market fluctuations · Autocorrelation coefficient · Leverage effect leverage effect

129.1 Introduction

In stock market, positive feedback trading means buying one stock as the price rises and selling one stock as the price drops. Due to Herd Behavior and the portfolio strategy using extrapolating expected method, large numbers of positive feedback traders exist in stock market. Their trading behavior can aggravate the fluctuation of securities' prices, increase the deviation degree between price and value, and even cause systemic risk in stock market. Thus, it is of great theoretic

J. Guo (✉) · Q. Yang · Q. Li
School of Business, Shandong University, Weihai, China
e-mail: gjq939@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_129,

meaning to research positive feedback trading in stock market. As we know, Chinese stock market is a new capital market with a short history. It has some features different from the mature markets. For example, investors' immature cognizance results in irrational investment and behavior bias. In Chinese stock market, nevertheless, it is still unclear whether there is positive feedback trading, an important form of irrational trading strategy. Therefore, on the empirical analysis of Chinese stock market, this paper examines the existence of positive feedback trading in this market and analyzes its influence to stock prices. Moreover, the author attempts to answer the following questions which are of enormous controversy in current academia: what is the relation between market fluctuation and autocorrelation coefficient of stock yield rate, and whether there exists asymmetry and leverage effect in Chinese stock market.

129.2 Literature Review

The research on positive feedback trading is based on the evolution of noise trading model. De Long et al. (1990) propose the noise trading model DSSW1. It points out that the investors in stock market can be divided into two classes: rational investors and noise traders. Noise traders can be affected by "noise trader emotions" and the mood is systematic and unpredictable [1]. This forms sustainable system risk in stock market and makes the deviation between stock price and its value. Moreover, because the rational arbitragers are risk averse and thus tend to reduce trading their securities, the deviation between price and value will persist. De Long et al. (1991) comes up with a new noise trading model DSSW2 that regards noise traders as a whole and claims that they can exist for a long time in stock market [2]. Afterwards, De long et al. put forward the positive feedback trading model, proving that the noise traders in stock market may use positive feedback trading strategy to invest—buy one stock when its price rises and sell one stock when its price falls. There are abundant empirical analyses on positive feedback trading behavior in stock market. Sentana and Wadhwanix (1992) provide evidence that the daily return and hourly return from stock market have positive (negative) autocorrelation when the fluctuation of stock price is low (high) [3]. Odean (1999) find investors tend to buy shares with extreme realization and if a stock can cash best or worst in the past, investors are more willing to buy it [4]. Ozdenoren and Yuan (2008) conclude that strong positive feedback trading can lead to excessive volatility of the stock market from the perspective that positive feedback trading affects asset prices and asset prices affect companies' cash flow [5].

The research on positive feedback trading starts later in China as Chinese stock market is in primary stage. The previous papers have used foreign theoretical models to test the existence of positive feedback trading in China and analyze the relationship between fluctuation of stock price and that. For example, Tang Yu et al. (2001) analyze the link between the positive feedback trading rules and the

autocorrelation of the Shanghai composite index's daily earnings. The results show that the positive feedback trading will lead to a negative autocorrelation of returns and the absolute value of correlation coefficient will be large with the fluctuation increasing [6]. Li Shaoping and Gu Guangcai (2007) also reach a similar conclusion by establishing an asymmetric EGARCH model [7]. Zhang Enzhong (2009) selects all the daily earnings data of Shanghai composite index from the establishment date of Shanghai Stock Exchange (December 19, 1990) to the date emerging highest point of 6124 (October 17, 2007). By using the GARCH model, the results show that positive feedback trading behavior does exist in the Shanghai stock market, but not very serious [8]. Luo Ying and He Xiaofeng (2005) conclude that the amplification effect of positive feedback trading deviates securities prices from their value and makes securities prices are very unstable [9]. Moreover, Su Yanli and Zhuang Xintian (2008), Chen Zhuosi et al. (2008) and Wang Meijin (2005) respectively take the view of securities investment fund behavior, institutional investors' behavior and investors' disposal effect to research the institutional investors' trading. They find that positive feedback trading behavior does exist in Chinese stock market and increases the volatility of stock price.

From the literature review we can see, China's research mainly uses the foreign positive feedback trading model for reference and the sample interval is generally short. Most papers just analyze positive feedback traders' behavior for a period time not for a long-term trading time. Thus, this paper will select a longer series data to analyze the three problems: (1) the existence of positive feedback trading behavior in Chinese stock market; (2) the relationship between the autocorrelation of market yields and the volatility of stock prices; (3) the existence of asymmetry and leverage effect in Chinese stock market under the premise of (1).

129.3 Research Design

129.3.1 Empirical Model

Positive feedback traders make investment decisions in line with stock's previous performances, then the present stock yield is related with the previous one and the autocorrelation will exist. Sentana and Wadhwani (1992) proposes a positive feedback strategic model, presuming that there are two kinds of investment traders in the stock market—positive feedback traders and rational traders. Positive feedback traders make the current investment decision using stock's previous yields, while rational investors decide their present investment by constructing a risk–benefit model including stock's expected return and stock market fluctuations. Empirical results from this model suggest that, when positive feedback traders and rational investors are both in the stock market, the autocorrelation exists in stock yields and will become more significant as the market fluctuates. This paper quotes the model and the demand function of positive feedback traders is as follows:

$$S_{1,t} = \gamma R_{t-1} \tag{129.1}$$

In this equation, $S_{1,t}$ is the ratio of positive feedback traders' demand accounting for the total market demand during t period. R_{t-1} is the stock's previous yield.$\gamma$is to distinguish the type of feedback traders, when $\gamma > 0$, the major feedback traders in stock market are positive ones, and the demand increases when the previous stock price rises and decreases when the previous stock price falls; when $\gamma < 0$, the major feedback traders are negative ones, changes in the demand are opposite with positive feedback traders. The demand function of positive feedback traders is as following:

$$S_{2,t} = [E_{t-1}(R_t) - \alpha]/u_t \tag{129.2}$$

$E_{t-1}(R_t)$ is the expected stock return of rational investors, who can make reasonable expectations, meaning: $R_t = E_{t-1}(R_t) + \varepsilon_t$; u_t is stock risk parameter related with stock fluctuations, σ^2 is a linear function of stock's conditional variance: $[u_t = u_2(\sigma^2)]$; α is risk-free rate, the demand of rational investors is positively correlated with stock's expected return, and negatively correlated with stock's risk premium. When the market is in equilibrium:

$$S_{1,t} + S_{2,t} = 1 \tag{129.3}$$

Taking the Eqs. (129.1) and (129.2) into the Eq. (129.3), the result is:

$$E_{t-1}(R_t) = \alpha + u_t - \gamma u_t R_{t-1} \tag{129.4}$$

$$R_t = \alpha + u_t - \gamma u_t R_{t-1} + \varepsilon_t \tag{129.5}$$

The above equations suggest that the stock return is related with risk-free rate, risk premium, previous stock yield and the type of traders. When the feedback traders are positive ones $(\gamma > 0,)$ then the autocorrelation coefficient $-\gamma u_t < 0$, meaning that stock yields are positively self-correlated; when the feedback traders are negative ones $(\gamma < 0), -\gamma u_t > 0$, meaning that the feedback traders are positive ones. Moreover, to describe the relationship of market fluctuations and positive feedback trading, Sentana & Wadhwani (1992) substitute the linear expression of u_t for the coefficient of R_{t-1}:

$$R_t = \alpha + u\left(\sigma_t^2\right) - \left(\gamma_0 + \gamma_1 \sigma_t^2\right) R_{t-1} + \varepsilon_t. \tag{129.6}$$

The Eq. (129.6) shows that when the market fluctuations are low (the value of σ_t^2 is small), then self-correlation coefficient is decided by γ_0,, when $\gamma_0 < 0 < \gamma_1$, meaning negative feedback traders are dominant, market yields are positive autocorrelation. But as the fluctuations strengthen, the autocorrelation coefficient is influenced by γ_1, then positive feedback traders are dominant and market yields are negative self-correlation.

129.3.2 EGARCH Model and Data Selection

While analyzing financial time series, the autocorrelation of series data is usual. Some researchers (Kraft,1983) find that, the stability of disturbance variance in the time series model is weaker than the hypothesis—heteroscedasticity. ARCH (autoregressive conditional heteroscedasticity) models are supposed to describe the variance correlation. This paper adopts EGARCH model (Nelson, 1991), which can test the asymmetry of fluctuations and leverage effect. The supposed model is as follows:

$$log\left(\sigma_t^2\right) = \omega + \beta \log\left(\sigma_{t-1}^2\right) + \theta\left|\frac{\upsilon_{t-1}}{\sigma_{t-1}}\right| + \delta\frac{\upsilon_{t-1}}{\sigma_{t-1}}. \tag{129.7}$$

The left is logarithm of conditional variance and its estimated value can't be negative. If $\delta \neq 0$, then conditional variance is asymmetry, the leverage effect can be tested by the hypothesis $\delta < 0$, meaning that bad news can induce lager fluctuations than equal amount of good news.

This paper chose daily yield rate of Shenzhen Component as the sample data, the sample period is from August 22, 1996 to May 16, 2012 and its number is 3,795. The data sources is Yahoo Finance (http://finance. yahoo.com/). And the daily yield rate is calculated by:

$$R_t = 100 * (\ln P_t - \ln P_{t-1}). \tag{129.8}$$

Here, R_t is the daily yield rate of Shenzhen Component in t period, P_t is the daily close price of Shenzhen Component in t period.

129.4 Empirical Analysis

129.4.1 Variables Descriptive Statistics

This paper uses EVIEWS 5.0 software to do empirical analysis. The descriptive statistics of Shenzhen Component daily yield rate during the sample period is shown as Fig. 129.1, suggesting the series distribution characteristics. The mean is 0.035705, the standard deviation is 1.921179, and the skewness is −0.296313(<0), meaning that the shape of the series distribution is left skewness compared with normal distribution. The kurtosis is 6.560051 higher than that of normal distribution whose kurtosis is 3, meaning that the distribution of yield rate series is saber and fat-tail. Jarque–Bera statistics is 2059.062, and P value is 0.000000, implying that the distribution of yield rate series is notably non-normal distribution. The statistic results suggest that there are sharp boom and bust in Shenzhen stock market, the times of below the average of yield rate are more, and the distribution is notably non-normal.

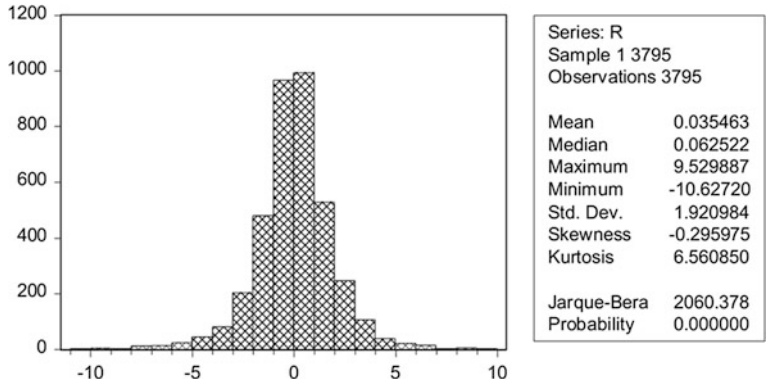


Fig. 129.1 Time series distribution

Moreover, to test whether the sample time series is stational, this paper made ADF stationary test and the t statistics value is −25.63828, P value is 0.000000, suggesting that the series is stational. Furthermore, this paper tests the autocorrelation and partial autocorrelation of the series, and the result suggests that the series is significantly autocorrelated at 5 % significant level, showing ARCH effect, meaning that the market fluctuations are influenced by the history ones and there exists concentration of fluctuations.

129.4.2 Empirical Model Regression Results and Analysis

The result is shown in Table 129.1 after ARCH Regression for Eqs. (129.6) and (129.7). It suggests that, α is significant in statistical sense, suggesting that there are rational traders in Chinese stock market; $\gamma_0 < 0, \gamma_1 > 0$, meaning that there are positive feedback traders in Chines stock market and they can largely influence market fluctuations.(The coefficient of estimated γ_1 is not significant, because this paper doesn't distinguish the rising and falling period of the market, and the existing research suggests that positive feedback trading behaviors in both periods are different.) When the level of market fluctuations is low (the value of σ^2 is small), the autocorrelation coefficient is more than zero $\left(\left(\gamma_0 + \gamma_1 \sigma_t^2\right) > 0\right)$ and the market yield rate is positively autocorrelated. However, when the level of market fluctuation strengthens (the value of σ^2 gradually increases), the autocorrelation coefficient is largely influenced by γ_1 and the negative feedback traders become dominant in the market, the yield rate is negative autocorrelation. In a word, the market fluctuation is negative correlated with the autocorrelation coefficient of the yield rate.

Table 129.1 Arch regression result

Parameter	Estimated value	Z statistics	P value
α	−1.886106 (0.015453)	−122.0539	0.0000[a]
γ_0	−0.084938 (0.032225)	−2.635821	0.0084[a]
γ_1	0.018848 (0.014286)	1.319299	0.1871
ω	−0.699809 (0.062947)	−11.11738	0.0000[a]
β	0.874066 (0.874066)	46.02965	0.0000[a]
θ	0.827953 (0.071566)	11.56905	0.0001[a]
δ	−0.140710 (0.034699)	−4.055120	0.0000[a]

[a] indicates statistical significance at the 1 % lever

The result shows that $\delta \neq 0$, it means that the market fluctuations is asymmetric; $\delta < 0$ confirms the leverage effect which suggests that the equal amount of bad news can induce lager scale market fluctuations than good news.

129.5 Conclusion

This paper uses daily close price data of Shenzhen Component from August 22, 1996 to May 16, 2012. The descriptive statistics of the time series of daily yield rate and the autocorrelation and partial autocorrelation test of the series confirm sharp boom and bust in Shenzhen stock market. The distribution is notably non-normal and the series has ARCH effect due to more frequencies below the average yield rate. By combining the positive feedback trading model of Sentana and Wadhwani (1992) and EGACH Model and using EVIEWS 5.0, this paper finds that there are both rational traders and feedback traders in Chinese stock market, and the latter largely influence the stock market fluctuations. Besides, negative relationship between the market fluctuation and autocorrelation coefficient of stock yield rate is also found. Finally, this paper points out that the asymmetry and leverage effect will make Chinese stock market experience greater fluctuation during descendant periods than rising periods.

The conclusion is of both theoretical and practical significance for the Chinese investors and government. Firstly, there are many irrational factors in Chinese stock market and the irrational traders don't make investment decisions according to the truly initial value of stocks. They make decisions by following the irrational strategies of buying when rising and selling when falling. The behaviors will aggravate the stock market fluctuations and reduce the market effectiveness. This finding is opposite to the conclusion of the efficient market hypothesis. Secondly, the behaviors of irrational traders may bring about stock bubbles, largely damage Chinese virtual and real economy and heavily influence the development of stock market. The government and relevant agencies should pay more attention to it. Thirdly, there are many reasons causing the behaviors, such as the current institution of the stock market, the information disclosure mechanism, the quality of

the listed companies and so on. These are important factors that can't be ignored. Future research on positive feedback trading may pay attention to how to control these behaviors.

References

1. De Long, J.B., Shleifer, A., Summers, L.H., Waldmann, R.J.: Noise trade risk in financial markets[J]. J. Political Econ. **98**(4), 703–738 (1990)
2. De Long, J.B., Shleifer, A., Summers, L.H., Waldmann, R.J.: The survival of noise traders in financial market [J]. J. Bus. **64** 1–19 (1991)
3. Sentana, E., Wahwani, S.: Feedback traders and stock return autocorrelations: evidence from a century of daily data [J]. Econ. J. **104**, 415–425 (1992)
4. Odean, T.: Do Investors trade too much? . Am. Econ. Rev. **89**, 1279–1298 (1999)
5. Ozdenoren, E., Yuan, K.: Feedback effects and asset prices. J. Finance **63**, 1939–1975 (2008)
6. Yu, T., Yong, Z., Xiaowo, T.: Empirical analysis of the positive feedback trading and the self-correlation of stock index returns. Acad. J. Electron. Sci. Technol. Univ. **30**(3), 300–303 (2001)
7. Shaoping, Li, Guangcai, Gu: Empirical research of positive feedback trading in Chinese stock market. Syst. Eng. **9**, 111–115 (2007)
8. Enzhong, Zhang: Effects of positive feedback trading on securities market: evidence from shanghai security market. Shandong Soc. Sci. **10**, 83–86 (2009)
9. Luo, Y., He, X.: The mechanism of positive feedback trading in Chinese stock market. Econ. Surv. 2 (2005)

Chapter 130
Asymmetric Foreign Exchange Rate Exposure of Listed Commercial Banks

Chi Xie and Liangqiu Zhou

Abstract This study investigates the impact of asymmetry foreign exchange risk on listed commercial banks and makes a comparative analysis on differences of foreign exchange rate exposure between introducing asymmetry and not introducing. Firstly, foreign exchange rate exposure of single bank is studied by GARCH model. Moreover, using unbalanced panel model to study the entire banking which includes both the exchange rate variable and the market variable. Empirical results indicate that listed commercial banks and the whole banking industry show different exchange rate exposure, while facing different exchange rate changes. For both single banks and the whole banking, the exchange rate exposure becomes more significant after asymmetry is considered. Exchange rate exposure can be described more accurately after considering asymmetry.

Keywords Exchange rate exposure · Asymmetry · GARCH model · Panel data

130.1 Introduction

The results of many early researches in the foreign exchange risk exposure show insignificant, the possible reason for this phenomenon maybe that researchers do not consider the asymmetric foreign exchange risk exposure, such as Adler and Dumas (1984), Jorion (1990), Griffin and Stulz (2001) and so on [1–3]. Subsequently, a large number of scholars study the foreign exchange risk exposure in different sectors. For example, Luo Hang and Jiang Chun (2007) analyze the

C. Xie (✉) · L. Zhou
College of Business Administration, Hunan University, Changsha, China
e-mail: xiechi@hnu.edu.cn

L. Zhou
e-mail: 839250710@qq.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_130,

foreign exchange risk exposure of Chinese listed companies after the exchange rate reform [4]. Empirical research shows that the companies' foreign exchange risk exposures of the two markets are low. Ding Hui et al. (2008), from the perspective of short-term and long-term, explore the impact of the RMB exchange rate on corporate value [5]. In addition to the financial sector, other sectors' foreign exchange risk exposure shows insignificant. Neglecting asymmetry may result to miscarriage of justice in the foreign exchange exposure.

Later studies have found that most of companies' foreign exchange risk exposure is not significant are actually not true, ignoring the asymmetry of foreign exchange risk exposure that will reduce their risk exposure significance. For example, the Tai (2005), the Koutmos and the Martin (2003), they report that the asymmetry of the foreign exchange risk exposure is mainly due to the company's behaviors [6, 7]. Their researches provide a theoretical and empirical support for the asymmetric foreign exchange risk exposure.

Some researchers use the OLS method or the regression model to study asymmetry of the foreign exchange risk exposure, such as Carter et al. (2003), they find the U.S. multinational foreign exchange risk exposure is nonlinear [8]. Pan and Liu (2012) use regression model to examine the asymmetry of the foreign exchange risk exposure of 20 U.S. manufacturing [9]. The empirical results show that long-term level of foreign exchange risk exposure is more obvious. Overall, these researchers do not consider the conditional variance of returns on assets.

Considering that capital return errors are conditional heteroskedastic, Júnior and José(2012) take the nonlinear STAR model and find the Brazilian non-financial sector's U.S. dollar foreign exchange risk exposure is asymmetry [10]. Muller and Verschoor (2006) find that ignoring the asymmetry causes foreign exchange exposure insignificant [11]. Brooks et al. (2010) show that Australian utilities sector's foreign exchange risk exposure is time-vary asymmetry, the technology sector's foreign exchange risk exposure is asymmetry [12]. Koutmos and Martin (2007) use a vector GARCH model and the dollar index to research asymmetry foreign exchange risk exposure of stock market returns [13]. The above researchers only consider a single exchange rate or exchange rate index on the value of the company, but some ignore foreign exchange risk exposure under the circumstance of different exchange rates.

Therefore, from the asymmetry point of view, this article attempts to study the listed commercial banks' foreign exchange risk exposure under the conditions of different exchange rates, expecting to fill the gaps of the field, more accurately revealing the listed commercial banks' foreign exchange risk exposure.

130.2 Model

Choi, Elyasiani and Kopecky propose a multi-factor model [14]:

$$R_{it}=\beta_{i0}+\beta_{im}R_{m,t}+\beta_{is}\Delta S_t+\varepsilon_{it} \tag{130.1}$$

where R_{it} is stock return of bank i in day t, $R_{m,t}$ is market portfolio index m in the day t, and generally take the market portfolio index returns as control variables to explain the impact of macroeconomic factors on stock returns, ΔS_t is the changes of exchange rates in the day t, ε_{it} is stock return residuals of bank i in day t. β_{im} means the impact of market factors on it, β_{is} is the exchange rate risk exposure coefficient of bank i in the sample during. Considering the foreign exchange risk exposure of asymmetry, Eq. (130.1) can be extended as follow:

$$R_{it}=\beta_{i0}+\beta_{im}R_{m,t}+(\beta_{is}+\beta_{iD}D_{i,t})\Delta S_t+\varepsilon_{it} \tag{130.2}$$

where β_{iD} is the coefficient of asymmetric foreign exchange risk exposure, $D_{i,t}$ is dummy variables, when $\Delta S_t < 0$, $D_{i,t} = 1$, otherwise, $D_{i,t} = 0$.

The presence of conditional heteroskedasticity may lead to dependency, and will result in insignificant parameter estimate and biased test statistics. Therefore, this article will set the residuals ε_t obey the GARCH (1, 1) process:

$$h_{\varepsilon,t}=\alpha_{\varepsilon,0}+\alpha_{\varepsilon,1}\varepsilon_{t-1}^2+\alpha_{\varepsilon,2}h_{\varepsilon,t-1} \tag{130.3}$$

where α_0, $\alpha_{\varepsilon,1}$ and $\alpha_{\varepsilon,2}$ is non-negative parameter. $\alpha_{\varepsilon,1}+\alpha_{\varepsilon,2}$ is the persistence of fluctuations, and $\alpha_{\varepsilon,1}+\alpha_{\varepsilon,2}<1$

130.3 Results and Analysis

Data are selected from 16 listed commercial banks on the Shanghai and Shenzhen stock market. The sample time period of banks that listed before exchange rate system reform was from the date of July 26, 2005 to April 20, 2012, The sample time period of banks that listed after exchange rate system reform is from the date of the listing date until April 20, 2012, sample data are all daily stock return data. Yield data of the market portfolio index is selected from daily yield data of the Shanghai A-share index returns and the Shenzhen A-share index returns, corresponding to the A-share listed companies in Shanghai and Shenzhen A-share listed companies. The sample time period of exchange rate data and market portfolio index data is selected from July 26, 2005 to April 20, 2012, including the RMB against the U.S. dollar, Euro and Japanese yen exchange rate daily data.

Figure 130.1 shows the return of RMB exchange rates, we can see that the fluctuation of RMB against U.S. dollar is the most obvious one, and the fluctuation of other two exchange rates keep relatively stable.

130.3.1 Foreign Exchange Risk Exposure of Single-bank

Because of the difference of size, mechanism, and capital adequacy ratio, every single listed commercial bank's degree of exchange rate risk exposure shows

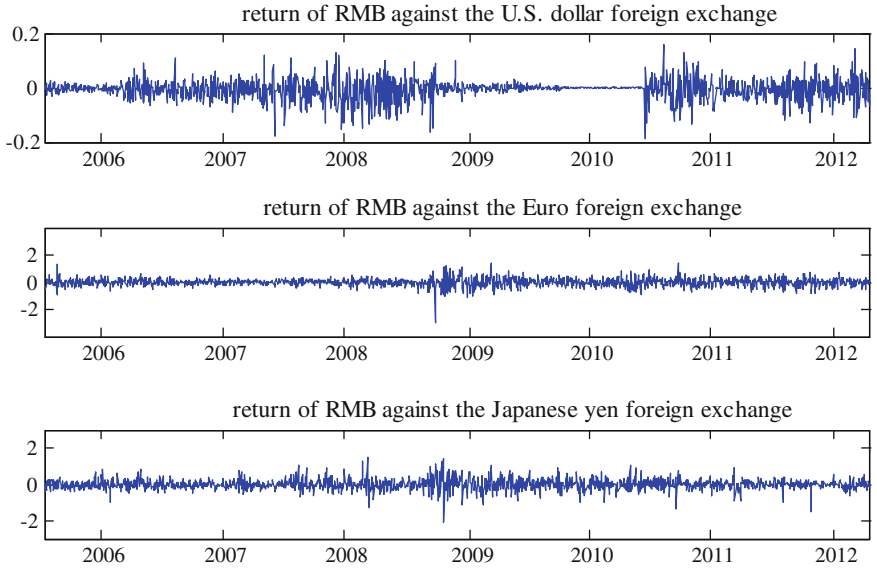


Fig. 130.1 Return of RMB exchange rates

different when facing exchange rate changes. Table 130.1 lists the 14 listed commercial banks' foreign exchange exposure coefficient (excluding two banks that its listed years are short). Table 130.2 shows the coefficient after considering asymmetric foreign exchange risk exposure.

Viewed from the RMB against the U.S. dollar, Euro foreign exchange risk exposure of Table 130.1 (excluding two banks of short time to market), the existence of foreign exchange risk exposure of listed commercial banks account for 28.57 %, RMB against the Japanese yen exchange rate risk exposure is the most obvious one which up to 71.43 %. It can be seen that the listed commercial banks face a certain degree of foreign exchange risk exposure, and the RMB against the Japanese yen foreign exchange risk exposure is the most striking one.

From the view of asymmetric coefficient of the listed commercial banks in Table 130.2, the proportion of listed commercial banks is 21.43 % according to the RMB against U.S. dollar asymmetry foreign exchange risk exposure. The proportion of facing RMB against Euro asymmetry foreign exchange risk exposure accounted for 28.57 %. The RMB against Japanese yen asymmetry foreign exchange risk exposure accounted for 64.29 %. It can be seen that the foreign exchange risk exposure of listed commercial banks have significant asymmetry.

It can also be seen from Tables 130.1 and 130.2, since facing different foreign exchange changes, listed commercial banks' foreign exchange risk exposure varies according to the "industry". After the asymmetry is considered, significance of foreign exchange exposure of the commercial banks which listed here is increasing to some extent. Thus, it is helpful for the listed commercial banks

Table 130.1 Foreign exchange risk exposure of single-bank

Bank Name	RMB/USD		RMB/EUR		RMB/JAP	
	β_m	β_s	β_m	β_s	β_m	β_s
SDB	0.9024***	−0.0009	0.8965***	−0.0011	−0.8962***	−0.0023***
BONN	0.8511***	−0.0006	0.8477***	0.0012**	0.8581***	0.0023***
SPDB	1.1301***	0.0096**	1.1293***	-0.0002	1.1233***	−0.0010*
HXB	1.1514***	0.0013	1.1584***	−0.0010**	1.1490***	−0.0003
CMSB	1.0454***	0.0072*	1.0486***	−0.0009*	1.0352***	−0.0015***
CMB	1.0605***	0.0006	1.0620***	−0.0002	1.0531***	−0.0015***
BONJ	1.0286***	−0.0064*	1.0290***	−0.0001	1.0380***	−0.0012**
IB	1.0986***	−0.0023	1.1046***	−0.0004	1.0965***	−0.0005
BOBJ	0.9897***	−0.0106***	1.0006***	−0.0008*	0.9833***	−0.0008*
BOC	0.9836***	0.0027	0.9802***	0.0005	0.9802****	−0.0006
ICBC	0.7928***	−0.0048	0.7904***	−0.0005	0.7847***	−0.0011***
CCB	0.7592***	−0.0028	0.7586***	0.0004	0.7569***	−0.0006
BOC	0.7829***	0.0045*	0.7836***	−0.0006	0.7772***	−0.0009**
CCTC	0.9960***	−0.0063	0.9906***	−0.0002	−0.9883***	−0.0020***

Note *, **, *** denote 10, 5, and 1 % significance level respectively

Table 130.2 Asymmetry foreign exchange risk exposure of single listed commercial bank

Bank name	RMB/USD		RMB/EUR		RMB/JAP	
	β_s	β_D	β_s	β_D	β_s	β_D
SDB	−0.0071	0.0094	−0.00002	0.0022	−0.0044***	0.0036*
BONN	−0.0086	0.01328	−0.0003	0.0028	−0.0019**	0.0080***
SPDB	−0.0100	0.0311**	−0.0011	0.0017	−0.0012	0.0005
HXB	−0.0033	0.0072	−0.0022**	−0.0020	−0.0016*	0.0027
CMSB	0.0010	−0.0100	−0.0024***	0.0030**	−0.0039***	0.0049***
CMB	−0.124	0.0213*	0.0002	−0.00085	−0.0011	−0.0007
BONJ	−0.0143	0.0134	−0.0021***	−0.0019**	0.0018***	0.0032***
IB	−0.1534	−0.0222	−0.0017	−0.0022	−0.0019*	−0.0030
BOBJ	−0.2833***	0.0457***	−0.0019*	−0.0022	−0.0017**	0.0025*
BOC	−0.0008	0.0057	0.0003	−0.0017	−0.0021**	0.0025*
ICBC	−0.0083	−0.0056	−0.0007	−0.0022*	−0.0021***	0.0019*
CCB	−0.0068	−0.0065	−0.0006	−0.0019*	−0.0019***	0.0021*
BOC	0.0027	−0.0028	−0.00007	−0.0001	−0.0019***	0.0015*
CCTC	0.0049	−0.0025	−0.0039***	0.0071***	−0.0022***	0.0004

Note *, **, *** denote 10, 5, and 1 % significance level respectively

more clearly to understand foreign exchange risk that they faced and take more rational foreign exchange risk prevention and response measures.

130.3.2 Whole Banking Sector's Foreign Exchange Risk Exposure

We take the unbalanced panel model to study the whole banking sector. Table 130.3 lists the parameter values of the banking sector's foreign exchange risk exposure and its asymmetries. From the situation of foreign exchange risk of the banking industry, the RMB against the U.S. dollar, RMB against the Euro foreign exchange exposure coefficients are not significant, while only the RMB

Table 130.3 Banking Industry's Foreign Exchange Risk Exposure Parameter

	Exchange rate	β_0	β_m	β_s	β_D
1	Rmb/usd	0.0005***	0.9772***	0.0011	–
	Rmb/eur	0.0052***	0.9771***	−0.00002	–
	Rmb/jap	0.0005**	0.9706***	−0.0012***	–
2	Rmb/usd	0.0009***	0.9772***	−0.0067*	0.0123**
	Rmb/eur	0.0011***	0.9761***	−0.0013***	0.0023***
	Rmb/jap	0.0013***	0.9692***	−0.0027***	0.0029***

Note The number *1* indicates the situation of foreign exchange rate exposure without asymmetry. The number *2* indicates the situation of foreign exchange rate exposure considering asymmetry

against the Japanese yen foreign exchange exposure coefficient β_s is significant at the 1 % level.

Taken into account of asymmetric foreign exchange risk exposure, the RMB against the U.S. dollar asymmetric foreign exchange risk exposure is the strongest one, the asymmetry of the RMB against the Euro foreign exchange risk exposure is relatively weak and the entire banking industry's foreign exchange risk exposure is asymmetric. Therefore, the whole banking sector's foreign exchange risk exposure significantly increases after considering the asymmetry.

130.4 Conclusion

With the degree of liberalization of the foreign exchange market continuing to accelerate and the banking industry in the context of the status of the national economy being increasingly important, listed commercial banks' foreign exchange risk exposure is becoming more and more prominent. The empirical study finds that:

After taking into account of asymmetry, foreign exchange risk exposure of listed commercial banks is significantly improved in addition to the RMB against the dollar. The number of listed commercial banks which facing RMB against Japanese yen foreign exchange risk exposure is the largest. Besides, after considering asymmetric foreign exchange risk exposure, the significance of whole banking sector's foreign exchange risk exposure improved significantly, especially the RMB against the U.S. dollar foreign exchange risk exposure.

Therefore, the innovation of this paper lies in: on the basis of considering the characteristics of asset returns sequence, the asymmetrical characteristic of the foreign exchange exposure is added in empirical model. It comparatively analyzes exchange rate risk exposure between different exchange rates and contributes to providing a new perspective for study of RMB exchange rates' foreign exchange risk exposure.

Acknowledgments This work was supported by Program for Changjiang Scholars and Innovative Research Team in University (IRT0916), and Science Fund for Innovative Groups of Natural Science Foundation of Hunan Province of China (09JJ7002).

References

1. Adler, M., Dumas, B.: Exposure to currency risk: definition and measurement. Finan. Manag. **13**(2), 41–50 (1984)
2. Adler, M., Dumas, B.: Exposure to currency risk: definition and measurement. Finan. Manag. **13**(2), 41–50 (1984)
3. Jorion, P.: The exchange-rate exposure of us multinationals. J. Bus. **63**(3), 331–345 (1990)
4. Griffin, J., Stulz, R.: International competition and exchange rate shocks: a cross-country industry analysis of stock returns. Rev. Finan. Stud. **14**(1), 215–241 (2001)

5. Luo, H., Jiang, C.: Foreign exchange risk exposure of listed companies in the RMB exchange rate formation mechanism. J. Zhongnan Univ. Econ. Law **163**(4), 78–81 (2007)
6. Ding, H., Xie, C., Chen, Q.: The empirical study of RMB exchange rate fluctuation affect the international enterprise value. Econ. Issues **8**, 102–108 (2008)
7. Tai, C.: Asymmetric currency exposure of US bank stock returns. J. Multinatl. Finan. Manag. **15**(4–5), 455–472 (2005)
8. Koutmos, G., Martin, A.D.: Asymmetric exchange rate exposure: Theory and evidence. J. Int Money Finance **22**(3), 365–383 (2003)
9. Carter, D., Pantzalis, C., Simkins B.J.: Asymmetric exposure to foreign-exchange risk: financial and real option hedges implemented by U.S. multinational corporations. Working paper (2003)
10. Pan, M., Liu, Y.A.: Exchange rate exposure: evidence from industry-specific exchange rates. Int. Res. J. Finance Econ. **84**, 121–132 (2012)
11. Júnior, R., José, L.: Understanding Brazilian companies' foreign exchange exposure. Emerg. Mark. Rev. **13**(3), 352–365 (2012)
12. Muller, A., Verschoor, W.F.C.: Asymmetric foreign exchange risk exposure: Evidence from U.S. multinational firms. J. Empir. Finance **13**(4–5), 495–518 (2006)
13. Brooks, R.D., Iorio, A.D., Faff, R.W., Fry, T., Joymungul, Y.: Asymmetry and time-variation in exchange rate exposure-an investigation of Australian stocks returns. Int. J. Comm. Manag. **20**(4), 276–295 (2010)
14. Koutmos, G., Martin, A.D.: Modelling time variation and asymmetry in foreign exchange exposure. J. Multinatl. Finan. Manag. **17**(1), 61–74 (2007)

Chapter 131
Influence of Highway Construction on Foreign Trade Based on Multivariate Regression Analysis

Rui Hua

Abstract Foreign trade contains business flow activities and logistics flow activities. With rapid e-commerce growth, the impact of logistics on foreign trade is increasing. As the infrastructure of logistics, the importance of highway is self-evident. Based on Heilongjiang's highway construction and foreign trade situation, the paper used the method of multiple regression analysis and analyzed the impact of highway construction on foreign trade in Heilongjiang and discussed the highway construction direction in Heilongjiang under the premise of enhancing the level of foreign trade, according to the analysis of paper, highway has a very important influence on total volume of imports and exports.

Keywords Multiple regression analysis · Highway construction · Foreign trade

131.1 Introduction

With the development of economy and transportation infrastructure progressing elaboration, the development of highway transportation is more and more quick. Highway transportation is flexible and rapid. With the dynamic flexibility and the fast of highway transport and strong adaptability, highway expands the scope of the commodity exchange and raises the speed of circulation of commodities.

High grade highway construction can shorten the time of commodity circulation, which actually the commodity circulation space distance is shortened relatively. Thus, it created conditions for enlargement of commodity circulation scale and development of trading market. Good transport condition can reduce the quantity of goods in transit and reserve, the number of goods in circulation, save

R. Hua (✉)
School of Management, Harbin University of Commerce, Harbin, China
e-mail: hsdglxyhr@126.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_131,

capital, improve economic benefits. The strength of regional economic relations broke closed state of the regional economy, so commodity can circulate in a broader space. High grade highway construction project can effectively reduce the circulation cost, narrow regional price differences. The phenomenon of price marked the smooth of the commodity circulation [1].

In recent years, with the formation of China Import and Export commodity distribution and logistics center and the rapid development of Heilongjiang Province ports trade, there appears a number of professional commodity wholesale market and service industry, but all of these are not lack of transportation development. In the initial stage of the development of frontier trade, the highway infrastructure is relatively backward after the trade development need, and railway transportation as the dominant mode of transport in trade development has made tremendous contributions.

Along with strategic partnership development between the China and Russia in the twenty-first Century, trade between the two countries and friendly exchanges continued to expand. Heilongjiang Province port transportation always upward momentum, according to this development tendency, railway transport is afraid very difficult to adapt to the further development. In order to avoid the traffic conditions of transport becoming bottleneck of trade development, excess capacity will be transferred to highway transportation. Therefore, the traffic condition and storage facilities are of great significance for border trade of Heilongjiang Province, and promote the development of local economic [2].

131.2 Current Situation of Foreign Trade in Heilongjiang Province

Heilongjiang Province as China's trade "bridgehead" with Russia has most border crossings in China. There are 25 the first ports allowed to open to the outside world in Heilongjiang Province, and 15 River Ports, 4 Port Road, 4 Aviation port Suifenhe port and Dongning port are most busy in the 25 port [3].

This makes the rapid development of Heilongjiang Province's trade with Russia. For decades, Heilongjiang Province to Russia trade always accounted for about 50 % of the total amount of foreign trade, accounting for about 20 % of Russia's national trade. In 2007, trade between Heilongjiang Province and Russia topped $10000000000 and reached a record of $11016300000000 in 2008 [4].

In recent years, the foreign trade of Heilongjiang Province has got rapid development. In 2009, the province's foreign trade had a brief period of decline, and began to rise again in 2010 and the total import and export volume exceed the 2008 peak, which showed a good development trend [5].

From the composition analysis of export and import products of Heilongjiang Province's foreign trade, we can find that agricultural products, textile clothing and footwear are the most and timbers are the most from the Russia [6].

131.3 Empirical Analysis of the Impact of Foreign Trade on Highway Construction

131.3.1 Estimation Model of the Total Import and Export as the Explained Variable

(1) The establishment and estimation of the model

Selecting each city's total import and export volume as the dependent variable JCK in Heilongjiang Province from 2006 to 2010, and checking Heilongjiang Province Branch highway mileage and road network density on Heilongjiang Province export total effect, we selects the representative index for highway mileage and road network density as independent variables. Specific indicators are as follows total mileage of highway ZLC, expressway mileage GS, highway mileage YJ, two highway mileage EJ, and road network density MD [7, 8].

According to the hypothesis, we set the following econometric model:

$$JCK = \alpha + \beta_1 ZLC + \beta_2 GS + \beta_3 YJ + \beta_4 EJ + \beta_5 MD + \mu \tag{131.1}$$

According to the highway construction drawings and the recent year's related data, we calculated the value of the dependent variable. According to the data from Statistical yearbook of Heilongjiang Province, we used Reviews software to estimate the model and get the results.

(2) Model checking and correction

In this model, ZLC and MD are not significant, coefficient C is negative and it does not meet the economic significance, so we shave it and get Table 131.1.

In the model, $R_2 = 0.66$, after adjusting R_2, we get $R_2 = 0.64$. Coefficient of determination is higher and model fitting degree is better.

Table 131.1 Adjusted Eviews regression results

Method: pooled EGLS (period weights)				
Linear estimation after one-step weighting matrix				
Variable	Coefficient	Standard error	t-Statistic	Probability
GS	253.9422	50.54896	5.023687	0.0000
YJ	206.2615	99.01711	2.083089	0.0425
EJ	91.77879	13.88232	6.611200	0.0000
Weighted statistics				
R-squared	0.662794	Mean dependent var		245073.2
Adjusted R-squared	0.649030	S.D. dependent var		273929.7
S.E. of regression	165504.9	Sum squared resid		1.34E + 12
Durbin-Watson stat	2.206427			
Unweighted Sstatistics				
R-squared	0.126811	Mean dependent var		141241.7
Sum squared resid	1.98E + 12	Durbin-Watson stat		1.777422

In the model, estimation results uses period weighting, estimation results of unweighted model $R_2 = 0.1268$, and weighted effect is obvious.

131.3.2 Estimation Model of the Total Export as the Explained Variable

(1) The establishment and estimation of the model

Selecting each city's total import and export volume as the dependent variable JCK in Heilongjiang Province from 2006 to 2010, and checking Heilongjiang Province Branch highway mileage and road network density on Heilongjiang Province export total effect, we selects the representative index for highway mileage and road network density as independent variables. Specific indicators are total mileage of highway ZLC, expressway mileage GS, highway mileage YJ, two highway mileage EJ, road network density MD.

According to the hypothesis, we set the following econometric model:

$$JCK = \alpha + \beta_1 ZLC + \beta_2 GS + \beta_3 YJ + \beta_4 EJ + \beta_5 MD + \mu \quad (131.2)$$

According to the highway construction drawings and the recent year's related data, we calculated the value of the dependent variable. According to the data from Statistical yearbook of Heilongjiang Province, we used Reviews software to estimate the model and get the results.

(2) Model checking and correction

In this model, the constant C, ZLC and YJ were not significant, so we shave it and get Table 131.2.

Table 131.2 Adjusted Eviews regression results

Method: Pooled EGLS (Period weights)					
Linear estimation after one-step weighting matrix					
Variable	Coefficient	Standard error	t-Statistic	Probability	
GS	96.16703	22.85831	4.207092	0.0001	
EJ	88.58881	14.60158	6.067069	0.0000	
MD	209.5814		80.16910	2.614241	0.0118
Weighted statistics					
R-squared	0.708223		Mean dependent var		198527.2
Adjusted R-squared	0.696313		S.D. dependent var		219600.1
S.E. of regression	108813.3		Sum squared resid		5.80E + 11
Durbin-Watson stat	2.526274				
Unweighted statistics					
R-squared	0.041773		Mean dependent var		101575.2
Sum squared resid	8.91E + 11		Durbin-Watson stat		1.740467

In the model, $R_2 = 0.70$, after adjusting R2, we get $R_2 = 0.696313$. Coefficient of determination is higher and model fitting degree is better.

In the model, estimation results uses period weighting, estimation results of unweighted model $R_2 = 0.041773$, and weighted effect is obvious.

131.3.3 Estimation Model of Total Imports as Explanatory Variable

(1) The establishment and estimation of the model

Selecting each city's total import and export volume as the dependent variable JCK in Heilongjiang Province from 2006 to 2010, and checking Heilongjiang Province Branch highway mileage and road network density on Heilongjiang Province export total effect, we selects the representative index for highway mileage and road network density as independent variables. Specific indicators are total mileage of highway ZLC, expressway mileage GS, highway mileage YJ, two highway mileage EJ, road network density MD.

According to the hypothesis, we set the following econometric model:

$$JCK = \alpha + \beta_1 ZLC + \beta_2 GS + \beta_3 YJ + \beta_4 EJ + \beta_5 MD + \mu \qquad (131.3)$$

According to the highway construction drawings and the recent year's related data, we calculated the value of the dependent variable. According to the data from Statistical yearbook of Heilongjiang Province, we used Reviews software to estimate the model and get the results.

(2) Model checking and correction

Table 131.3 Adjusted Eviews regression results

Method: pooled EGLS (period weights)				
Linear estimation after one-step weighting matrix				
Variable	Coefficient	Standard error	t-Statistic	Probability
GS	184.4357	16.84515	10.94889	0.0000
YJ	133.0711	36.96428	3.599992	0.0007
EJ	4.785062	1.510952	3.166920	0.0027
Weighted statistics				
R-squared	0.833191	Mean dependent var		102158.8
Adjusted R-squared	0.826383	S.D. dependent var		161298.5
S.E. of regression	76753.84	Sum squared resid		2.89E + 11
Durbin-Watson stat	1.590585			
Unweighted statistics				
R-squared	0.225016	Mean dependent var		49160.67
Sum squared resid	4.20E + 11	Durbin-Watson stat		2.105440

In this model, coefficient of C, ZLC and MD is negative and they do not meet the economic significance, so we shave them and get Table 131.3.

In the model, $R_2 = 0.833$, after adjusting R2, we get $R_2 = 0.826383$. Coefficient of determination is higher and model fitting degree is better.

In the model, estimation results uses period weighting, estimation results of unweighted model $R_2 = 0.225$, and weighted effect is obvious.

131.4 Conclusion and Suggestions

In conclusion, highway has a very important influence on total volume of imports and exports, imports, which proved that highway as a main line of highway traffic, its effect was very important, and construction of highways also will promote the development of foreign trade. Level 1 highway has significantly influence on imports, but not significant influence on exports, which may be decided by product features of Heilongjiang import and export. Level 2 highway as the main body of the port road had the influence on total exports than that on the total import, which shows commodity export transit is more dependent on highway transportation than import transit and means that the port road construction can promote export growth. Because the export commodities are agricultural products, textile and apparel products, highway density had large influence on the export.

Through the above theoretical and empirical analysis, in order to promote foreign trade in Heilongjiang Province, paper gives the following suggestions.

(1) Accelerating the construction of Expressway

Highway as traffic trunk road has very important influence on import and export. The construction of highways can make the relationship between ports more closely and make the goods materials circulation quicker, turnover bigger, lower, which has a positive meaning to promote the regional economy and foreign trade.

(2) Building products processing zone of highway and first class road as the core

At present, lumber resource is main imported products in Heilongjiang Province, and these products of high cost, large transportation quantity need high grade highway to transport. From the model, the constructions of the highway and level 1 road have main impact on import and export. Due to the limited resources, it is impossible to build high grade highway in a relatively large range of Heilongjiang Province. So researchers make highway, level 1 road as the core, and build processing area to reduce the cost of transportation and promote local economic development with the advantages of industrial clusters.

(3) Increasing the construction of port road, and appropriate the level of port road

In the analysis of total export, the influence of the level 2 road is very significant, which shows Port Road as the main body has an important effect on foreign

trade. Increasing port road construction can promote the port of the material circulation and the development of foreign trade. At the same time, from the contrast between the total volume of imports and exports of level 2 highway mileages, researchers find that import is less dependence on level 2, but more on railway transit.

Acknowledgments The work is supported by the Department of education projects for Humanities and Social Sciences of Heilongjiang Provincial (Project number:12522088).

References

1. Rui, H.: Research on logistics industry based on economic and trade cooperation between China and Russia in Heilongjiang. Research on the third industry structure optimization and innovation in Heilongjiang Province, vol. 4, no. 7, pp. 12–14 (2009)
2. Rui, H., LinShi, X.: Research on economic integration between Heilongjiang Province logistics industry development and the northeast Asia regions. J. Harbin Commer. Coll. **5**(3), 23–25 (2009)
3. Bo, W.: Logistics prediction and development planning of Heilongjiang province highway port. Northeast forestry university, Harbin (2010)
4. Heilongjiang bureau of statistics. Heilongjiang NBS survey office, the national bureau of statistics. Heilongjiang Province Statistical Yearbook (2008–2011)
5. Rietveld, P.: Infrastructure and regional development: a survey of multiregional economic models. Ann. Reg. Sci. **5**(6), 23–25 (2004)
6. Buton, K.J.: The Channel Tunnel-the economic implications for the South-East of England. Geogr. J. **7**(3), 13–16 (2005)
7. YanSong, H.: Impact assessment of border trade on Hailar-manzhouli highway construction in Inner Mongolia region. Jilin University, Jilin (2007)
8. YaDong, L.: Research on development relationship between Foreign trade and modern logistics development based on the empirical analysis from 1985 to 2007 statistical data. Heilongjiang Foreign Econ. Trade **8**(4), 23–25 (2009)

Chapter 132
Construction of Heilongjiang Forest Products Processing Industry E-Commerce Platform

Ying Cao

Abstract Judging by the fact that the forest product information features strong timeliness and other characteristics, the integration of existing e-commerce system has become the key to the development of forest product processing industry. In this paper, SOA employed to reorganize the existing resources in the manner of services. And loosely coupled architecture will be designed so as to tackle with the problems occurring in the forest product processing e-commerce system.

Keywords SOA · Web service · Forest products processing industry e-commerce platform

132.1 Introduction

As one of the largest forestry provinces in China, Heilongjiang Province has been endowed with the unique advantages of resources in developing forest products processing industry. In recent years, the forest products processing industry cluster began to take shape in Heilongjiang Province. The construction of Forest products processing industry e-commerce platform is conducive to the information of integrate forest products and the expansion of marketing channels, and is also the inevitable trend to adapt to the economic development.

Y. Cao (✉)
Collage of Economics and Management, Northeast Forestry University, Harbin, China
e-mail: 37263755@qq.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_132,

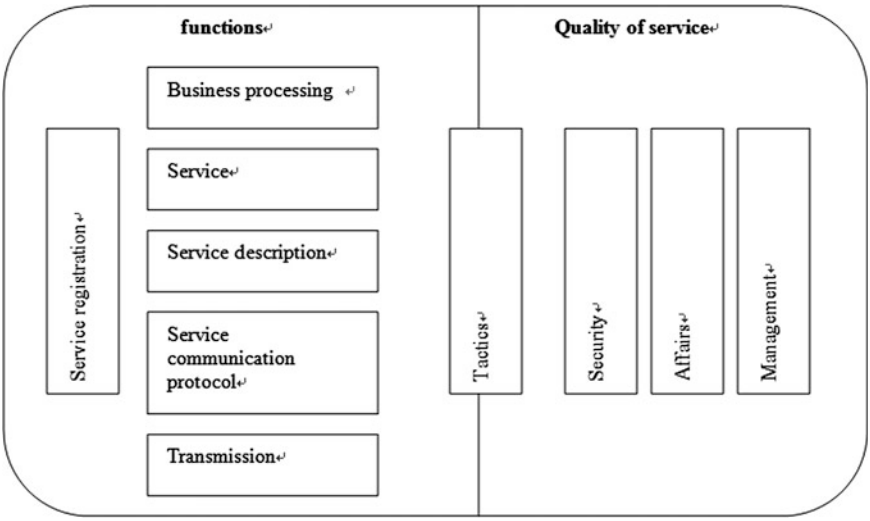


Fig. 132.1 Architecture element of SOA

132.2 SOA

The definition of SOA given by OASIS standards organization in reference model is that SOA (Service Oriented Architecture) is a software architecture model, which can organize and use distributed functions under the control of different owners. There are various perspectives in understanding SOA. In technical perspective, SOA is an architecture and it describes a kind of IT infrastructure, which makes it possible to exchange data with each other in different business services, participate in business processes and complete the specific business operation through a flexible way in collaborate with each other [1]. The business services are independent of the programming language, the implementation method and the operating environment. It combines the functionality and the quality of service organically, as shown in Fig. 132.1.

132.3 The Role of Forest Products Processing Industry E-Commerce Platform

Forest products processing industry e-commerce platform helps to create cooperation opportunities for domestic and foreign corporations with products production, trade, procurement, processing and downstream products, and make fair businesses with ultimate customers. The advantages can be included in the following aspects:

132.3.1 Information Disclosure

The platform makes it possible to convey the information without geographical restriction and the disclosure of information for forest products is open and transparent.

132.3.2 Fair Trade

The deal made on the same platform according to the principle that the products with best quality and the most preferential price comes first on the basis of different brands of forestry products, so it can avoid unfair bidding phenomenon.

132.3.3 Simple and Convenient

Because it is beyond geographical restriction, it makes it possible to achieve the long-distance trade in a short period of time, avoiding the inconvenience of traditional spot trading, and reducing the time cost significantly [2].

In short, forest products e-commerce platform plays a neutral role in both supply and demand parties, improves the credit of forest products in the market, and promotes the sustainable development of the forest products processing industry cluster.

132.4 The Design of the Forest Products E-Commerce Platform in Heilongjiang Province

132.4.1 Architecture

The e-commerce system architecture is divided into four levels: the presentation layer, the business layer, the service layer and the application layer, as shown in Fig. 132.2.

(1) Presentation layer: It is responsible for interacting with users, receiving users' input information from the application subsystem through browser and passing the information to the business level [3].
(2) The business layer: It is mainly responsible for categorizing related information according to specific business requirements and classifying and organizing services as workflow, and conveying the business information to the service layer.

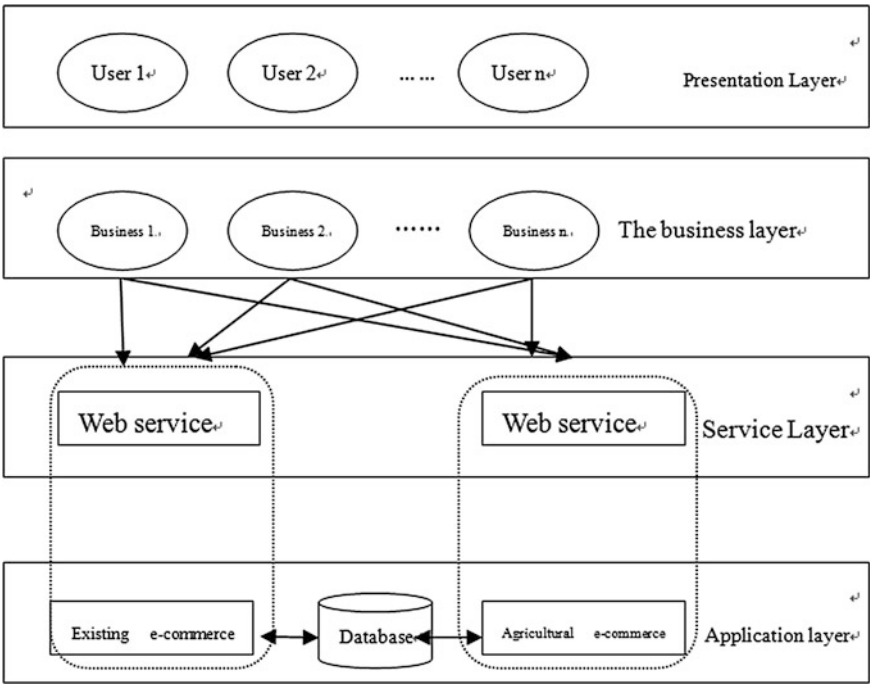


Fig. 132.2 Architecture figure

(3) Service layer: This layer has such functions as management, organization and implementation of various services, including registration services, demanding for distribution services and location-based services and so on.
(4) Application layer: It is responsible for integrating the existing e-commerce platform and the service requirements of the Web service, and then providing the usage of e-commerce platform to the services layer, in the manner of Web service to realize and employ the e-commerce platform which is designed by SOA [4].

132.4.2 System Function

According to the basic requirements of the e-commerce system, the whole system can be divided into following functional modules, namely user management module, security management module, supply and demand information management module, trading services module and repository support module, as shown in Fig. 132.3. Specific module functions are as follows.

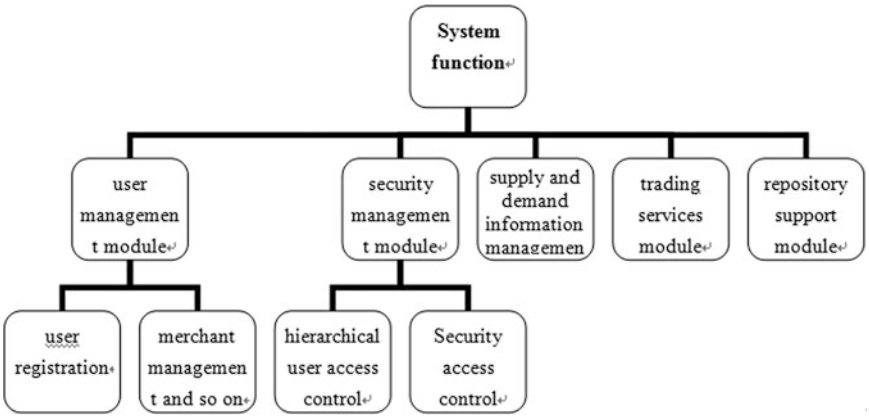


Fig. 132.3 Architecture of system function

132.4.2.1 User Management Module

The user management module mainly manages the user's information which can be divided into the user registration, common user management, hierarchical administrative privileges control, merchant management and so on [5].

132.4.2.2 Security Management Module

Security management module is mainly responsible for the management of security measures which can be divided into hierarchical user access control, security access control.

132.4.2.3 Supply and Demand Information Management Module

The key to the e-commerce platform design is to release supply and demand information and to establish trading relationship, which requires the publisher must be sure of the authenticity and accuracy of the published information, and the publisher must register. In addition, the supply and demand information should include forest products names, grades, quantities, unit price, origin places, contact information and relevant quality inspection certification when necessary. Procuring information includes forest products names, grades, quantities, prices and contact information. In addition, any users can have access to the supply and demand information in the validity time [6]. The platform provides two query methods: simple query and advanced query. For simple query, the user just needs to input the vegetable names, while the advanced query allows users to query combined information of multiple conditions, the obtained results are more accurate and more in line with the needs of users.

132.4.2.4 Business Service Module

The forestry product's e-commerce platform integrates the scattered forestry products commerce together in Heilongjiang province so as to form a virtual agricultural products market and a "virtual forest products market "on the Internet. The services that the platform can provide for trading parties are mainly online purchase, seller's delivery, buyer's receipt, the capital flows depending on the third party payment platform and so on. Online purchase services provide a chance for the customers to negotiate prices according to the needs of supply and demand of both parties and form electronic orders and electronic contract after negotiation, and pay the money to a third party in order to ensure the security of the transactions. The seller delivers the product when the third party receives the payment. Then the seller prepares for the relevant supply in accordance with the contract, and gives the products to the logistics enterprises that are responsible for transporting and at last sends the delivery information to the platform [7]. Buyer's receipt services means testing the quality of the received goods. If the products conform to the requirements of the contract, the buyer receives the goods and puts them in the stock, otherwise refuses warehousing and leaves the deal to logistics companies and the seller. The services of the capital flows depending on the third party payment platform can adequately guarantee the security of capital flows of both buyers and sellers.

132.5 Key Technologies

The key to implementing the forestry products e-commerce platform is to realize the information integration and interaction by means of Web Services as well as establishing the architecture of the Web Services. NET platform fully supports Web services and employing the built-in Web services development functions can easily develop Web services. Moreover, in the development process, the development of logic module and the generation of WSDL service definition file can be completed through the NET platform. Web services are established on the basis of the industry-standard, such as SOAP, XML and WSDL. Web services can solve the message interoperability problems on the Internet by means of SOAP, which adopts two widely used protocols: HTTP and XML. HTTP is used for SOAP message transmission, while XML is the encoding mode of the SOAP. When the information and data interchange between systems, you can output the data in the database for the WSDL document that conforms to the XML standard, and then bind the WSDL with the SOAP, finally complete the communication by HTTP-GET/POST agreement. Obviously, as long as the client sends a compliant SOAP message according to the format of the service description, then the client can use the Web service that has been released. After completing the Web services development, it also needs to be registered on the Internet for the discovery, application and integration of other subsystems.

132.6 Conclusion

The forestry product e-commerce design integrates the features of Web Services technology and SOA architecture, thus unifies existing e-commerce systems. Thanks to loose coupling, coarse-graining, strong mutual interoperability and other features of the SOA structure system, the system is of good scalability, compatibility and adaptive capacity. It also effectively integrates the dispersible resources of forestry products.

Acknowledgments The authors would like to express sincere thanks to support of Natural Science Foundation of Heilongjiang (G201105) project.

References

1. Zheng, H., Liu, L., Li, Y.: An information fusion-enabled third party e-commerce platform based on SOA. J. Softw. **1**, 50–57 (2009)
2. Anonymous. Research and Markets: SOA Applications Middleware Market at $3.5 billion in 2009 is anticipated to reach 8.2 billion by 2016. M2 Presswire (2010)
3. Anonymous. Research and Markets: Worldwide Services Oriented Architecture (SOA) infrastructure market shares, strategy, and forecasts, 2009–2015. M2 Presswire (2009)
4. Zhang Q.: School of Glorious Sun of Business and Management. Heterogeneous system integration of supply chain based on SOA & WSC (2010)
5. Liu, L., Wei, H., Tan, Z.: Build forest product electronic commerce platform. Hunan For. Sci. Technol. **4**, 100–103 (2010)
6. YangPing.: Agricultural product electronic commerce platform construction. Hunan Agri. Mach. **1**, 85–87 (2012)
7. Zeng, J., Liu, B.: Regional manufacturing industry cluster of e-commerce platform construction. Commercial Era. **32**, 50–55 (2010)

Chapter 133
Multiple Case Studies of Global Enterprise System Implementation in China

Roger L. Hayen and Zhenyu Huang

Abstract This study examines and compares multiple global enterprise operations in China. Business ownership arrangement, business products and services, and IT support for those agreements and products are considered. Findings indicate business ownership as a moderator impacts the relationship between business processes and IT support choices. All global enterprises in the study were found to use the SAP enterprise software for their transaction processing and reporting. Furthermore, enterprises also use proprietary software in the creation or delivery of their primary product or service. Wholly-owned businesses are more likely to use proprietary software for their key businesses or specialty.

Keywords Global enterprise · Information technology · China, business ownership · SAP · Enterprise software · Transaction processing software · Proprietary software · Case studies

133.1 Introduction

No country in the world has seen greater expansion in its automotive industry over the past 25 years than China. Vehicle sales grew by 32.37 % year-on-year to 18.06 million in 2010, making the country the world's biggest auto market for the second consecutive year [1]. The number of state-owned companies in China is decreasing. State-owned companies account for just over a quarter of the total

R. L. Hayen (✉)
Grawn Hall 302A, Central Michigan University, Mount Pleasant, MI 48859, USA
e-mail: hayen1rl@cmich.edu

Z. Huang
Grawn Hall 319, Central Michigan University, Mount Pleasant, MI 48859, USA
e-mail: Zhenyu.Huang@cmich.edu

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_133,

number of companies. However, they produce 45 % of China's industrial values and more than half of the industrial profits [2]. Foreign companies make up 20 % of the total number of companies producing nearly 60 % of China's exports. Every one of China's top twenty companies, measured by revenues, is state-owned [2]. It is in this environment that global enterprises must operate in China.

A global enterprise is a legal entity that does business around the world. It has a world-wide presence. Global enterprises amplify the complexity of many different kinds of business partnerships. Each venture is different so different information technology (IT) support is needed. Business organizations and operations are considered in examining the environment in which enterprises deploy IT resources. This view provides insight to underlying requirements for IT support [3]. It is important to understand business organizations' operations when comprehending the association of IT support to its business processes [4]. This study examines core business processes together with their observed IT support. The analysis is conducted using a case study approach. A framework of the analysis is presented followed by the research methodology. This is applied to three global enterprises in China. The alignment of the framework is presented to summarize the observations of the operations and IT support of global enterprises in China.

133.2 Framework

Tse [2] presents a product-market-freedom matrix for China. This matrix has a classification scale from restricted to free for business segments. None of the global enterprises examined are in the restricted product market or have restricted ownership, as these are Chinese government controlled ventures. Applying Tse's framework, global enterprises fall into two categories: joint-venture or wholly-owned. Ownership type is considered and compared to this framework for each company studied. A question to be answered in assessing IT deployment is whether ownership has an observed impact on a company's IT and how this ownership status impact IT support choices.

IT software used by global enterprises in China can be categorized as global or local. Global is generally software used by many different companies and in many different countries. Local is the software used by a single company, but leveraged in many different countries. Enterprise software (ES) exemplifies global, while local is unique, proprietary software, such as that used in manufacturing control. Within ES, its deployment may be categorized as separate instances, separate clients, or same clients. The use of separate instances means that application link enabling (ALE) is needed for sharing data [5].

133.3 Methodology

The case-based research method is followed in this study. The approach used examines characteristics and features of the business, corporate environment, and IT deployment. Case-based research provides a channel for investigating phenomenon in information systems and is particularly appropriate for exploratory studies [6]. Applied to IT and global enterprises, case-based research affords a schema for studying application characteristics. According to Voss, Tsikriktiss [7] and Meredith [8], case-based research is one of the most powerful methods for generalizing conclusions about a field of study. Results from case-based research can have a very high impact that leads to new and creative insights with a high validity for practitioners—the ultimate user of research.

The companies selected in this study were based on global enterprise personal contacts through the Information Systems Advisory Board at the authors' university. The varied membership of this Advisory Board provided a random selection of global enterprises. Each site visit was uniquely organized for this study with a particular focus on both facility operations and their IT deployment.

The data collection method was personal site visits to facilities of each selected enterprise in the Shanghai, China area [3, 4]. The visit duration was approximately one-half day and included both a facility tour and a meeting with managers including those in IT. Three locations were visited: Shanghai GM final assembly, Delphi China manufacturing, and Coca Cola China administration and bottling. The time period for these visits was during the 1-month period from mid-May to mid-June. 2011. Data describing each enterprise is reported and analyzed here.

133.4 Global Enterprise Cases

Three global enterprise case applications are presented here to gain an understanding of their organization and IT support. The company cases are Shanghai GM, Delphi China, and Coca-Cola China.

133.4.1 Shanghai GM

The site visit to Shanghai GM (SGM) took place at their South Plant facility located in JingQiao Town, Pudong District, Shanghai. It began at the SGM IT building and was conducted by a member of the SGM IT team. The tour concentrated on the final vehicle assembly area where the manufacture of a car is completed.

SGM has 13,000 employees within China [9]. It is a joint venture of General Motors Company (GM) and Shanghai Automotive Industry Corporation (SAIC).

SAIC is unusual in the Chinese auto industry in that it is a state-owned enterprise of the Shanghai city government. Most other auto companies in China are owned by or a joint venture with China's central government. SGM began assembling the venture's first vehicle, the Buick Regal, in Shanghai, China in 1999 [10].

In 2010, GM sold more vehicles in China than in the US [11]. In the first half of 2011, SGM saw sales jump 25 % from the prior year to a record 600,002 units. Strong demand for the Buick brand jumped 28 % to 324,919 vehicles [12]. The Buick brand in China is exceptionally strong. Approximately two Buicks are sold in China for every one sold in North America. To supply this demand, SGM final assembly has a production rate of approximately 40 vehicles per hour. It runs two shifts per day and frequently operates 6 days a week. None of the vehicles produced at SGM are exported outside of China. The demand for vehicles is so great in China there is no reason to export them.

SGM is one of the most advanced vehicle manufacturing operations on the planet with its advanced flexible manufacturing system. At the time of the visit, four different vehicle models were being assembled on the same final assembly line. The model types were intermixed in any combination. These vehicle models are premium brands that include Buick. It is in this business environment that SGM IT provides software support.

SGM deploys the SAP ECC ES as the overall transaction processing application. This software handles the build schedule of products. However, the actual final assembly operation is processed using separate proprietary software on the manufacturing floor. This software is unique to GM and is provided and supported by a third-party consulting organization. The software is shared with other GM assembly plants, which leverages the deployment of this specialized software.

Final assembly is divided into 60 major work centers. A work center is a place on the assembly line where a significant and measurable operation takes place. Final assembly receives a completed vehicle body from the body build and paint shop. The attachment of a physical paper document to the hood of the vehicle is the first operation. This document contains bar codes used for unique vehicle identification including the vehicle identification number (VIN). The VIN is the ultimate vehicle identifier throughout its entire life cycle. At this first station, the doors are removed from the body to be completed in a separate assembly operation and will later be placed back on the same vehicle. The bar code confirms arrival of the vehicle body.

At each work center, a bar code reader is used to collect the VIN and to identify the part(s) being added to the vehicle. Some parts have unique identifiers, such as the engine or transmission. This unique number is recorded and included in the vehicle build record. For batches of parts, such as headliners, then the batch number is recorded. A complete record is created of all parts going into the final vehicle assembly. For a future warranty, parts can then be traced back to the individual subassembly or batch. Clearly, without IT support for vehicle assembly it would be impossible for SGM to build a vehicle. Bar code collection is first, and actual assembly of the part is second.

As vehicles are "born"—they are started and driven from the assembly line. During inspection, vehicles with any identified problem go to a holding area for remediation. The last final assembly work center is the water test. Completed vehicles are driven into a special car wash. It is then inspected inside and outside for any water problems. A repair area also exists for this final inspection. However, it was observed that few vehicles require repair and many occurred quickly.

Overall, GMS IT support encompasses caring for the SAP ECC ES and the GM proprietary software used in vehicle assembly. The finally assembly software is the pillar of flexible manufacturing and its related activities. GM leverages this proprietary software by using it throughout its global operations.

133.4.2 Delphi China

The site visit to Delphi Packard Electric Systems Co., Ltd., commonly known as Delphi China (DC), occurred at their Anting Town, Jiading District, Shanghai facility. Hosting was by a business planning associate and included the operations director, IT director, and toolroom manager.

Delphi operates 17 wholly owned entities and joint ventures in China with about 12,000 employees. DC is one of China's largest auto parts producers. It supplies almost all automakers in China, including SGM, Shanghai Volkswagen, Ford, Toyota, and Chery. Delphi produces three major product lines in China including the electronic/electric distribution system manufactured at the Anting facility [13]. SGM and other GM manufacturing plants consume only about one-fourth of Delphi's production capacity. Only about 5 % of DC's production is exported to North America. Most is consumed in China or elsewhere in Asia and in Europe.

The Anting facility is a wholly owned operation. It does plastic injection molding, metal stamping, and assembly operations. Furthermore, one of the most advanced tool making operations in the world is located here. Plastic injection molding machines make vehicle electrical connectors and fuse blocks (panels). Metal stamping machines produce the blade connectors used with printed circuit boards in fuse blocks. Fuse blocks are assembled using injection molded outer cases and printed circuit boards. The plastic injection molding machines are all computer controlled with a single operator monitoring up to six machines. Completed parts are automatically ejected with the sprue and runner (a channel through which plastic is poured into a mold) re-ground and re-used at the same machine. The metal stamping machines are similarly controlled with these machines operating at the rate of up to 8,000 strokes per minute. Eight thousand parts can be produced per minute. Very high precision is required when metal stamping occurs at these speeds.

DC runs its operations using the SAP ECC ES. SAP is configured to integrate Delphi's operations globally in a single instance. A most interesting point was viewing a web-based screen showing the current production status of injection

molding machines and stamping machines globally—a strategic enterprise management application. The operations director displayed the screen reporting the status of every machine in Delphi's world operations—not just one plant. Was that machine in production, down for changeover, or down for maintenance? This was most impressive and illustrates how ES is deployed to monitor global manufacturing operations in real-time. Clearly, this underscores the need for global businesses to exploit world-class ES for transaction processing and reporting.

DC is the home to one of five of the most advanced robotic toolrooms in the world. A toolroom is a place where tools and dies are made. For example, when doing plastic injection molding, a metal mold is the tool which is made for that process. Liquid plastic is injected into the mold (tool) and then cooled to form the desired part. Vast numbers of plastics items use molds of this kind. DC's toolroom makes their molds. A robotic machine exploiting computer-aided engineering (CAE) is housed in toolroom for making molds. Square corners are a challenge in making electrical connectors. Many tools used to make molds are round cutters. It is difficult to make a square corner with a round cutting tool. So, a method of electrical discharge creates square corners with the robotic CAE machine. This applies a special language to control tool making. Delphi finds it difficult to hire individuals with this specialized skill. Nonetheless, in China, individuals will do self-study and learn to write CAE programs. In general, they do not find a similar situation for North American skilled workers. Hence, producing tooling in China is easier for Delphi. The DC toolroom produces tooling used in 80 % of Delphi's operations world-wide. Approximately 40 completed tools per month are shipped from DC to other Delphi plants.

Overall, DC uses the SAP ES in managing its day-to-day operations. Delphi has configured the SAP software to provide global monitoring of their production resources. Special purpose software is used in the control of production machines, while proprietary software is used in creating production tooling.

133.4.3 Coca-Cola China

The Coca-Cola Company (TCCC) is celebrating 125 years of producing their star beverage Coca-Cola. This site visit of the Coca-Cola Bottling Investment Group China, commonly known as Coca-Cola China (CCC), was at their Shanghai headquarters in Jinqiao Export Processing Zone, Pudong New District, Shanghai. Hosting was by the group head of public affairs and communications and included one of eight group managers and the IT manager.

Unlike the other enterprises visited in Shanghai, CCC was unique in that it contains a Coca-Cola museum with an auditorium for greeting visitors. That is, CCC is more open to visitors than the other enterprises. However, their visits are typically limited to the museums and the bottling lines. Our visit was customized and specialized with the management meeting.

The CCC facility is a wholly-owned business enterprise of TCCC. However, TCCC also has many joint-venture operations in China. TCCC has arranged China into fundamentally three areas. The southeast with Shanghai is the CCC area. The northeast with Tianjin is a joint-venture with the Chinese government. And, the west is a special arrangement with a non-governmental Chinese company. TCCC is a direct majority partner in four bottling joint ventures: one in Shanghai, two in Tianjin, and one in Hainan Province. However, to maintain control over the ingredients and formulas of its drinks, TCCC produces the concentrates for its international-brand soft drinks at CCC, the wholly foreign-owned factory in Shanghai and home to the company's China headquarters [14].

TCCC is committed to the China market with 1.3 billion consumers. The CCC bottling plant produces about 3 billion bottles of Coke product a year. This is only about 10 % of China's consumption of these products and one and a half days of Coke's global production. In China, the current per capita consumption is only 32 bottles of Coke products annually. This compares with a consumption of 150 in Hong Kong and more than 500 in the US. Clearly, the China market has outstanding growth potential for TCCC with the soft drink industry growing at a 12.8 % compounded annual growth rate [15].

At the bottling plant visited, each different beverage is produced on a separate bottling line. There is one line for Coca-Cola and another for Sprite and so on for each of the other products. There is one exception and that is with Coke Light (China's equivalent of Diet Coke) and Coke Zero which may be produced on the same bottling line because of the similarities between them. No changeover of a bottling line occurs for drink flavors. Demand variations are satisfied by running a bottling line for longer periods of time. It is common to run the bottling operations 6 days a week during the summer months—peak demand time. CCC produces beverages packaged in plastic containers. Blank plastic bottles are made at this plant for use in product packaging. The final forming of plastic bottles occurs during a blow molding operation at the beginning of the bottling line. This allows a common size blank, for example for a half-liter, to be used for all different products. That is, the form of a Coca-Cola bottle is different form a Sprite bottle. However, the same blank bottle can be used for each of these drinks. This simplifies the production of bottles.

CCC deploys the SAP ECC ES as the overall transaction process application. This software is used in managing their bottling operations. As a wholly-owned business there is a tight integration of SAP ES. TCCC and CCC are included in the same SAP client. This indicates a highly integrated, one-company global operation. Using the same client leverages the IT support at the corporate level. However, this requires the processing of the CCC transactions on the same instance (computer system) as those of TCCC. Distances and time differences between these facilities present a challenge with a single client implementation. For example, a 2:00 a.m. system up data at TCCC headquarters occurs at 2:00 p.m. at CCC, a prime processing time there. With SAP implementations a common arrangement for global enterprises is to use separate clients when they

have disparate legal and reporting requirements in different countries and encounter time zone issues.

Overall, CCC IT support encompasses support for the SAP ECC ES that shares the same instance and client of TCCC. Deployment with a single client delivers a very tight integration among global enterprise operations. On the other hand, time zone issues present situations that must be considered by IT managers in the utilization of ES.

133.5 Analysis and Discussion

All the global enterprise operations in this case research were found to use the SAP ECC ES for their core business processing and reporting requirements. Two of the operations also exploited proprietary software to support overall business requirements (Table 133.1). In this study, SAP ECC ES is categorized as global, whereas the proprietary software is local. Observation of the deployment of proprietary software found it to be a primary IT resource in the enterprise's product manufacturing or service delivery. Or, that software had a very specialized use, such as in the robotics toolroom operations of DC. Proprietary software deployed by SGM and DC is of particular importance as it is fundamental system by which their product or service is produced. The proprietary software is used to support their key or specialty business. Generally, IT support for these enterprises is diverse and requires a vision and provision for more than just ES.

Based on the product market and ownership matrix of Tse [2], the expected relationship for the enterprises in this study is shown in Table 133.2. CCC requires additional consideration. Table 133.3 shows the observed product market and ownership relations of the global enterprises in this study. The business units visited support that was proposed by Tse. However, the entire operations of TCCC

Table 133.1 Software deployment

Enterprise	Deployment	
	SAP ECC (global)	Other (local)
Shanghai GM	X	X
Delphi China	X	X
Coca-Cola China	X	–

Table 133.2 Expected product market and ownership

Enterprise	Ownership	
	Joint venture	Wholly owned
Shanghai GM	X	–
Delphi China	–	X
Coca-Cola China	–	X

Table 133.3 Actual product market and ownership

Enterprise	Ownership	
	Joint venture	Wholly owned
Shanghai GM	X	–
Delphi China	–	X
Coca-Cola Chin	X	X

in China are much more intricate than the site visited in Shanghai. Like a spider web, it is a complex intertwining of wholly-owned operations, joint-ventures with Chinese companies, and joint-ventures with the Chinese government. This places TCCC's China operations in a partially restricted product market and ownership position, rather than in the free position indicated by Tse. Nonetheless, overall, there is general alignment between expected results and observed results.

Enterprise ownership is important because of its impact on the relationship of business processes with IT support choices. Business processes overall determines the IT choice. For example, due to the scale and diversity of business and logistic processes, global enterprises have to count on robust enterprise systems like SAP ECC to help manage daily transaction. This is especially critical as it relates to the necessary reporting and legal requirements of the enterprise. The relationship between business processes and IT choices is moderated by ownership status. The role of the Chinese government in some, but not all joint ventures appears to add complexity to the reporting and legal requirements, which in turn impacts the IT choices. This is evidenced by the deployment of separate SAP ECC instances in the case of SGM. CCC is more difficult to assess as the whole of business in China by TCCC is a combination of wholly-owned operations and joint-venture arrangement. However, the wholly-owned operations, which include the head-quarters for China operations, are tightly integrated with TCCC's operations through the single client arrangement within the SAP ECC ES. The observed impact of ownership on IT support is summarized in Fig. 133.1.

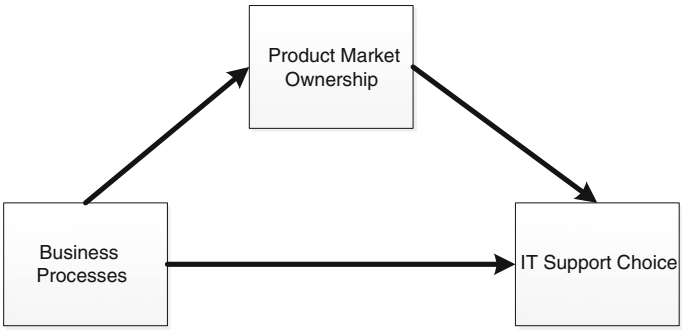


Fig. 133.1 Path to IT support

133.6 Conclusion

The delivery of IT support in global enterprises in China is complex. This complexity emanates from the ownership arrangements within that business environment. The global enterprise pact in China may be wholly-owned, a joint-venture with a government Chinese entity, a joint-venture with a non-government Chinese entity, or some combination of these business relations. Different agreements foster varied reporting and legal requirements. The enterprises surveyed all use the SAP ECC ES, which is driven by the complexity and need of global business transactions.

This study is useful for students learning about IT deployment in organization with a particular focus on global enterprises. It should help them better understand the environment in which IT is delivered together with an appreciation for business issues that impact that IT utilization and career preparation.

Considering this study's findings, future research may take two directions. The first is a more in depth study of an individual global enterprise, such as all of The Coca-Cola Company's various ownership relationships and their IT support. The second is a more comprehensive study across a larger set of global enterprises operating in China and including the framework put forth here.

The study presents three case studies. However, the global enterprises chosen for this study are based on the authors' personal connections. These enterprises, though convenient samples, represent three different industries and present valid and adequate information for the case studies. They provide significant values for scholars to investigate the IT choice and the effect of market product ownership status.

References

1. China Car Times. Delphi aiming for growth in China. Chinese Car News (2011)
2. Tse, E.: Is it too late to enter china? Harvard Bus. Rev. **88**(4), 96–101 (2010)
3. Hayen, R.L., Huang, Z. (eds.): Information Technology of Global Enterprises in China: A Case Perspective. The International Academy of Business and Public Administration Disciplines, Orlando (2012)
4. Hayen, R.L., Huang, Z.: Enterprise software applications in China (Unpublished raw research data) (2011)
5. Hayen, R.: SAP R/3 Enterprise Software: An Introduction. McGraw Hill Higher Education, London (2007)
6. Yin, R.K.: Applications of Case Study Research. Sage Publications, Newbury Park (1993)
7. Voss, C., Tsikriktiss, N., Frolich, M.: Case research in operations management. Int. J. Oper. Manag. **22**(2), 195–219 (2002)
8. Meredith, J.: Building operations management theory through case and field research. J. Oper. Manag. **16**, 441–454 (1998)
9. Top Employers. China's Top Employers 2011—General Motors. http://www.topemployers.com.cn/en/ChinasTopEmployers/ChinasTopEmployers2011/C/tabid/5333/C/444/ShanghaiGeneralMotors.aspx (2011)

10. Pelfrey, W.: The start of Shanghai GM. http://history.gmheritagecenter.com/wiki/index.php/1982-1999_Globalization (2008)
11. Ash, H.: No shocking news: GM sells more cars in China than it does in the USA. http://www.chinacartimes.com/2011/01/26/no-shocking-news-gm-sells-more-cars-in-china-than-it-does-in-the-usa/ (2011)
12. Associated Press. General Motors says 1H China sales up 5.3 pct. http://www.huffingtonpost.com/huff-wires/20110705/as-china-gm/ (2011)
13. Wu, J.: Corporate news: Beijing gives its nod to GM on Delphi deal. Wall Street J. (Eastern edition), B.3 (2009)
14. Weisert, D.: Coca-Cola in China: quenching the thirst of a billion. China Bus. Rev. (Internet). http://www.chinabusinessreview.com/public/0107/weisert.html (2001)
15. Associated Press. Coke adds billion dollar brand from China. http://biz.thestar.com.my/news/story.asp?file=/2011/2/2/business/7918410&sec=business (2011)

Chapter 134
Service-Based IT Management Model with Emphasis on Existing Outsourcing Areas

Michal Sebesta and Jiri Vorisek

Abstract Ensuring effective business/IT alignment is considered a critical task for the contemporary IT managers. One of the ways to achieve better results is to use a service-based approach. However, existing methods such as ITIL or Enterprise Architecture frameworks are very extensive and broadly oriented. General lack of appropriate methods and tools limits the use of the approach among organizations. This paper presents a service-based IT management model named SPSPR which represents a lightweight solution suitable for a variety industries and business segments including the small and medium-sized enterprises. The main goal of the proposed model is to assist with aligning business and the IT within the whole organization, and to enable effective IT support of business processes using ICT services. This paper describes a view on outsourcing areas and on their relation to the presented SPSPR model. Within the analysis, it maps these areas onto specific layers and thus supports the potential use of the SPSPR model within organizations that decide to utilize outsourcing in any of its available forms.

Keywords Outsourcing · SPSPR · Business process · ICT service · Business/IT alignment

134.1 Introduction

Growing complexity of enterprise IT demands methods and tools that could be used to effectively manage resources while reflecting related business processes. Aligning business strategy with IT strategy, reviewing business processes, and subsequently rearranging resources is crucial and has been discussed by a number

M. Sebesta (✉) · J. Vorisek
Department of Information Technologies, University of Economics, Prague, Winston Churchill Square 4 130 67 Prague 3, Czech Republic
e-mail: michal.sebesta@vse.cz

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_134,

of researchers such as Henderson and Venkatraman [1], Chan et al. [2], or Duncan [3]. Effective business/IT alignment can be from our viewpoint ensured by utilising the service-based view of the organization and its IT.

Service-based IT management[1] is a concept that is slowly but surely gaining its place within the contemporary management practice. A good definition of the concept is provided by Bon et al. [4] who state that it is "a management of all processes that cooperate to ensure the quality of live IT services,[2] according to the levels of service agreed with the customer." Moreover, they add that the "Initiation, design, organization, management, provision, support, and improvement of IT services tailored to the needs of an organization are addressed."

Recent example of a service-based approach currently used in practice can be the Information Technology Infrastructure Library (ITIL).[3] The ITIL [5] introduced several methods around the service-based view and is popular especially within the segment of large enterprises. However, the ITIL approach itself is very broad, demanding in terms of resources, and therefore not very suitable for small and medium-sized enterprises (SMEs). Other approaches including Enterprise Architecture (EA) frameworks[4] in most cases share a similar attribute of extensiveness and over-complexity. Due to these factors and due to subsequent high overhead costs associated with the use of these approaches, we can expect their future use mainly within the segment of large enterprises. From our viewpoint, the lack of available approaches for small and medium enterprises is currently a significant gap within IT management research.

Our long-term aim is therefore to develop a model for service-based IT management that might be useful for organizations from a variety of segments. Special emphasis should be put on efficiency and clarity of such model in order to enable its use within the growing SME segment.

Important part of our model is the area of ICT services that we primarily discuss in the first part of our paper. The reason why it is important is that the perception of ICT service significantly affects understanding of our presented service-based IT management concept.

In the second part of our paper, we present our concept named SPSPR,[5] which provides a hierarchical view of the position of ICT services within organizational architecture. This concept provides a global view on strategy, processes, ICT services, ICT processes, and ICT resources of the whole organization. It also helps with identification of the possible interconnections between objects from various layers in the hierarchy. Together with the model, we present a taxonomy of ICT

---

[1] Sometimes referred as IT Service Management.

[2] In this paper, we will use the terms IT services and ICT services interchangeably.

[3] Currently in its version 3 (ITIL V3) [5].

[4] An example of contemporary EA frameworks could be TOGAF [6], DoDAF [7], or FEAF [8].

[5] Model of business and IT alignment, which divides the business and IT management into five layers: S—Strategy, P—Business Processes, S—ICT Services, P—ICT Processes, and R—ICT Resources.

services according to service content summarizing the service types that might be used within the ICT services layer.

Third part of the paper is devoted to an analysis and mapping of the outsourcing areas to the presented SPSPR model. Main aim is to summarize existing areas of outsourcing and outline their relation to particular layers of the hierarchy. Identification of these areas and the associated layers is crucial. Decision-making about outsourcing within various layers of the model is built on different principles. Clear differentiation of the possible outsourcing areas within the SPSPR model can help organizations with maintaining a better picture of the organizational architecture. Most importantly, it can provide a sound base for the decision-making about outsourcing.

Our main aim within this paper is to:

- Briefly discuss the ICT service definitions (Sect. 134.2).
- Describe our approach to service-based IT management—the SPSPR model (Sect. 134.3).
- Outline classification of ICT services used within the model (Sect. 134.3.1).
- Present outsourcing areas and describe their relation to particular layers within the model hierarchy (Sect. 134.4).

134.2 ICT Services

In the service-based IT management, ICT services play a central role as a communication interface between business and IT. Therefore, it is of high importance to discuss the term ICT service in more detail.

We think that defining the ICT service primarily requires clarification of the general definition of service. The concept of services has been discussed for a number of years; for instance, Lovelock [9] provided a comprehensive analysis of the term and presented a general classification of services. For our aim, we have identified several useful definitions. For example, Kotler [10] defines the service as "a deal, which one party may offer to the other party and which is basically immaterial." We may further deepen the definition. For instance, Booth et al. [11] define service as "an abstraction of some source, which is represented by source capability of task processing with coherent functionality from the point of view of service provider as well as of service consumer, where in order to use the service it must be realized by a particular provider's agent."

The aforementioned definitions provide necessary space for a separate ICT service definition, which is essentially its subtype that bears specific attributes. From the definitions we analyzed the most suitable seem the ones introduced by Vorisek [12], Skala [13], ITIL [5], Gala et al. [14], and Schekkerman [15].

For instance Vorisek [12] presents that "an ICT service is represented by activities and/or information provided by an ICT service provider to ICT service consumer."

Schekkerman [15] defines the ICT service as "an implementation of well-defined business function, which is executed independently of the state of any other service defined in a system."

On the other hand, some of the definitions are aimed more on the hierarchical aspect of ICT services where Skala [13] defines the ICT service as "a particular functionality, provided by ICT service provider that enables execution of a particular business process." As added by Gala et al. [14] in order to use the ICT service, it must be realized by a particular provider's resource.

Finally, the ITIL [5] mentions that an ICT service is based on ICT use, and that it supports organizational business processes. Moreover, they present that an ICT service is created by people, processes, and technology and should be specified in a service level agreement (SLA). This perspective is also supported by Schekkerman [15] where he also mentions that "services present a well specified set of interfaces and are executed based on agreement between service client and service provider that is specified beforehand." Also from our viewpoint, the existence of SLA is a very important aspect that benefits especially the monitoring and control of the organizational ICT services.

According to our view on ICT services, we could summarize our own definition of ICT services as follows:

> *ICT service is represented by coherent activities and information delivered by ICT service provider to service consumer. ICT service is implemented by ICT processes, which consume ICT resources (hardware, software, data, people,* etc.*) during their execution. Service is realized on the basis of agreed business and technological conditions.*

This definition is a cornerstone of our perception of service-based IT management that we depict within the following section.

134.3 Service-Based IT Management Using the SPSPR Model

After presenting our definition of ICT services, we can further describe our approach to service-based IT management, the SPSPR model.

The preliminary concept of the SPSPR model was introduced in 2001 by Vorisek and Dunn [16] and was further developed by Vorisek et al. [17]. The model itself can be used for solution of the relationship between enterprise process management and IT management. Especially, it is distinctive in its hierarchical approach to the organizational architecture, where it uses five interacting layers.

These layers are namely:

S Strategy,
P Business Processes,
S ICT Services,
P ICT Processes,
R ICT Resources.

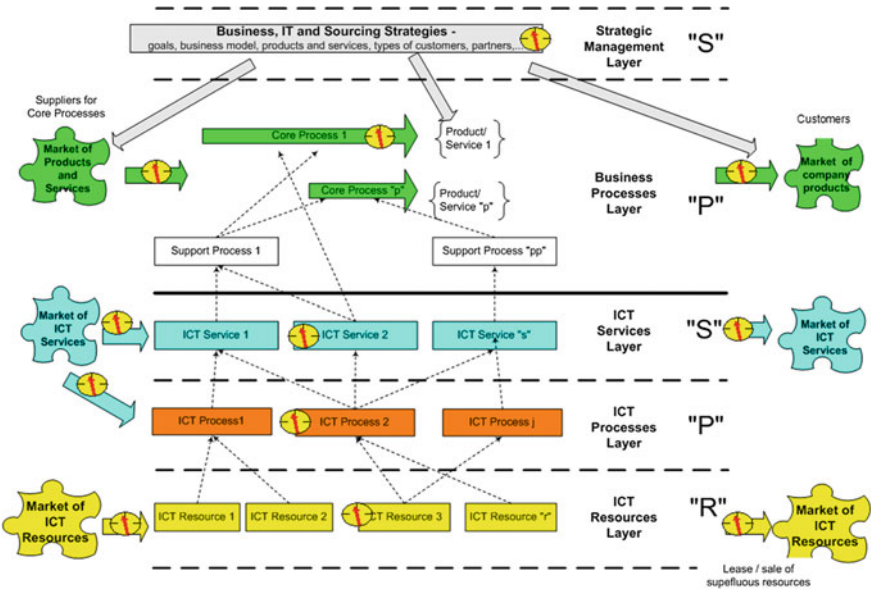


Fig. 134.1 The SPSPR model

The hierarchy and particular layers can be seen in Fig. 134.1 where we present the basic concept of the SPSPR model. Its primary aim is to ensure effective IT-business alignment while maintaining integrity of the global architectural solution. Essentially, the decisions about organizational architecture should be connected to a certain business and IT strategies that should influence other architectural decisions. While defining and managing business processes, it also provides their interconnection with the relative ICT services.

The role of ICT services in this model is significant, because they are defined as an interface between business and IT departments (in Fig. 134.1 this interface is represented by the bold line). Important aspect is that the ICT services support individual business processes or their partial activities.

ICT service may be delivered internally or externally. In case of internal service provision, the IT department has to provide both ICT processes for service delivery as well as all the necessary ICT resources (hardware, software, data, and people). In other words, it has to provide ICT resources needed for execution of the particular ICT process or processes. External service provision usually means buying the service on particular ICT service market. In this case, the provision and management of ICT resources and ICT processes needed for the particular ICT service delivery is under the responsibility of an external service provider.

Bring it either choice of service provision, there exist certain possibilities to measure performance, quality, volume, efficiency, and further characteristics. Within the model, we have also indicated locations suitable for measuring these

characteristics using clock images. These measurements support decision-making concerning the service, process, and resource outsourcing.

In the SPSPR model, ICT processes support ICT services. Unlike business processes, these ICT processes are not directly connected with the business, are rather connected with ICT operations, and utilisation of the ICT resources. Examples could be specific methods used within the IT management field such as ITIL or COBIT. Subsequently, ICT resources are a base layer that is directly connected to ICT processes. These resources might be hardware, software, data, infrastructures, but also employees operating particular ICT process.

134.3.1 Types of ICT Services in the SPSPR Model

In order to use the model in practice, an organization needs to be aware of the existence of various ICT services. Due to the fact that the model aims mainly on the business/IT alignment, only certain ICT services can be considered to be directly involved within the model.

In this regard, we can identify various classifications of ICT services. A taxonomy that we deem interesting and that covers a number of aspects is the one introduced by Jelinek et al. [18]. They distinguish five different classifications based on: service content, methods of service consumption, types of service consumer, types of service provider, and on required resources and knowledge.

For SPSPR model, we consider the classification based on service content the most crucial. As we have expressed earlier the main aim of SPSPR model is an effective management of business/IT alignment. The importance of the mentioned service content classification is caused by its orientation towards the actual thing that is delivered by service provider to service consumer, and also towards relationship of delivered service to the business of the consumer. From the viewpoint of service content, the ICT services can be from our viewpoint classified into two categories: Business oriented ICT services, and ICT development oriented ICT services.

- *Business oriented ICT services* are services that directly support business processes and end-users. This fact is emphasized in the SPSPR model with the interconnections between business processes and ICT services. This category includes various service types, which we briefly discuss further in this section.
- *ICT development oriented ICT services* are services used for advancement of current ICT services or for development of new ICT services and are not directly consumed by the business. This category includes a number of services, especially software development, application implementation, integration, technology infrastructure advancement, and consulting. Their full inclusion within the model can be spotted specifically in case of software development organizations as in that case their main aim is to support the main (core) business processes of such organizations. However, in case of customer

organization, role of these services can be considered peripheral and their inclusion in the model should be therefore only partial.

The SPSPR model is capable of operating a variety of ICT services from both categories. However, the primary orientation of the SPSPR model is on the business side of ICT (business/IT alignment), and therefore the "Business oriented ICT services" can be considered the most important in this particular case. As we already mentioned, these services can be further divided into several groups: information services, application services, infrastructure services, supporting services, and mixed services.

- Information services

Service provider delivers specific information to the service consumer. Information should be delivered in a required structure, data format, and an agreed time (or time intervals). The received information could be subsequently used by the customer organization in its business processes that require information to support particular decisions. The service provider is responsible for information quality such as information relevant to a certain business process. An example of such service could be current stock price, exchange rate, weather forecast, map retrieval, image recognition, or legal case search. Specific features that can be identified are mainly: potential service replication with low cost, existence of service legal restrictions (i.e. Intellectual property law) for the service consumer, irrelevance of software application by which it is delivered and aim on the service content.

- Application services

Service provider delivers specific functionality of business application. Such applications could be accounting, asset management, customer relationship management, business intelligence, human resource management, or enterprise content management. There exist three models of data ownership; Data processed by the application may be owned by the organization, by the provider, or bear a combined ownership model. Generally, the service provider is responsible for data transformation provided by the application functionality, whereas the data owner is responsible for the quality of the input data. The application service may support one or more activities of business process depending on the service granularity and related functionality. Application service may often be delivered as a whole together with supporting services such as user training, service support, or possible customisation. When we link this category to the cloud computing classification presented by Mell and Grance [19], the application services are represented by the Software-as-a-Service (SaaS) concept.

- Infrastructure services

Service provider delivers infrastructure required for application processing and its proper operation. The delivery usually consists of implementation and delivery of the ICT infrastructure. Such infrastructure can be for instance servers, networks,

operating systems, or databases. This category is also mentioned by Ross and Weill [20], where they also further divide this category into several groups. Essentially, we may identify several subcategories. One of them could be services ensuring technological resources administration that according to the mentioned authors ensure provisioning and administration of end-user devices, and operation of platform for new business application development and implementation. Another group could be from our viewpoint a unified subcategory of communication services that typically include management and integration of all electronic communication channels, and provisioning of communication and networks between places and/or applications. Specific subcategories are data administration services, which provide an environment for data management[6] independent of applications, and risk/security management services that are used to ensure specified levels of information security.

Example of infrastructure services within the mentioned cloud computing classification as mentioned by Mell and Grance [19] could be the Platform-as-a-Service (PaaS) concept, and Infrastructure-as-a-Service (IaaS). Specifically the PaaS concept includes services together with development platform and integration tools, whereas IaaS concept includes services without development platform and integration tools.

- Supporting services

Activities performed by the service provider in order to support information, application, and infrastructure services can be defined as supporting services. They include namely user training, implementation, application customisation, integration, help desk services, consulting services concerning service design, or service contract generation.

- Mixed services

Often the above-mentioned services are tightly connected in practice, resulting in mixed services. Examples could be for instance specific cloud computing environments providing application service for activities such as data administration, or integration of the services, together with the necessary ICT infrastructure.

134.3.2 Role of Outsourcing in the SPSPR Model

As we have presented in this section, the SPSPR model represents a hierarchical approach to organizational architecture within the enterprise with emphasis on business/IT alignment. In its context, we have outlined a content-based

[6] Such as accessibility, storage, archiving, replication, and restoration after failure.

classification of ICT services, which illustrates various service types that can be used within the model.

These services can be often provided internally and externally. In relation with the ICT service provisioning, the SPSPR model could be further extended by adding the sourcing dimension. Although ICT services are an important part of the whole organizational architecture, the sourcing decisions take place on various layers within the presented hierarchy—namely on the layers of processes, ICT services, ICT processes, and ICT resources.

Therefore, it is important to outline existing outsourcing areas and their relation to the SPSPR model. We further discuss these areas within the next section.

134.4 Outsourcing Areas

This section is devoted to the actual analysis of outsourcing areas and their connection with the aforementioned service-based IT management model. As we have presented earlier, one of the key questions that an organization has to ask regarding its organizational architecture is the sourcing of particular model components. Due to the fact that the SPSPR model contains specific layers that may not be directly related to the contemporary outsourcing research areas, we need to outline possible mapping of these areas to the model. This mapping should allow a full utilization of outsourcing research results and methods within the SPSPR model and its particular layers.

From our viewpoint, we may differentiate three general areas of outsourcing:

- *Business Process Outsourcing* (*BPO*),
- *Information Technology Outsourcing* (*ITO*),
- *Software Development Outsourcing* (*SDO*).

This classification especially emphasizes a separation of organizational IT and its business processes, which is often used within various Enterprise Architecture frameworks. While the BPO is more common within large organizations, the ITO could be considered as the most enticing area for contemporary SMEs. This is mostly thanks to the recent emergence of cloud computing technologies and the related Software-as-a-Service (SaaS) concept.

Particular areas are described in more detail within the following subsections. Within these subsections, we also discuss the position of individual outsourcing areas in the SPSPR model.

134.4.1 Business Process Outsourcing

Although Business Process Outsourcing is widely used within both academia and practice, it is hard to find its exact definition. According to Gartner glossary [21],

the Business Process Outsourcing could be defined as "a delegation of one or more business processes to an external provider that, in turn, owns, administers, and manages the selected processes." A suitable definition of business process was provided by Davenport and Short [22] where they define it as "a set of logically related tasks performed to achieve a defined business outcome." In our view, possible business processes are not limited to IT-backed processes, since there can exist a process which operates without an IT support. For instance, Sparrow [23] also supports this idea. On the other hand, he also mentions that these processes are interconnected to a core process, which is usually backed by some sort of computer system.

The fact of interconnection is a very important aspect of the SPSPR. When deciding about possible business process outsourcing, it is necessary to inspect the existence of any major interconnections with ICT services or ICT resources that are otherwise needed to remain internal. Using the SPSPR model presented in this article, an organization can possibly map the system to an extent where these interconnections become evident. The organization is able to avoid an unfavourable situation of outsourcing a process that is possibly highly integrated in the company. Similar features and assistance can be also provided by most of the currently available Enterprise Architecture approaches.

As apparent, this area of sourcing aims on the business process layer in the SPSPR model. When the organization outsources its whole process, the related ICT services, ICT processes, and ICT resources, if not connected to other processes that remain within the organization, often become unnecessary. This is important especially during the decision-making about possible sourcing solutions as mentioned earlier; The costs associated either with running partially unused servers or potential liquidation of the ICT resources need to be included in the calculation.

134.4.2 Information Technology Outsourcing

The Information Technology Outsourcing (ITO) can be from our viewpoint defined in accordance with Loh and Venkatraman [24] who define IT outsourcing as "a significant contribution by external vendors of the physical and/or human resources associated with the entire or specific components of the IT infrastructure in the user organization."

Its difference from BPO is that in this case, only the IT function is outsourced while the business process remains in the organization.

Many further taxonomies exist, for instance Vorisek et al. [25] propose that this sourcing type can be further divided into partial IT outsourcing and complex IT outsourcing. With partial IT outsourcing the organization outsources an ICT service, ICT process, or ICT resource. Complex IT outsourcing then means total outsourcing of organizational IT, which essentially means outsourcing of most of the organizational ICT services, ICT processes, and ICT resources. According to

Kern et al. [26], IT outsourcing could be further categorized into various forms based on resource utilization as follows:

- *Insourcing*—Using internal resources under internal management.
- *Buy-in*—Bringing in external resources to run under in-house control.
- *Traditional outsourcing*—Supplier taking ownership of customer resources and managing those resources on behalf of the customer.
- *ASP*—Renting supplier-owned resources to customers and delivering over the Internet.

When we have a look at the SPSPR model, this area of sourcing mainly aims on the ICT services layer. However, decisions within this layer also indirectly influence ICT processes and ICT resources layers. Given the forms mentioned by Kern et al. [26], in case of insourcing and buy-in forms, ICT resources are located in-house, whereas in case of traditional outsourcing and ASP, the related ICT resources are located externally. This also means that unless these resources are used by other ICT process and/or service, they do not need to be mapped in the SPSPR model.

134.4.3 Software Development Outsourcing

The Software Development Outsourcing (SDO) is essentially a separate category of outsourcing. Using a managerial lens on this term, we see it as an alternative to in-house development that can be possibly combined with the areas BPO and ITO. It also shares some of the characteristics with both of the mentioned, such as the need of precise contractual terms.

When we talk about customer or end-user organizations: instead of buying the application service as a ready-made solution, the organization could develop the application in-house (in case when it has its own IT department), or use the mentioned Software Development Outsourcing, or a combination of these approaches.

A situation when the SDO is typically applied is the case of an organization acting as a service provider, a system integrator, or a service broker. Use of the Software Development Outsourcing overall emerges more often on the side providing the software or application service than on the side of a pure customer organization. One reason could be that managing such development needs a significant amount of experience and knowledge of relevant methods. As mentioned by Richmond and Seidmann [27], the delivery task is partitioned into two consecutive stages: system design and software development. The parties can contract for each stage separately or specify an initial contract that covers both stages.

The role of this sourcing area within the SPSPR approach lies mostly in insight to the cost calculations connected with in-house and bespoke ICT solutions. The impact of this sourcing area within SPSPR is therefore mainly in ICT resource layer and partially in ICT process layers. Its results then appear usually in the ICT services layer.

134.5 Conclusion

Service-based IT management can be identified as one of the trends of future management of information technology. One of its specifics is the orientation towards business requirements with regard to various technological aspects. Its main potential benefit therefore is the use of services as a mediator between business and IT, where it ensures effective business/IT alignment.

The currently used approaches based on services, such as Information Technology Infrastructure Library (ITIL) or some of the Enterprise Architecture (EA) frameworks, are however very demanding on resources, broadly oriented, and generally unsuitable for small and medium-sized enterprises (SMEs).

General lack of methods and tools forms a significant gap in both academic and practitioner research. Ideally should be these methods available and accessible across a variety of business segments and industries.

We have decided to propose a lightweight service-based IT management solution suitable for a variety of business segments including small and medium enterprises (SMEs). Moreover, the related problem of ICT services outsourcing is reflected by mapping various outsourcing areas onto the proposed solution. The result is that an organization is able to analyze its organizational architecture including processes, ICT services, ICT processes, and ICT resources, then connect these objects with its business and IT strategy, and consequently identify layers affected by the potential use of outsourcing.

The paper can be divided into three parts.

In the first part, we have discussed an integral part of every service-based IT management approach, the ICT services concept. Due to the fact that a universal definition of ICT service does not exist, we have provided a selection of some interesting viewpoints, and subsequently proposed a definition that covers our perception of problem.

In the second part of the paper, we have presented our approach to service-based IT management, the SPSPR model. This model is built by using hierarchical design and works with various layers/objects and their interconnections. Its main goal is to assist with aligning business and IT within the whole organization. An important aspect that we have emphasized is that the SPSPR clearly defines the role of information technology (ICT services) as supportive of a particular business process or business processes. The mentioned ICT services layer acts as a mediator that facilitates proper business/IT alignment. Services included in this layer may be of various types, while not all services within the organization might be suitable for this model. We have therefore summarized some of the existing ICT service categorization. Subsequently, we have examined the content-based categorization, and identified service types that might be used within the SPSPR model.

Finally, in the third part of the paper we have presented our view on the outsourcing areas and outlined their relation to the presented SPSPR model and its layers. The existing outsourcing areas of Business Process Outsourcing (BPO), Information Technology Outsourcing (ITO), and Software Development

Outsourcing (SDO) all have different contexts that need to be reflected within the current IT management methods. Mapping the areas indicated that they are distributed among various layers of the model. Especially, the ITO can be regarded as closely connected with the identified trend of service-based IT management. This is due to the fact that this area of sourcing mainly aims on the ICT services layer. However, the identified relations in the outsourcing areas enable us to understand the broader context of outsourcing decisions.

In the future research, we plan a further development and testing of the presented model. In addition, a series of experimental applications and case studies is currently scheduled in the segment of small and medium enterprises. Special emphasis is planned to be put on the outsourcing areas and their possibilities.

Acknowledgments This contribution was supported by a research grant of the Czech Science Foundation GACR P403/10/0092.

References

1. Henderson, J.C., Venkatraman, N.: Strategic alignment: leveraging information technology for transforming organizations. IBM Syst. J. **32**(1), 4 (1993)
2. Chan, Y.E., Sabherwal, R., Thatcher, J.B.: Antecedents and outcomes of strategic IS alignment: an empirical investigation. IEEE Trans. Eng. Manag. **53**(1), 27–47 (2006)
3. Duncan, N.B.: Capturing flexibility of information technology infrastructure: a study of resource characteristics and their measure. J. Manag. Inf. Syst. **12**(2), 37–58 (1995)
4. van Bon, J., et al.: IT Service Management: An Introduction Based on ISO 20000 & ITIL V3. Van Haren Publishing, Zaltbommel (2007)
5. ITIL Lifecycle Publication Suite Books: Version 3. Office of Government Commerce. The Stationery Office, London (2007)
6. TOGAF: The Open Group Architecture Framework. The Open Group (2009)
7. DoDAF Architecture Framework Version 2.02. U.S. Department of Defense (2010)
8. FEA Practice Guidance. Federal Enterprise Architecture Program Management Office, OMB (2007)
9. Lovelock, C.H.: Classifying services to gain strategic marketing insights. J. Mark. **10**(1), 9–20 (1983)
10. Kotler, P.: Marketing Management: Analysis, Planning, Implementation and Control, 6th edn. Prentice Hall, Englewood Cliffs (1988)
11. Booth, D., et al.: Web Services Architecture. W3C Working Group Note (2004)
12. Vorisek, J.: Jak clenit informaticke sluzby a navrhovat jejich architekturu. In: Proceedings of the 16th international conference on systems integration, Prague, Czech Republic (2008)
13. Skala, J.: Best Practice rizeni ICT sluzeb a ICT infrastruktury. In: ICTM2005 conference proceedings, Prague, Czech Republic (2005)
14. Gala, L., Pour, J., Toman, P.: Podnikova informatika. Grada Publishing, Prague (2006)
15. Schekkerman, J.: How to survive in the jungle of Enterprise Architecture Frameworks, 2nd edn. Trafford Publishing, Bloomington (2003)
16. Vorisek, J., Dunn, D.: Management of business informatics—opportunities, threats, solutions. In: Proceedings of the 9th conference on systems integration, Prague, Czech Republic (2001)
17. Vorisek, J., Jandos, J., Feuerlicht, J.: SPSPR model—framework for ICT services management. J. Syst. Integr. **2**(2), 3–10 (2011)

18. Jelinek, P., Sild, V., Vorisek, J.: Information services classification and architecture. Systemova integrace **14**(3), 7–23 (2007)
19. Mell, P., Grance, T.: The NIST Definition of Cloud Computing. National Institute of Standards and Technology, Gaithersburg (NIST Special Publication), pp. 800–145 (2011)
20. Ross, W., Weill, P.: Six IT decisions your IT people shoulnd't make. Harv. Bus. Rev. **80**(11), 84–91 (2002)
21. Gartner IT Glossary, Sept. 23, Gartner.com (2012)
22. Davenport, T.H., Short, J.E.: The new industrial engineering: information technology and business process redesign. Sloan Manag. Rev. **31**(4), 11–27 (1990)
23. Sparrow, E.: Successful IT Outsourcing: From Choosing a Provider to Managing the Project. Springer, London (2003)
24. Loh, L., Venkatraman, N.: Diffusion of information technology outsourcing: influence sources and the Kodak effect. Inf. Syst. Res. **3**(4), 334–358 (1992)
25. Vorisek, J., et al.: Principy a modely rizeni podnikove informatiky. Oeconomica, Prague (2008)
26. Kern, T., Willcocks, L., Lacity, M.C.: Application service provision: risk assessment and mitigation. MIS Q. Executive **2**(1), 113–126 (2002)
27. Richmond, W.B., Seidmann, A.: Software development outsourcing contract: structure and business value. J. Manag. Inf. Syst. **10**(1), 57–72 (1993)

Chapter 135
An Empirical Study of Customers' Intentions by Using Logistics Information Systems (LIS)

Yu Liu

Abstract In order to gain new insights into the determinants of behavioral intention to use LIS, this paper proposes a theoretical model that augments the technology acceptance model (TAM) with three new constructs: trust, flow experience and social influence. The paper examines issues related to behavioral intention to use LIS from the viewpoint of customers. Also within the model framework, the paper investigates the effect of behavioral attitude on behavioral intention. To test the model, structural equation modeling is employed to analyze data collected from 248 respondents in Korea. Empirical results show that consumer's attitude in LISs' use is determined by technology and social factors. There is also a significant impact of customer's attitude on customer's intention. This research provides a theoretical foundation for academics and also practical guidelines for logistics service providers in dealing with LIS aspects.

Keywords Logistics information system (LIS) · Technology acceptance model (TAM) · Trust · Flow · Experience · Social influence

135.1 Introduction

135.1.1 Research Background and Purpose

Logistics information system (LIS) is becoming more and more important as it provides efficient and effective logistics management that aims to reduce cost and cycle time for its customers on the supply chain. However, many small and medium-sized logistics service providers still focus on internal operation performance but

Y. Liu (✉)
The School of Finance and Economics, Tibet University for Nationality, Xian Yang, China
e-mail: ly710@hotmail.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_135,

lack the vision of collaborating with other supply chain participants on improving overall supply chain performance. Moreover, these logistics service providers (LSPs) are unable to synchronize information with trading partners in real time for making timely decision or providing responsive services. To support the different kinds of supply chain, logistics information systems (LISs) are organized rapidly to adapt the requirement of the response in the logistics enterprise according to the current customer requirements. The rapid development of LIS calls for an investigation to discover what key factors motivate consumers to use LIS.

Therefore, the main purpose of this study is to develop and empirically test a theoretical model of the determinants of intention to use LIS. The proposed model integrates trust, social psychology, and flow experience into the theory of TAM. TAM suggests that behavioral intention is a function of an individual's attitude toward the behavior. This work applied a structure equation model (SEM) to assess the empirical strength of the relationships in the proposed model.

Our research uses Korea as the site of the empirical investigation because the supporting infrastructure required for LIS development has been put in place. According to the annual report of EC published by the Korea Ministry of Commerce in 2007, the total EC market size in Korea was USD 507.42 billion with a growth of 34.6 % compared to the previous year [1]. Since Korea is the world's second-fastest-growing IT market, LIS will play an important role in executing wide-ranging activities and actively confronting changing logistics conditions.

135.1.2 Research Methodologies

Measurement assessments are used to validate our model. Following the recommendations of prior studies for developing and validating measurement instruments, our study conducts a three-stage procedure. The first stage is conducted through a review of the relevant literature and corresponding scales. In stage two, a set of sample items is generated for each construct and assessed for the reliability and content validity. In stage three, we precede with an extensive confirmatory analysis LIS by testing and validating the refined scales for the reliability and construct validity. We also verify convergent validity and the goodness-of-fit of our research model.

135.2 Theoretical Backgrounds

135.2.1 Overview: Background for LIS

Logistics is the process of strategically managing the acquisition, movement and storage of materials, parts and finished inventory (and the related information flows)

through the organization and its marketing channel in such a way that current and future profitability is maximized through the cost-effective fulfillment of orders.

135.2.2 Review of the Literature on Behavioral Intention Issues in LIS

135.2.2.1 Technology-Oriented Influence

TAM has received considerable attention of researchers in the information system field over the past decade. The TAM was first developed by Davis to explain user acceptance of technology in the workplace [2]. Davis adapted the TRA by developing two key sets of constructs that specifically account for technology usage: (1) Perceived Usefulness and Perceived Ease of Use, and user's attitude, behavioral intentions and actual computer usage behavior.

135.2.2.2 Trust-Oriented Influence

Consumer trust is a human viewpoint toward e-commerce and part of the human aspect of information systems. Researchers believe that trust is the foundation of e-commerce. Based on an investigation of multiple disciplines, Mayer et al. (1995) have defined trust as a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another [3]. In this paper, trust is approached based on this definition.

135.2.2.3 Flow-Oriented Influence

Flow has been studied and identified as a possible measure of on-line user experience. This definition suggests that flow consists of four components—control, attention, curiosity, and intrinsic interest. Considered LIS as Task-oriented information technology, here, flow is defined as an involved experience, in e-commerce with control, attention focus, curiosity and intrinsic interest.

135.2.2.4 Social-Oriented Influence

Social factors profoundly impact user behavior. Kelman's study of social influence was motivated by his interest in understanding the changes brought in individuals' attitude by external inputs, such as information communicated to them. Kelman distinguished between three different processes of social influence that affect individual behavior: compliance, identification, and internalization. [4]

According to the literature review, we can categorize the factors that influence consumers' attitudes in the use of LIS into four perspectives: technology, trust, flow experience, social influence (Fig. 135.1).

135.3 Research Model and Hypotheses

Based on the extended TAM (Fig. 135.1), the backbone of the conceptual model is followed as Fig. 135.2. Hypothesized relationships were proposed in the context of LIS.

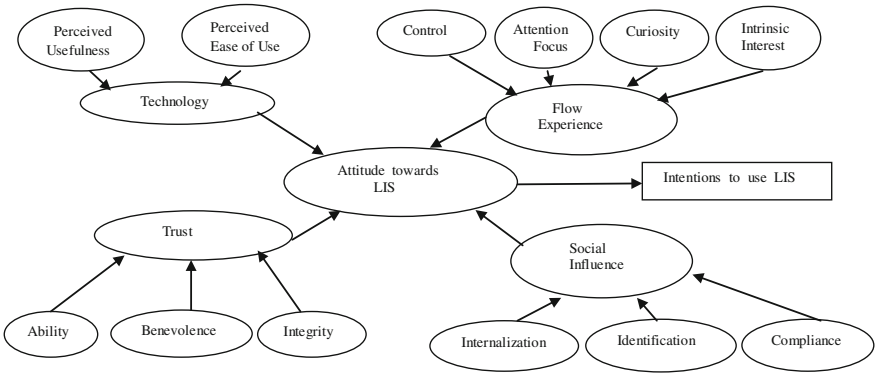


Fig. 135.1 The extended TAM

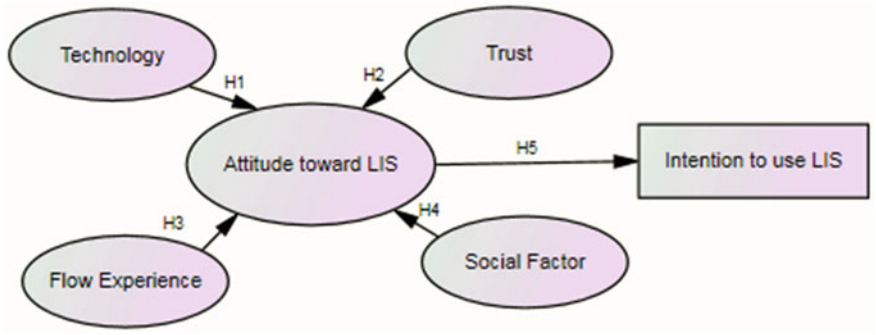


Fig. 135.2 The conceptual model

135.4 Empirical Analysis and Result

135.4.1 Data Collection

The survey respondents for this study were recruited from November 2011 to January 2012 and the participants were solicited by distributing questionnaire via email. Altogether, 245 questionnaires were collected by personal visits and email. 13 questionnaires were eliminated due to invalid answers, leaving 232 questionnaires for our empirical analysis (a response rate of 23.2 %).

135.4.2 Reliability and Validity Tests

135.4.2.1 Reliability Test

Reliability is determined by Cronbach's alpha, a popular method for measuring reliability [5]. Nunnally (1978) suggests that for any research at its early stage, a reliability score or alpha that is 0.60 or above is sufficient [6] (Table 135.1).

Table 135.1 Reliability coefficient test

Variables+[3]	Number+[3]	Alpha+[3]
Technology+[3]	6+[3]	0.818+[3]
Trust+[3]	6+[3]	0.809+[3]
Flow experience+[3]	6+[3]	0.841+[3]
Social factors+[3]	6+[3]	0.880+[3]
Attitude+[3]	2+[3]	0.650+[3]
Intention+[3]	2+[3]	0.672+[3]
Average+[3]		0.792+[3]

Note n = 248

135.4.2.2 Validity Test

Twenty-eight survey items in the questionnaire were relevant to factor analysis. To determine the underlying structure, the correlation matrix was initially examined to determine how appropriate it was for factor analysis. As shown in Table 135.2, we concluded that the data were approximately multivariate normal data. Furthermore, the correlation matrix contained sufficient covariation for factoring.

Based on the Screen test and the Eigen values that were greater than one, six factors were accepted as interpretable factors. These factors accounted for 60.05 % of the variance. Table 135.3 shows the results of our factor analysis.

Table 135.2 KMO and Bartlett's test

Kaiser-Meyer-Olkin measure of sampling adequacy		0.749
Bartlett's test of sphericity	Approximate Chi Square	2791.161
	df	378
	Significance	0.000

Table 135.3 Factors analysis-rotated component matrix

Items	Components					
	1	2	3	4	5	6
Trust 1	−772					
Trust 2	−646					
Trust 3	−602					
Trust 4	−705					
Trust 5	−692					
Trust 6	−798					
Technology 1		−718				
Technology 2		−756				
Technology 3		−708				
Technology 4		−729				
Technology 5		−633				
Technology 6		−762				
Inflow 1			−818			
Inflow 2			−845			
Inflow 3			−344			
Inflow 4			−828			
Inflow 5			−741			
Inflow 6			−786			
Social 1				−637		
Social 2				−736		
Social 3				−810		
Social 4				−880		
Social 5				−819		
Social 6				−802		
Attitude 1					−774	
Attitude 2					−600	
Intention 1						−597
Intention 2						−766
Eigen value	4.730	3.687	3.261	2.563	1.487	1.229
% of variance	16.893	13.167	11.646	9.154	5.312	4.391
Cumulative %	16.893	30.059	41.705	50.859	56.171	60.362

Extraction Method: Principal Component Analysis
Rotation Method: Varimax with Kaiser Normalization

135.4.3 Structural Equation Modeling

As suggested in the literature, the model fit is assessed by indices such as the Comparative Fit Index (CFI), the Goodness of Fit Index, and the Root Mean Square Error of Approximation (RMSEA) [7] (Table 135.4).

Table 135.4 Analysis of model-fit

Indices in SEM analysis	Default model	Fit the model	Fit standard
Chi square/DF	826.509/343	Good fit	<3
RMR	0.054	Good Fit	<0.08
GFI	0.805	Not good fit	>0.90
CFI	0.810	Not Good fit	>0.90
RMSEA	0.078	Good fit	<0.08

135.4.4 Hypotheses-Path Testing

This section presents the statistical results of the measurement validation and hypothesis testing. The effects of technology, trust, social factors and flow experience on customers' attitude and intention in LIS were assessed through AMOS 6.0. Our empirical results are shown in Table 135.5.

Table 135.5 Path coefficients and regression weights

Relations	Estimate	S.E.	C.R.	P-Value
Attitude<—Technology	0.320	0.078	4.081	***
Attitude<—Trust	0.102	0.081	1.263	*0.207*
Attitude<—Flow	0.076	0.119	0.637	0.524
Attitude<—Social	0.216	0.062	0.3466	***
Intention<—Attitude	0.136	332	3.419	***

*P < 0.1; **P < 0.05; ***P < 0.01

Overall, the path coefficients of H1, H4, and H5 were significant at a level of $p < 0.01$, thereby indicating support for these hypotheses. Hypothesis 2 and 3 are not supported. Figure 135.3 shows a summary of our results for each hypothesis in the research model.

135.5 Conclusions and Implications

This paper examines usage intention issues in the context of LIS from the viewpoint of consumers. For academic research, this study contributes to a theoretical understanding of the factors that promote task-oriented IT usage such as LIS. The

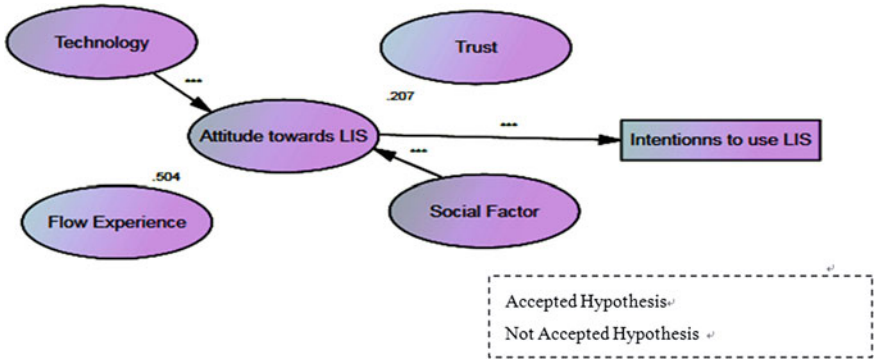


Fig. 135.3 Output path diagram of the research model

research presented the implications for future research: First, while most of past studies found consistently technology an important predictor in TAM model, the author has found that this is not always true. LIS is a task technology, which is different from a problem-solving technology. As a result, the role of technology and social influence will become important in behavioral intentions to use LIS. Second, social influences may play an important role. Users intend to use task technology continuously where they are completely and totally immersed. Increasing usability through dialogue and social interaction, access, and navigation, is the key to successful management of on-line communities.

For practice, this study has key implications. First, the findings suggest that technology is an important antecedent to both consumers' attitude and intention, implying practitioners must consider the element of technology if they are to provide users with attractive LIS. Technology is a form of extrinsic motivation. When users achieve technology for LIS, they are more likely to have positive attitudes toward using LIS and, most important, they will be motivated to return frequently. Second, social influences also have the strong impact on consumers' attitude toward LIS. Interpersonal interaction among LIS' users creates a community in which business value can be created by improving customer loyalty. When users use LIS intensely, the interaction with other users will cause more to join in. The more users in LIS, the more user-generated experience it is likely to exchange and thus the more users it will attract. This idea, called the dynamic loop.

In conclusion, this study was conducted to examine factors influencing consumers' attitude in using LIS. The research model and hypotheses are based on TAM and prior literature on trust, social influence and flow experience. The results of this study confirm the important roles of technology and social influences in predicting consumers' attitude toward LIS. Also the results emphasize the significant impact of technology and social influences on consumers' attitude toward using LIS. The insignificance of the link from trust and flow experience to consumers' attitude indicates the need for further research on flow experience in the context of LIS.

This study is not free from limitations. Firstly, although the research comes up with some significant findings from the viewpoint of consumers, it does not include all the factors that affect consumers' use of LIS. For example, factors, such as specific LIS functions and social and individual factors can be taken into consideration in future research. Special attention should also be paid to human factors, management, education, awareness, and other non-technology factors in order to prevent security risks. Secondly, all the participants in the samples in our research had experience in LIS use. Pre-interaction factors, such as brand reputation, advice, or experience from trusted sources of information (e.g. word of mouth and traditional media) were not considered in our research model. It would be interesting for further research to focus on other factors that give more detailed information on consumers' intention to use LIS. Finally, the subjects are self-selected and only those users with experience answered the questionnaires, a condition that may potentially limit the applicability of our research findings in other settings or populations. Therefore, additional research is required to examine the generalizability of the model and its findings to a wide array of settings and populations. Considering these limitations, the research constitutes an important stepping-stone for future research in different national settings in which it involves an investigation of the factors that influence intentions to use LIS.

Acknowledgments The project number is 12myQP06.

References

1. Yoon, S.: The antecedents and consequences of trust in online-purchase decisions. J. Interact. Mark. **6**(2), 36–45 (2002)
2. Davis, F.D., Bagozzi, R.P., Warshaw, P.R.: User acceptance of computer technology: a comparison of two theoretical models. Manag. Sci. **35**, 982–1003 (1989)
3. Mayer, R.G., Davis, J.H., Schoolman, F.D.: An integrative model of organizational trust. Acad. Manag. Rev. **20**(3), 709–734 (1995)
4. Kelman, H.C.: Compliance, identification, and internalization: three processes of attitude change. J. Conflict Resolut. **2**, 51–60 (1958)
5. Mukherjee, A., Nath, P.: A model of trust in online relationship banking. Int. J. Bank Mark., pp. 5–15 (2003)
6. Nunnally, J.C.: Psychometric theory. McGraw-Hill, New York, pp. 23–45 (1978)
7. Steiger, J.H.: Structural model evaluation and modification: an interval estimation approach. Multivar. Behav. Res. **25**, 173–180 (1990)

Chapter 136
Measurement of Gender Segregation in Chinese Industry

Dingquan Yang, Zongwei Xu and Lihua Ma

Abstract In order to measure the degree and trend of gender segregation in Chinese industry, the paper uses five occupational segregation measure methods, D index, Ds index, Ip index, Square Root index and A index, to measure Chinese urban units' industry gender segregation from 2000 to 2010, and to obtain gender segregation index of Chinese urban units industry. Studies have shown that Chinese industry gender segregation degree is far below the world average, gender segregation of Chinese industries is on the rise, and is becoming larger. The industries of the highest gender segregation in China are construction, manufacturing, education, public management and social organization, sanitation, social security and social welfare. The lowest industries are information transfer, computer and software, resident services and other services, culture, sports and entertainment, agriculture, forestry, farming of animals and fishery. It shows that gender segregation is more easily formed in the gender-dominated industry.

Keywords Gender segregation · Industry gender segregation · Occupational segregation · Measure

136.1 Introduction

Occupational segregation is a product of social and economic development to a certain stage, which is the first concern, and the highest attention is occupational gender segregation [1], because gender segregation is the most direct and most

D. Yang (✉) · Z. Xu
School of Management, Hefei University of Technology, Hefei, China
e-mail: yangdingquan@163.com

D. Yang · L. Ma
Management Department, Guangxi University of Technology, Liuzhou, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_136,

external manifestations in occupational segregation. Serious occupational gender segregation will lead to social injustice and inefficient allocation of resources, exacerbate hierarchies curing and aggravate the tense social and labor relations. At present, the primary methods of gender segregation measure is D index and Ds index in China, rarely dealing with the Ip index, O (x) index and A index. The study sample also mainly concentrated on certain years, departments and industries, so this single method and small sample research almost does not reflect the whole picture of China's occupational gender segregation, and final evaluation of the results of reliability and validity has been widely questioned. The paper is mainly based on the five main methods of measure, and it studies Chinese urban units' industry gender segregation from 2000 to 2010 in order to clarify the present situation of Chinese industry gender segregation, explore occupational gender segregation development trends of China, and propose solutions to the plight of response strategies.

136.2 Research Method and Data Sources

136.2.1 Research Method

Occupational gender segregation measure is an important part of the study of gender segregation, from the existing literature; there are some major occupational gender research methods.

136.2.1.1 D Index

D index is the most commonly used indicator of occupational gender segregation, which was put forward in 1955 by Duncan, and it is widely used in occupational gender segregation measure [2]. D index is calculated as follows:

$$D = \frac{1}{2}\sum_{i=1}^{j} |(W_i/W) - (M_i/M)| \tag{136.1}$$

In the Eq. (136.1), W_i: the number of women in occupation I, W: female employment, M_i: the number of men in occupation I, M: male employment, j: the total number of occupational. If the D index is equal to zero, it means that women and men have the same proportion of employment in different occupations. If the D index is equal to 1, it means that men and women completely are isolated in different occupations, that is to say, men and women are not in the same occupation. D index could be interpreted to eliminate gender differences between occupations, and women (or men) will want to change careers proportion.

136.2.1.2 Ds Index

Since D index does not take into account the inter-annual differences between samples in the occupational scale, the gender composition of the labour market and other questions, Gibbs and Gross put forward that D index should be corrected and proposed standardization of Duncan index (abbreviates Ds). Ds index is calculated as follows:

$$D_S = \frac{1}{2}\sum_{i=1}^{j}\left|\left[(W_i/T_i)/\sum_{i=1}^{j}(W_i/T_i)\right] - \left[(M_i/T_i)/\sum_{i=1}^{j}(M_i/T_i)\right]\right| \qquad (136.2)$$

In the Eq. (136.2), T_i: the total number of occupation i, $T_i = M_i + W_i$, definitions and explanations of the remaining variables are same as Eq. (136.1). Ds is not affected by the different occupational scale, and it eliminates occupational scale change with time values impact. Therefore, D_S index is widely applied in the use of time series data and the cross-phase study.

136.2.1.3 Ip Index

In order to further study occupational gender segregation, Karmel and Maclachian created the Ip index, and the formula is as follows:

$$I_P = \left(\frac{1}{T}\right)\sum_{i=1}^{j}|W_i - a(W_i + M_i)| = \left(\frac{1}{T}\right)\sum_{i=1}^{j}|(1-a)W_i - aM_i| \qquad (136.3)$$

In the Eq. (136.4), T: total employment, a: women account for the proportion of total employment, definitions and explanations of the remaining variables are same as Eqs. (136.1) and (136.2). Ip index shows that male and female ratio remained unchanged in the occupational structure and in the total employment structure, for the same as the proportion of the ratio of men and women in each occupation, what percentage of people need to replace their work [3].

136.2.1.4 A Index

In order to carry out research on occupational gender segregation in different countries, Charles and Grusky proposed multiplication model and associated index. The method is the improvement of the conventional sum index, which can measure the degree of occupational gender segregation and reveal the type of occupational segregation. Deduced A index is as follows:

$$A = \exp\left(\frac{1}{J}\sum_{j=1}^{J}\left\{\ln\left(\frac{W_{jk}}{M_{jk}}\right) - \left[\frac{1}{J}\sum_{j=1}^{J}\ln\left(\frac{W_{jk}}{M_{jk}}\right)\right]\right\}^2\right)\frac{1}{2} \qquad (136.4)$$

In the Eq. (136.4), i: gender, j: occupation, k: environment, definitions and explanations of the remaining variables are same as Eqs. (136.1) and (136.2). A index indicates that the gender ratio of a career deviates from the average degree of all occupational gender ratio. If there is no occupational segregation in all occupations, A index is equal to 1, the greater the occupational gender segregation, the greater A Index value is [4].

136.2.1.5 Square Root Index

Hutchens raised the square root of the index that can be decomposed. It can decompose different occupational segregation, and it can clearly understand the segregation degree between occupational categories and occupational groups within. Its formula is as follows:

$$O(x) = -\sum_{j=1}^{T} S_{2j}\left[\left(S_{1j} - S_{2j}\right)^{0.5} - 1\right] = 1 - \sum_{j=1}^{T} \sqrt{\left(S_{2j}\right)\left(S_{1j}\right)} \tag{136.5}$$

In the Eq. (136.5), S_{ij} (i = 1,2): women and men in the proportion of the j-th class career, T: the total number of occupational categories [5], definitions and explanations of the remaining variables are same as Eqs. (136.1) and (136.2). O(x) has decomposability and homogeneity.

136.2.2 Data Sources and Description

The study required data from 2000 to 2010, such as "China Labor Statistical Yearbook", data compiled by the 11 years between the 19 major industries (2000–2002 according to 16 industries). 2000–2002 China national economy according to the 16 industries classification: Farming, Forestry, Animal Husbandry and Fishery(S1), Mining and Quarrying(S2), Manufacturing(S3), Production and Supply of Electricity, Gas and Water(S4), Construction(S5), Geological Prospecting & Water Conservancy(S6), Transport, Storage, Post and Telecommunications(S7), Wholesale and Retail Trade & Catering Services(S8), Finance and Insurance(S9), Real Estate Trade(S10), Social Services(S11), Health Care, Sporting and Social Welfare(S12), Education, Culture and Arts, Radio, Film and Television(S13), Scientific Research and Polytechnical Services (S14), Government Agencies, Party Agencies and Social Organizations(S15), Others(S16). Since 2003 by 19 industry classification, Agriculture, Forestry, Farming of Animals and Fishery(S1), Mining(S2), Manufacturing(S3), Production and Distribution of Electricity, Gas and Water(S4), Construction(S5), Traffic, Transport, Storage and post(S6), Information Transfer, Computer and Software(S7), Wholesale and Retail Trade(S8), Accommodation and Restaurants(S9), Finance(S10), Real Estate(S11), Tenancy and Business Services(S12), Scientific Research, Technical Service and

Geologic Perambulation(S13), Management of Water Conservancy, Environment and Public Establishment(S14), Resident Services and Other Services(S15), Education(S16), Sanitation, Social Security and Social Welfare(S17), Culture, Sports and Entertainment(S18), Public Management and Social Organization(S19). This shows that numbers of statistical indicators are not fully consistent, and statistics need to pay attention to these changes [6].

136.3 The Results of Industry Gender Segregation Measure in Chinese Industry

The paper uses the above discussed five kinds of gender segregation measure, D index, Ds index, Ip index, O(x) index and A index method, to measure gender segregation from 2000 to 2010 in urban units of China (due to space limitations, the specific calculation process omitted). Calculation results in Tables 136.1, 136.2 and Fig. 136.1.

136.3.1 D Index Measure Results

D index average is 0.2088 from 2003 to 2010, the highest D index of industries are construction(S5), manufacturing (S3), education (S16), and D index are 0.0430, 0.0302 and 0.0298. The lowest D index of industries are information transfer, computer and software (S7), resident services and other services (S15), agriculture, forestry, farming of animals and fishery (S1), and D index are 0.0003, 0.0004 and 0.0008.

Table 136.1 2000–2010 industry gender segregation Index table in China's urban units

Years	D	Ds	Ip	O(x)	A
2000	0.1755	0.1743	0.0827	0.0225	1.0064
2001	0.1805	0.1786	0.0849	0.0239	1.0062
2002	0.2054	0.1884	0.0966	0.0498	1.0039
2003	0.1929	0.1777	0.0908	0.0345	1.0002
2004	0.1995	0.1830	0.0941	0.0295	1.0007
2005	0.2050	0.1815	0.0965	0.0318	1.0010
2006	0.2110	0.1855	0.0994	0.0338	1.0013
2007	0.2144	0.1899	0.1008	0.0354	1.0010
2008	0.2129	0.1964	0.0998	0.0355	1.0002
2009	0.2164	0.1947	0.1011	0.0374	1.0002
2010	0.2182	0.1991	0.1020	0.0389	1.0001

Table 136.2 2000–2010 industries gender segregation index table in China's urban units

Industry	2000–2002				2003–2010			
	D	Ds	Ip	A	D	Ds	Ip	A
S1	0.0013	0.0042	0.0006	1.0936	0.0008	0.0011	0.0004	1.1406
S2	0.0130	0.0143	0.0061	1.6542	0.0156	0.0185	0.0073	2.2017
S3	0.0364	0.0112	0.0171	1.0255	0.0302	0.0056	0.0142	1.0397
S4	0.0033	0.0058	0.0015	1.2902	0.0042	0.0086	0.0019	1.3762
S5	0.0300	0.0249	0.0141	2.8207	0.0430	0.0265	0.0202	4.3452
S6	0.0022	0.0124	0.0010	1.5459	0.0111	0.0113	0.0052	1.5141
S7	0.0119	0.0106	0.0056	1.4603	0.0003	0.0010	0.0001	1.1190
S8	0.0123	0.0124	0.0058	1.0164	0.0073	0.0084	0.0034	1.0169
S9	0.0042	0.0115	0.0020	1.0224	0.0057	0.0188	0.0027	1.0150
S10	0.0008	0.0030	0.0004	1.2166	0.0080	0.0125	0.0038	1.0021
S11	0.0044	0.0088	0.0021	1.0402	0.0012	0.0047	0.0005	1.2314
S12	0.0183	0.0290	0.0086	1.0411	0.0018	0.0047	0.0008	1.2304
S13	0.0206	0.0119	0.0097	1.0194	0.0028	0.0070	0.0013	1.3094
S14	0.0015	0.0040	0.0007	1.2420	0.0011	0.0033	0.0005	1.0651
S15	0.0270	0.0153	0.0127	1.7173	0.0004	0.0041	0.0002	1.0601
S16	0.0002	0.0000	0.0001	1.1313	0.0298	0.0123	0.0140	1.0022
S17					0.0215	0.0248	0.0101	1.0765
S18					0.0009	0.0042	0.0004	1.0547
S19					0.0239	0.0117	0.0112	1.5338

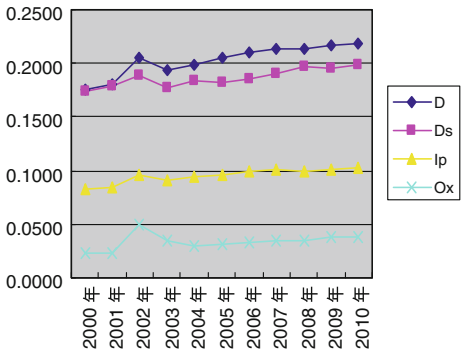


Fig. 136.1 2000–2010 industry gender segregation index trends in China's urban units

136.3.2 Ds Index Measure Results

Ds index average is 0.1885 from 2003 to 2010, the highest Ds index of industries are resident services and other services (S15), sanitation, social security and social welfare (S17), accommodation and restaurants, Ds index are 0.0265, 0.0248 and 0.0188. The lowest Ds index of industries are information, computer and software(S7), agriculture, forestry, farming of animals and fishery(S1), management of water conservancy, environment and public establishment(S14), Ds index are 0.0010, 0.0011 and 0.0033.

136.3.3 Ip Index Measure Results

Ip index average is 0.0981 from 2003 to 2010, the highest D index of industries are construction (S5), manufacturing (S3), education (S16), and Ip index are 0.0202, 0.0142 and 0.0140. The lowest Ip index of industries are information transfer, computer and software (S7), resident services and other services (S15), culture, sports and entertainment (S18), and Ip index are 0.0001, 0.0002 and 0.0004.

136.3.4 O(x) Index Measure Results

O(x) index average is 0.0346 from 2003 to 2010, the index is increasing year by year, and reached a high of 0.0389 in 2010. O(x) index is slow growth rate of annual average 0.0006.

136.3.5 A Index Measure Results

A index average is 1.0006 from 2003 to 2010, the highest A index of industries are construction (S5), mining(S2), public management and social organization(S19), which are 4.3452, 2.2017 and 1.5338 respectively. The lowest A index of industries are finance (S10), education (S16), accommodation and restaurants (S9), which are 1.0021, 1.0022 and 1.0150.

136.4 Discussion of the Results and Methods

Compared with overseas research, Chinese industry gender segregation showed the two significant features. Firstly, Chinese industry gender segregation degree is far below the world average; secondly, contrary to many countries of the world, Chinese industries gender segregation are on the rise, and becoming larger. The industries of the highest gender segregation in China are construction, manufacturing, education, public management and social organization, sanitation, social security and social welfare. The lowest industries are information transfer, computer and software, resident services and other services, culture, sports and entertainment, agriculture, forestry, farming of animals and fishery. It shows that gender segregation is more easily formed in the gender-dominated industry, which is affected by people's traditional concept of "gender roles".

Industry gender segregation is constrained by industry technical requirements, industry remuneration, industry fluidity, laws and other factors, and it will be widespread over the long term. Industry gender segregation will lead to serious

problems of resource allocation inefficiencies and social injustice; the rising industry gender segregation should attract the attention of the relevant departments of China.

From the actual results of the industry gender segregation measure, D index and Ip index measure the results and interpretation of the results has demonstrated a high degree of consistency, Ds index and O (x) index can play a complementary role. A Index does not seem to apply to the study of gender segregation in the inter-industry, its measure results are often contrary to the perception of the reality of the people. Recommend that the researchers of the gender segregation prefer the D index and the Ip index, and select a method or both which are used simultaneously in the Ds index and O (x) index, thus can more effectively measure gender segregation.

136.5 Conclusion

The paper uses the five occupational segregation measure methods of D index, Ds index, Ip index, Square Root index and A index to measure Chinese urban units' industry gender segregation from 2000 to 2010. The study finds out that Chinese industry gender segregation degree is far below the world average, but gender segregation of Chinese industries is on the rise, and becoming larger. It shows that gender segregation is more easily formed in the gender-dominated industry. These trends should attract enough attention from researchers and related sectors.

References

1. Li, C.: Status and trends of occupational sex segregation in China. Jiangsu Soc. Sci. **03**, 9–16 (2009)
2. Duncan, O.D., Duncan, B.: A methodological analysis of segregation of indices. Am. Sociol. Rev. **20**, 210–217 (1955)
3. Karmel, T., MacLachlan, M.: Occupational sex segregation- increasing or decreasing. Econ. Rec. **64**, 187–195 (1988)
4. Weiguo, Yang, Yujie, Chen: Measure of occupational gender segregation[J]. Chin. J. Popul. Sci. **03**, 77–87 (2010)
5. Yi, D., Liao, S.: Inspection and analysis of occupational gender segregation in Chinese industry. Chin. J. Popul. Sci. **04**, 40–47 (2005)
6. Yang, D., Xu, Z.: Study on measure and trends of occupational gender segregation in Chinese industry. Chin. Hum. Res. Dev. **02**, 16–23 (2012)

Chapter 137
Construction of Linguistic Resources for Information Extraction of News Reports on Corporate Merger and Acquisition

Wenxin Xiong

Abstract Detecting real time corporate merger and acquisition information from publicly available text data and feeding it to the decision-making module are essential for an applicable e-business management system. There are plenty of machine learning algorithms in text categorization and information extraction to address the problem, requiring different feature selection methods. Among them, linguistic features are key issues to accomplish this task. The acquisition of IBM's PC division by Lenovo in 2004 was chosen as a case, and news reports of this event were gathered from the Internet to build a Corporate Merger and Acquisition (M&A) mini-corpus. Comparing the M&A corpus with larger general corpora, we constructed a feature word list by applying Term Frequency and Inverse Document Frequency strategies and augmented it by introduction of the word groups from thesauri. Typical patterns, which highlighted the event of M&A, were collected by employing regular expression matching on these words acquired in the former step. By means of the accumulated language resources, the precision and recall for predicting the amount of the M&A in Chinese and English texts are 61.76 %, 65.22 % and 84 %, 71.43 % respectively.

Keywords Corporate merger and acquisition (M&A) · Information extraction · News reports · Language resources

137.1 Introduction

In today's global economic era, there are plenty of rapidly growing enterprises, which are eager to make a quick action response to the fast-changing situation, by deploying e-business management systems [1]. An e-business management system

W. Xiong (✉)
National Research Centre for Foreign Language Education,
Beijing Foreign Studies University, Beijing, China
e-mail: xiongwenxin@bfsu.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_137,

traditionally is powerful in providing data manipulation and decision-making function based on acquired data, but it often lacks acquisition of real-time data, especially textual data information [2]. Those data should be collected, encoded and keyboarded into the system manually so that they can be analyzed as inputs by a business intelligence module. This work seems labor-extensive, error-prone and time-consuming.

Intelligent text processing techniques can be of help in finding and extracting proper information from large quantities of text data. Feature selection plays an important role in achieving better results by applying sophisticated data mining or machine learning algorithms in such tasks. Among them, linguistic feature is one of the most significant features for recognition of events expressed in natural language. In the study of the case of Corporate Merger and Acquisition (M&A), we explore the construction of linguistic database (feature word list, word occurrence template bank, etc.) for an information extraction module, which can be integrated into an e-business management system to improve the precision of data extraction.

137.2 Related Work

Most traditional data mining tools inside business intelligence systems deal with only numerical data and can be hardly extended to handle textual information. Different from pages returned by search engines, a natural language question answering system returns a more concise and precise information pertinent to user's need. No matter which kind of information access, for machine or for human being, text information extraction is necessary for assimilation of the amount of information existed in heterogeneous sources [3].

There are three typical approaches to text information extraction task. The first is hand-made templates enriched with regular expressions. It requires linguist's expertise and needs more time to accumulate. The granularity and conflict between templates will be solved before they can be applicable to practical systems. The others are statistics-based methods. They can be categorized into two different types, i.e., classifier and sequence labeling. Then the former is further differentiated with a generative classifier (Naïve Bayes classifier) [4] and a discriminative classifier (Maximum Entropy model) [5]. The latter can be deemed as sequence models, implemented by Hidden Markov Model [6], Maximum Entropy Markov models [7] and Conditional Random Fields [8]. Among these distinct frameworks, feature selection is one of the most important procedures. Due to the nature of expressions in natural language, some linguistic characteristics may be of help for proper feature selection and can be applied to these systems.

From the perspective of theoretic linguistics, a predicate-argument structure can be dominated by a predicate, which holds a number of arguments and determines what semantic roles these related arguments play in this structure. Inspired by this theory, linguistic resources such as FrameNet and PropBank [9, 10], recording

more patterns and structural information, have been built and tested in various natural language processing projects with positive out-puts. These repositories provide a wealth of contextual information, which absolutely are beneficial to understand the whole sentence, rather than the combination of the meanings of single words. Some researchers have conducted pilot studies on automated extraction of predicate-argument structure from raw corpora [11].

An information extraction task comprises different sub-tasks, such as named entity's recognition, co-reference resolution and relationship determination. Most of them require more natural language processing techniques. Fortunately, some open-source tools serve the purpose [12]. Domain adaptation should be considered while different kinds of texts are processed. Construction of ontology may be of help to domain transfer [13].

In a word, there have been some progresses in employing linguistic features in information extraction, especially in recognition of expressions for specific events. We aim to explore the language resources building suitable for detection of Corporate M&A event expressed in Chinese and English.

137.3 Language Resources Construction

Language repositories, such as word list and word occurrence template bank, when constructed and used properly, play important roles in intelligent text processing tasks. It is commonly recognized that statistical data or linguistic rules can be elicited from massive authentic and natural texts, a.k.a. corpus.

Taking the corporate M&A event as an example, we first accumulated some documents pertinent to such an event as a specialized mini-corpus; then calculated the importance of each word in it and a large-scale general-purpose corpus by computing its TF*IDF; and highlighted the differences between the two word lists by comparing their metrics. A feature word list specific to M&A domain has thus been constructed. The expression template bank has been built using the combination of feature words by eliminating and scrutinizing all possible patterns in related documents. All this processing was conducted offline.

In our case, the acquisition of IBM's PC division by Lenovo in 2004 has been chosen as a sample corpus; from which linguistic features are learned.

137.3.1 Data Collection

In order to build language repositories used for extraction of M&A events, some prerequisite resources should be prepared in advance.

1. **Corpora** Two kinds of corpora can be utilized for highlighting the topic word list. One is a specialized corpus narrowed down in a specific domain, i.e.,

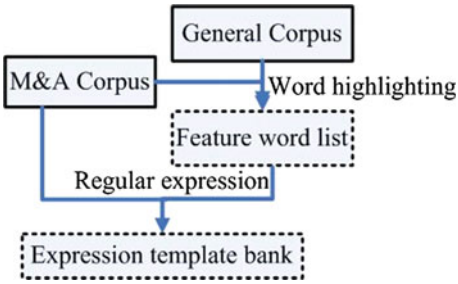


Fig. 137.1 The work flow of language resource building

Table 137.1 M&A Mini-corpus data

News	No. of texts	Total tokens
Chinese	126	48125
English	109	65437

Corporate M&A. The other is large-scale general-purpose balanced corpora, such as British National Corpus for English and Chinese National Corpus for Chinese.

2. **Language resources** Due to terms mismatching between query and texts, thesauri can be of help to expand related word groups, which share the same meanings but with no identical forms. *Tongyici cilin* is a well-known Chinese thesaurus, which groups a list of words based on lexical meaning [14]. *Wordnet* is its English counterpart [15]. Both thesauri are widely used in natural language processing.
3. **Tools** Without delimiters between words in Chinese texts, segmentation is the first step for almost all Chinese text processing systems. In our case, a General-purpose Word Segmentation (GPWS) system was employed [16]. Other tools, such as regular expression matching, pattern extraction, were also employed to accomplish different tasks.

The procedure and prerequisite resources are depicted in Fig. 137.1. Named entities, such as Lenovo, IBM as obligatory terms, and predicates, such as acquire, buy et al. as optional terms, were chosen as search terms and sent into search engines. The returned results were collected and saved as a specialized mini-corpus. Baidu was utilized for fetching Chinese news while Google for English news. The time range was set in 2004, in which year, the acquisition was completed. After manual examination, the mini-corpus is depicted in Table 137.1.

137.3.2 Feature Word List

Word information is the most commonly-used feature for text categorization tasks. Selecting the proper terms is crucial for them. The more times a word occurs in a

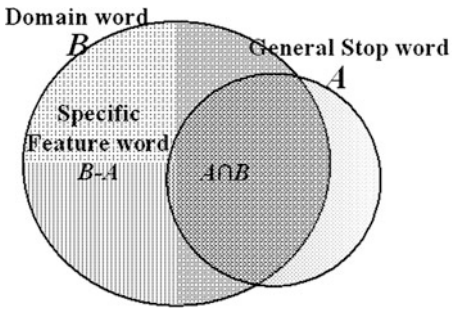


Fig. 137.2 Determination of feature word for specific domain

text, the more important it is; the fewer documents a word appear in, the more specific it is. The former is counted as term frequency (TF), while the latter is quantified as inverse document frequency (IDF). Combining these two properties of words is a conventional metric of term importance.

A salient word set B in M&A mini-corpus can be obtained by counting their frequencies and the numbers of texts in which they occur. TF*IDF is used for calculating term's importance. The general stop word set A can be generated over a large-scale general-purpose corpus in the similar way. The difference between B and A is the domain-related salient word set, which may be helpful for determining the topic of texts. The domain-specific feature word set is defined as B-A, as Fig. 137.2 shows.

Due to the limited size, thesauri were deployed for avoiding the possible problem of data sparseness. WordNet and Tongyici cilin were used as references for getting words unseen in the mini-corpus. After this step, an extended feature word set was accomplished.

The words, expressing the meaning of predicates buy and sell, were extracted by applying TF*IDF metric, and formed as set K. Some other words, sharing same meaning with the words in K but not appearing in the mini-corpus, were fetched by looking up original words in thesauri. The retrieved words formed a new extended word set $\bar{K}$, utilized to improve the recall rate. Table 137.2 shows the original word set K and extended word set $\bar{K}$ in italic type.

The superset comprised of K and $\bar{K}$ was indeed a word set, in which each element was considered as a positive feature for various data mining systems.

Table 137.2 Extended feature word set

Verb	Buy	Sell
Transitive verb	*K* {buy, buy over, buy out, take over, take ownership of, get ownership of, acquire, purchase, pursue,}	*K* {sell, sell off, spin off, dispose of, divest, transfer}
	$\bar{K}$ {get, buy up, quest for, go after, quest after, hunt for, incorporate, pick up, ……}	$\bar{K}$ {surrender, cede, deliver, give up, relinquish possession or control over, get rid of, cease to hold, deprive, strip, disinvest, shift}

137.3.3 Word Co-occurrence Expression Template Bank

Each text is simplified as a set of discrete words in BOW (Bag Of Word) strategy. It fails to capture the nontrivial meaning of the text. Due to the lack of contextual information, there exist plenty of ambiguities in word level. Feature word co-occurrence expression template can produce a better result than the individual words, for its broader view on contextual situation.

Because the theme of all texts in the mini-corpus was related to Lenovo's acquisition of IBM's PC division, all sentences, in which both corporate names were mentioned, expressed the same trade events. Thus, we employed regular expression techniques to locate all these sentences, and extracted the most frequent patterns to create a template bank. According to FrameNet, such an expression is connected with a basic commercial transaction involving a Buyer and a Seller exchanging Money and Goods. Thus, all obligatory arguments related to a predicate can be described as a quadruple:

predicate <***Buyer***, ***Seller***, ***Goods***, ***Money***>

There are different ways for expressing such a semantic scheme. For example, from buyer's perspective, taking the predicate as *buy*, the most typical expression is ***Buyer*** **buys** ***Goods*** **from** ***Seller*** **for** ***Money*** (*Lenovo buys the PC division from IBM for $1.75 billion*). While from seller's angle, taking the predicate as *sell*, the most common expression is ***Seller*** **sells** ***Goods*** **to** ***buyer*** **for** ***Money*** (*IBM sells its PC division to Lenovo for $1.75 billion*).

Although the numbers and properties of arguments involved in the case frame of a predicate are not unanimous, these patterns can be used as skeletons to deal with those variants existing in natural languages, empowered with fuzzy matching of regular expression.

For each transaction verb in the feature word set, combined with two corporate names Lenovo and IBM, we formulated a search query, and acquired possible template instances. All authentic templates extracted from the corpus composed a template bank, which will be used in next online matching.

137.4 Experiments and Results

We have accumulated a feature word list and a feature word co-occurrence template bank from corporate M&A news reports. Based on the mini-corpus, and the feature word set, we have concluded 8 English and 8 Chinese expression templates for detecting M&A events separately. In order to check the availability of these templates, we downloaded 20 Chinese and English news reports on *Geely's* acquisition of *Volvo* as test corpus respectively.

Firstly, we picked up these sentences which described the acquisition event from test bed, and searched test corpus with these candidate templates from

Table 137.3 Performance on prediction of templates over test corpus

Language	Precision (%)	Recall (%)	F (%)
Chinese	84.00	61.76	71.18
English	71.43	65.22	68.18

template bank, and collected all returned sentences, then calculated the precision and recall rates. The measures are defined as Eqs. (137.1)–(137.3).

$$\text{Precision} = \frac{\text{correct} \,\&\, retrieved}{retrieved} \tag{137.1}$$

$$\text{Recall} = \frac{correct \,\&\, retrieved}{relevant} \tag{137.2}$$

$$\text{F} = 2 \cdot \frac{Precision \cdot \text{Re}call}{\text{Pr}\, ecision + \text{Re}call} \tag{137.3}$$

The results are shown in Table 137.3. Although the overall measures were not high enough for automating feeding structural data into a business intelligence system, the approach served a computer-assisted information acquisition tool; especially some templates contributed much to the last performance, which meant those knowledge-enriched elaborated linguistic rules did enhance the recognition of corporate M&A event. The limitation may exist in how to determine granularity and feasibility of feature words and templates for specific events, and how to capture the balance between recall and precision.

137.5 Conclusion

The linguistic features play important roles in intelligent text processing. The author collected news reports pertinent to Lenovo's acquisition of IBM's PC division, obtained the domain-specific core word set by contrasting the words' distribution in different corpora, and acquired those frequent transaction expression templates by applying regular expression matching strategy. Because the contextual situation is considered, the feature word co-occurrence template can serve as a good indicator for a specific event. The experiment showed an encouraging result. Some commonly-used templates contribute high precision.

These language resources (feature word set and word co-occurrence template bank) are helpful for most text processing tasks. In the future, researchers will refine the template finding procedure and incorporate it into statistical machine learning frameworks to validate its feasibility. Some fundamental tasks, such as corporate names' recognition, predicates subcategorization, and identification of variants of the same expressions, should be paid more attention.

Acknowledgments The research was supported by Fundamental Research Funds for the Central Universities (2009JJ054) from Beijing Foreign Studies University, the Project of Humanities and Social Sciences of Ministry of Education (09YJA740013) and National Social Science Fund (11BYY051) and Program for New Century Excellent Talents in University (NCET-11-0591)

References

1. Liautaud, B., Hammond, M.: E-Business Intelligence: Turning Information into Knowledge into Profit. McGraw-Hill, New York (2000)
2. Maynard, D., Saggion, H., Yankova, M. et al.: Natural language technology for information integration in business intelligence. In: Proceedings of BIS 2007, pp. 366–380 (2007)
3. Srihari, R., Li, W.: A question answering system supported by information extraction. In: Proceedings of ANLC 2000, pp. 166–172 (2000)
4. Lewis, D.: Naive Bayes at forty: the independence assumption in information retrieval. In: Proceedings of EML1998, pp. 4–15. New York (1998)
5. Chieu, H., Ng, H.: A maximum entropy approach to information extraction from semi-structured and free text. In: Proceedings of AAAI 2002, pp. 786–791 (2002)
6. Freitag, D., Mccallum, A.: Information extraction with HMM structures learned by stochastic optimization. In: Proceedings of AAAI 2000, pp. 584–589 (2000)
7. Mccallum, A., Freitag, D., Pereira, F.: Maximum entropy markov models for information extraction and segmentation. In: Proceedings of ICML 2000, pp. 591–598 (2000)
8. Sarawagi, S., Cohen, W.: Semi-Markov conditional random fields for information extraction. Adv. Neural Inf. Process. Syst. **17**, 1185–1192 (2004)
9. Baker, C., Fillmore, C., Lowe, J.: The Berkeley FrameNet project. In: Proceedings of COLING-ACL1998, pp. 86–90 (1998)
10. Palmer, M., Gildea, D., Kingsbury, P.: The proposition bank: an annotated corpus of semantic roles. Comput. Linguist. **31**(1), 71–106 (2005)
11. Kawahara, D., Kurohashi, S.: Acquiring reliable predicate-argument structures from raw corpora for case frame compilation. In: Proceedings of LREC 2010, pp. 1389–1393 (2010)
12. Saggion, H., Funk, A., Maynard, D. et al.: Ontology-based information extraction for business intelligence. In: Proceedings of BIS 2007, pp. 25–27 (2007)
13. Daya, C., Dou, D.: Ontology-based information extraction: an introduction and a survey of current approaches. J. Inf. Sci. pp. 306–323 (2010)
14. Mei, J., Zhen, Y., Gao, Y., et al.: Tongyici Cilin. Shanghai Lexicographical Press, Shanghai (1983). (in Chinese)
15. Miller, G.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
16. Lou, T., Song, R., Li, W. et al.: Design and implementation of general-purpose interface of modern Chinese word segmentation system. J. Chin. Inf. Process. (5) (2001)

Chapter 138
Research on the Improvement of Business Age Model

Long Liu, Yanmei Xu, Lucheng Huang and Xiang Yao

Abstract Based on the theory of ecology, enterprise is viewed as life entity in the social environment. What enterprise bionics concern most is how to measure the vitality of those enterprises. This article modifies the original index system based on the theory of enterprise bionics, and chooses the real estate industry to do the empirical research. Analyzing enterprises' business age is a warning to those enterprises whose vitality is weak. From the comparison of enterprises' business age and nature age, this article has found that the business age of some enterprises like Vanke, Shenzhen special Economic Zong Real Estate, AVIC and Gemdale are between 30 and 40, and those enterprises' life vitality is strong and competitive; but some enterprises' business age like RongAn, DingLi Technology and Yihua are between 60 and 70, which means those enterprises' life vitality is weak and not competitive. The warning is serious.

Keywords Business age model · Enterprise vitality · Enterprise bionics · Real estate industryreal estate industry

138.1 Introduction

From the view of bionics, enterprise is viewed as life entity. The life cycle of enterprise include prepared, established, expanded, summit, declined and withered away, which is just like people's life cycle including birth, growth, maturing, becoming old and death. Just like people have nature ages and psychological ages,

L. Liu (✉) · Y. Xu · X. Yao
School of Management, University of Chinese Academy of Sciences, Beijing, China
e-mail: long525@126.com

L. Huang
School of Economics & Management, Beijing University of Technology, Beijing, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_138,

enterprises have their nature ages and business ages. Enterprise's nature age is used to measure the time from its birth to death; Enterprise's business age is used to measure its vitality, which illustrates its operating status and developing direction [1]. Business age is determined by enterprise's growth rate, competitive ability, management ability and operating benefit. Business age is only related to its vitality, which has nothing to do with enterprise's nature age. That means, enterprises, with old nature age, may not have old business age, vice versa.

In order to measure the enterprise's vitality, Japanese scholars first put forward the first quantificational model: business age model, but the model is not perfectly reflecting enterprise's vitality [2]. Therefore, some Chinese scholars put forward the improvement suggestions. In this article, the original business age model is improved on index selection, and evaluates system's construction and the interval of enterprise's business age.

138.2 Literature of the Business Age Model

138.2.1 Japanese Scholars' Research on Business Age Model

Japanese scholars use three indexes to calculate enterprises' business age, which are average sales growth rate in the current 5 years, employees' average age, age of the machinery's used. The meaning of the index is as below:

Average sales growth rate (X_1), reflects enterprise's growth rate and profit-earning ability;

Employees' average age (X_2), reflects enterprise's management ability and innovation ability, and Japanese scholars believe enterprise's innovation ability will increase as the decrease of employees' average age;

The age of Machinery's used (X_3), reflects the current technological ability. The technological ability will increase with the decrease of the age of machinery's used.

When one index is beyond the 3 times standard deviation, use to replace the index's standardized value.

Japanese business age model was constructed by data converting to calculate enterprises' business age. And business age is used to evaluate enterprise's vitality, thus enterprise vitality can be converted to business age by linear interpolation. Business age model is a useful tool to evaluate enterprises' vitality.

Evaluation on Japanese business age model.

The merits of the Japanese business age model:

First, innovation ability. As the first business age model, it is a breakthrough in quantitative method to evaluate enterprise's vitality.

Second, macroscopic analysis. Researchers can calculate different industries' business ages, because the data can easily be got from enterprises. Thus, people

can get to know where the industry's business age is going ahead, and that's also the meaning of the research on business age model [3].

The shortcomings of Japanese business age model:

The design of the index system has its defects. Business age is used to evaluate enterprise's vitality, which includes enterprises' growth rate, competitive ability, management ability and economic profit, but there are only 3 indexes in the business age model, which is not enough. Moreover, index cannot reflect enterprise's business age, for the formula is not the real age of machinery.

Second, the calculation of composite index value has its defects. The calculate formula is from linear regression, but the variable's economic meaning is not clear. The model just uses quantitative methods, lacking the qualitative methods.

138.2.2 Chinese Scholars' Improvement on Business Age Model

138.2.2.1 Research on the Improvement Method

The first improvement of business age model dates back to 2,000, some scholars put forward two different methods to modify the original business age model.

One method uses sales growth rate, employee average age, R&D input and capital earning rate to evaluate enterprise's business age. Moreover, the business age is limited from 1 to 30 [4].

The other method is use composite quotient value to evaluate the business age. The composite quotient includes four second level index (enterprise's IQ, EQ, CQ and nonmeasure variable) and 10 third level indexes. Moreover, the business age is limited from 20 to 80 [5].

Some scholars used business age model to do some research of applications.

On the basis of business age of enterprise, logistic regression model of the credit-risk state-probability model of life distribution were studied and the case calculation of credit risk control was given [6].

Another method regards enterprises as economic body, and takes the life of the enterprise as its research object, discusses the problems of enterprise age and growth strategies in enterprise bionic study by analyzing theories, building mathematical models and demonstration study [4].

Based on the theory of life span, the new division standard of different life stage and key factors that decide enterprises' growth are put forward [7].

After reviewing some foreign and domestic articles about the life-span of enterprise and with a clear understanding of the studies on life-span of enterprise, the article gives some suggestions to prolong it [8].

According to the investigation of Chinese enterprises' survival state, the article put forward the new measurement of enterprises' business age on the basis of Japanese research [9].

138.2.2.2 Evaluation of Improved Business Age Model

The improvement has amended some defects of the original business age model, but it still has some defects:

First, the index system is still not well-established, and it still lacks enough indexes to evaluate enterprises' vitality, such as finance index, corporate governance variable and so on.

Second, the calculation method and age interval of the business age model is not unified.

This article is focused on the construction of index system and the calculation interval of business age model, and aims to explain the theoretical foundation of the index system based on the theory research on enterprise bionics.

This article focuses on the construction of index system and the calculation interval of business age model, and aims to explain the theoretical foundation of the index system based on the theory research on enterprise bionics.

138.3 The Improved Business Age Model

138.3.1 Index Selection and Construction of Index System

From the view of bionics, enterprise vitality includes survive ability, growth ability and regeneration ability. The relationship among the three elements is shown in the graph one. Survive ability reflects the survival ability of enterprise and it is the foundation of growth ability and regeneration ability; Growth ability is the precondition of enterprise's survival, and enterprise should keep its vitality quality by continuous growth. Moreover, growth ability is the foundation of regeneration ability; Regeneration ability reflects enterprise's self-improvement ability, which is why enterprise can live longer.

Therefore, enterprise's vitality can be evaluated by survival ability, growth ability and regeneration ability (Fig. 138.1).

Through the related coefficient to select indexes and eliminate indexes which has significant correlation with another index, item analysis is used to eliminate low discrimination indexes; finally the index system is as below:

Enterprise's survival ability can be reflected by debt-paying ability and operating ability; enterprise's growth ability can be reflected by profitability and growth capacity; enterprise's regeneration ability can be reflected by innovation ability and transition ability. Thus, the below index system is shown in Fig. 138.2.

The meaning of some non-financial indexes:

(1) Interest Cost (X_{10}), which is the ratio of state-owned shares and total shares, reflects proportion of state-owned shares;
(2) Agent Cost (X_{11}), which is the ratio of management fees and total asset, reflects operating agent fees;

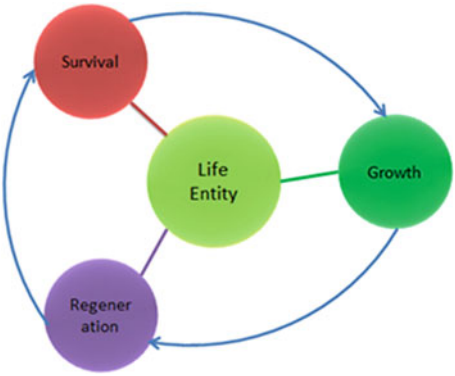


Fig. 138.1 Expression form of life vitality

(3) R&D Condition (X_{12}), which is the ration of R&D input and revenue, reflects enterprise's innovation input and innovation ability. As the R&D input is not revealed on the corporate annual report, intangible assets are used to replace R&D input because the R&D input will finally transform to intangible asset;
(4) Employee's academic condition (X_{13}), which is the ratio of employees with bachelor degrees and the total number of employees, reflects enterprise's innovation ability;
(5) Patent condition (X_{14}), which is the number of patents the enterprise has, reflects the enterprise's technology advantage.

138.3.2 Definition of Business Age Interval

From the theory of bionics, people reach its peak when they are between 30 and 40, so enterprises with strong vitality should be in the age interval between 30 and

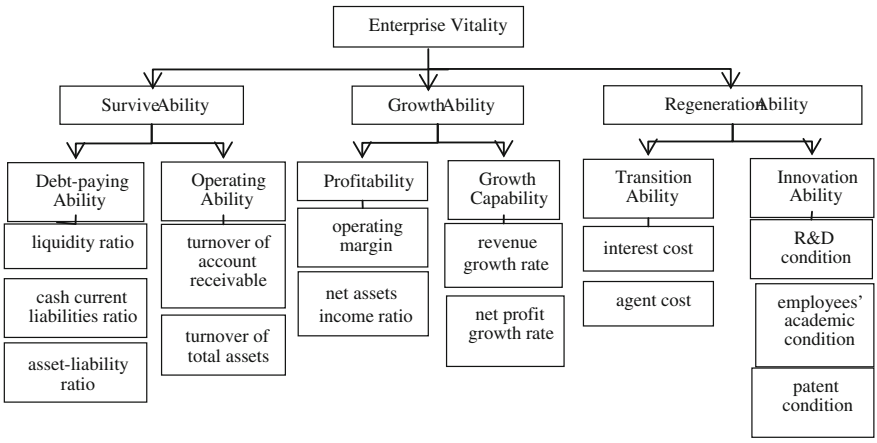


Fig. 138.2 Enterprise vitality index system

40. Besides, if enterprise's nature age is lower than the industry average nature age, its business age should be 0–30; if enterprise's nature age is higher than the industry average nature age, its business age should between 40 and 70.

The calculation of improved business age model is as below:

When $Z_i > Z_{1/4}$, business age SL_i is calculated as below:

$$SL_i = 30 + \frac{Z_i - Z_{1/4}}{Z_{\max} - Z_{1/4}} \times 10 \tag{138.1}$$

$Z_{\max}$ is the highest composite index value in the industry, $Z_{1/4}$is the upper quartile composite index value in the industry.

When $Z_i < Z_{1/4}$,

If the enterprise's nature age is lower than the industry average nature age, business age $SL_i$is calculated as below:

$$SL_i = 30 - \frac{Z_{\frac{1}{4}} - Z_i}{Z_{\frac{1}{4}} - Z_{\min}} \times 30 \tag{138.2}$$

$Z_{\min}$ is the lowest composite index value in the indstry.

If the enterprise's nature age is higher than the industry average nature age, business age SL_i is calculated as below:

$$SL_i = 70 - \frac{Z_{\frac{1}{4}} - Z_i}{Z_{\frac{1}{4}} - Z_{\min}} \times 30 \tag{138.3}$$

Through the converting, the interval of business age is limited between 0 and 70, and enterprise with business age between 30 and 40 has strong vitality.

138.3.3 The Relationship Between Business Age and Enterprise Vitality

From the research of corporate lifespan theory, this article believes that the relationship between enterprise's vitality and its business age is just like normal distribution curve. When enterprise's business age is low, the business age will grow as it improves its management ability, accumulates its capital; when its business age reaches the peak, the vitality will decrease as the bureaucratic problems occur in the enterprise.

The relationship between business age and enterprise's vitality is just like reverse "U" curve. When business age is lower than 30, enterprise's vitality increases as its business age increases; enterprise's vitality reaches its peak when business age is between 30 and 40; when business age is higher than 40, enterprise's vitality decreases as its business age increases.

Through the research of bionics, enterprise's best business age should between 30 and 40. So the enterprise's goal is to keep its business age between 30 and 40.

138.4 Empirical Research of Business Age in Real Estate Industry

From the research above, the business age model is affected by 14 third level indexes. Now, the constructed model is used to analyze the business age of real estate industry.

138.4.1 Sample Enterprises Selection

By the end of 2010, there are 106 enterprises in the real estate industry. All the financial data is downloaded from TianXiang database. Nine enterprises is deleted from the sample because of lacking some data in indexes such as turnover of account receivable, operating margin, net assets income ratio and net profit growth rate. One enterprise is deleted from the sample because of its turnover of account receivable is significant abnormal. Eighteen enterprises are deleted from the sample because the employee academic conditions cannot be obtained from the annual reports. In order to guarantee the completeness of data, the final sample includes 78 enterprises.

138.4.2 Index Selection and Index Weight

138.4.2.1 Index Selection

Based on the quantitative principle, the 14 third level indexes are selected and the patent condition index is deleted because no enterprise makes public its patent data. In order to test the indexes' discrimination, every index is tested through item analysis. The item analysis result is shown in Table 138.1.

The significant difference between the two groups is under the 1 % significance level, which means all the indexes on the left have passed the item analysis.

138.4.2.2 Index Weight

The article uses interviews and questionnaires to get the comments of 13 professors. The discriminating matrix is calibration from 1 to 9, and all data is input into Expert Choice 11.5 to get every index's weight from each professor. Five questionnaires were deleted for its data cannot pass the consistency test and finally get 8 effective questionnaires. The final weight is the average of eight weights from eight professors. The final weight is shown in Table 138.2.

Table 138.1 Item analysis result of each index

Index	Group	Average	Standard deviation	P value
Liquidity ratio	Group 1	3.48	2.37	7.42E-07***
	Group 2	1.30	0.06	
Cash current liabilities ratio	Group 1	83.43	2227.70	3.76E-07***
	Group 2	13.21	37.64	
Asset liability ratio	Group 1	80.06	36.43	4.65E-16***
	Group 2	44.62	118.31	
Turnover of account receivable	Group 1	0.45	0.02	9.06E-13***
	Group 2	0.13	0.00	
Turnover of total assets	Group 1	593.11	401681.85	0.000399***
	Group 2	10.07	20.98	
Operating margin	Group 1	38.64	497.38	1.69E-06***
	Group 2	7.99	11.03	
Net assets income ratio	Group 1	25.18	33.88	4.89E-17***
	Group 2	4.04	6.18	
Revenue growth rate	Group 1	166.99	19985.35	3.04E-07***
	Group 2	−33.49	693.96	
Net profit growth rate	Group 1	193.74	38314.91	3.8E-06***
	Group 2	−44.26	801.55	
R&D input condition	Group 1	0.08	0.02	0.025599***
	Group 2	0.00	0.00	
Interest cost	Group 1	0.32	0.05	2.62E-06***
	Group 2	0.00	0.00	
Agent cost	Group 1	0.03	0.00	8.9E-05***
	Group 2	0.01	0.00	
Employee academic condition	Group 1	0.71	0.01	6E-24***
	Group 2	0.15	0.00	

Comment ***represent 1 % significance level

Table 138.2 The index's final weight

Index	$IndexX_1$	$IndexX_2$	$IndexX_3$	$IndexX_4$	$IndexX_5$	$IndexX_6$	$IndexX_7$	$IndexX_8$	$IndexX_9$	$IndexX_{10}$	$IndexX_{11}$	$IndexX_{12}$	$IndexX_{13}$
Weight	0.038	0.059	0.059	0.134	0.140	0.048	0.051	0.069	0.125	0.073	0.072	0.082	0.049

138.4.3 Calculation of Sample Enterprise Business Age

Standardize all indexes according to the formula as below

n—number of samples, m—number of variables.

Adjust the standardized data, multiply (−1) to guarantee asset-liability ratio, interest cost and agent cost indexes the higher the better.

Calculate each enterprise's composite index value, and through the formula below calculate each enterprise's business age.

138.4.4 Result of the Improved Enterprise Business Age Model

Through the calculation, the average nature age of real estate industry is 17.05, and the average business age of real estate is 37.66. The distribution graph is shown in Figs. 138.3 and 138.4.

As can be seen from Fig. 138.3, the nature age of real estate industry is between 15 and 25, which means the real estate industry is a growing industry; from Fig. 138.4, the business age of real estate industry is between 30 and 40, which means the real estate industry is in a healthy condition. But there are still some enterprises' business ages that are between 60 and 70, which means the life vitality is not enough and they are in a dangerous condition of being obsolete.

The analysis of some enterprises' business ages is shown in Table 138.3.

This article just analyzes the real estate industry, and the enterprises nature ages and business ages are shown in Table 138.3. From the comparisons of enterprises' business ages and nature ages, this article found that the business ages of some

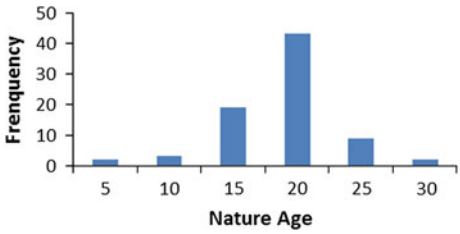


Fig. 138.3 Nature age distribution graph of real estate industry

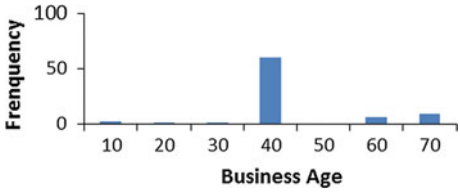


Fig. 138.4 Business age distribution graph of real estate industry

Table 138.3 Some enterprises' business ages and nature ages

Enterprise name	Nature age	Business age	Warning
RongAn Co., Ltd	22.00	69.96	Serious
Shanghai DingLi Technology Co., Ltd	19.00	68.00	Serious
Yihua Real Estate Co., Ltd	18.00	67.68	Serious
Gemdale Co., Ltd	15.00	36.35	No
AVIC Real Estate Co., Ltd	17.00	33.68	No
Shenzhen special Economic Zone Real Estate Co., Ltd	18.00	32.50	No
China Vanke Co., Ltd	23.00	32.46	No

enterprises like Vanke, Shenzhen special Economic Zong Real Estate, AVIC and Gemdale are between 30 and 40, and those enterprises' life vitality is strong and competitive; but some enterprises' business ages like RongAn, DingLi Technology and Yihua are between 60 and 70, which means those enterprises' life vitality is weak and not competitive, so the warning are serious.

138.4.5 Comparison between Business Age Model and Financial Model

First, financial model just uses history financial indexes to evaluate enterprises' current operating condition, and the evaluate standard is whether the enterprise is becoming ST stock in the next 2–3 years. But whether an enterprise becoming ST is determined by its operating profitability, which is not enough for us to evaluate an enterprise's operating condition. As is known to all, some enterprises, like Amazon, 360BUY, have always deficit in its first starting years, but they are successful enterprises. So it is not correct to evaluate enterprises only based on financial data. Improved business age model evaluates enterprises' life vitality based on the combination of financial indexes and nonfinancial indexes. Enterprises' business age give warning to those enterprises whose business ages are too old or too young, which overcomes shortcomings of the financial model.

Second, financial model uses multiple regression, logistic model, fisher model and artificial neural network model to evaluate whether the enterprise needs to be warned. Those models evaluate enterprises as normal and abnormal, and there is no transition between the normal and the abnormal. Actually, enterprises are distributed on every stage. Business age model evaluated enterprise's life vitality by business ages, which are distributed on every stage. Thus, the transition can be reflected from business ages.

In conclusion, business age model overcomes the shortcomings of financial model, but it still need empirical research on other industries to prove its features.

138.5 Conclusions and Future Research

Through the discussion of business age model, this article puts forward the new and modified index system and new business age model based on the theory of bionics. Then, this article uses the real-estate industry as a sample to do the empirical study. Through the empirical study, the real-estate industry is still on its midstream of life, whose average business ages are 37.66. This article gives warning to those enterprises whose business age is too old or too young, such as RongAn, DingLi Technology and Yihu. Business ages can reflect the enterprise's life vitality.

Above all, improved business age model can reflect enterprises' operating condition better than financial model. In the next research, researchers should develop the industry feature predicting of the business age model.

Acknowledgments Beijing Natural Science Foundation<The Early Warning Research of Regional Science and Technology Innovation Capability In View of Business Ecosystem>(9102021).

References

1. Xu, Y., Wang, L., Gu, L.: Evaluation and improvement of business age model. China Econ. Rev. **09**(22), 13–17 (2003)
2. Japan GuangShi Editorial Department, Two Thousand Japanese Excellent Enterprise Introduction. China Zhanwang Press, Beijing(1987)
3. Gu, L., Han, F., Xu, Y.: Business age research. Foreign Econ. Manag. **12**(12), 8 (2000)
4. Wen, S.: Enterprise age and growth strategy research. Master Dissertation of Beijing Industry University, Beijing (2001)
5. Diao, Z., Li, Z.: Business age model and its measure method. Sci. Technol. Prog. Polidy **09**, 137–138 (2003)
6. Zhao, G., Zhuang, X., Huang, X.: Enterprise risk analysis based on business age. J. Northeast Univ. **5**(23), 488–489 (2002)
7. Zhou K., Gu L.: Research on enterprise life cycle divide methods, divide foundation and growth variables. Mark. Mod. **02**(457), 119 (2006)
8. Team of life measurement and empirical study of Chinese enterprises. Theory and practice of measurement of enterprise's life. Stat. Res. **04**(25), 20–32 (2008)
9. Han, L., Zhao, B.: Research on enterprise vitality evaluation mathematic model. J. Hebei Univ. Sci. **2**(31), 81–86 (2010)

Chapter 139
Sales Forecast Using a Hybrid Learning Method Based on Stable Seasonal Pattern and Support Vector Regression

Fei Ye and J. Eskenazi

Abstract An obvious seasonality appears in customer demand of many industries. It can have a repetition period from a month to a year. In this paper, researchers use a hybrid learning method to improve sales forecast and supply chain management. This hybrid method combines Stable Seasonal Pattern (SSP) and Support Vector Regression (SVR) analysis. It provides a flexible approach which gives accurate forecast for budget and manufacture planning of companies.

Keywords Sales forecast · Stable seasonal pattern · Support vector regression

139.1 Introduction

Supply Chain Management integrates supply and demand management within and across companies. It coordinates major business functions such as marketing, sales, distribution, and finance. Their goal is to maximize the value offered to customers, which can be characterized by cost, quality, service, and lead time.

In order to optimize both production and delivery processes, customer demand data is first collected from retail network, and then processed by sales planners to finalize the forecast. A higher forecasting accuracy will make reducing costs and improve customer retention easier so it will help increase profits.

Statistic methods have been widely used for sales forecasting of different businesses. There is a lot of literature about qualitative and quantitative forecasting. Qualitative forecasting mainly relies on experts' knowledge and experience to make judgments based on historical analogy and marketing surveys.

F. Ye (✉) · J. Eskenazi
Shanghai Jiaotong University, Shanghai, China
e-mail: sophyning@gmail.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_139,

Quantitative forecasting uses machine learning theory and methods to study hidden patterns in available data through supply chain.

We can see very obvious seasonality in the sales data of some specific industries. Since demand fluctuations are most commonly influenced by periodical factors like weather, calendar, and special events, this seasonal character should be well analysed and taken into account when predicting future trend.

Support Vector Machine (SVM) is a promising technique for solving a variety of machine learning, classification, and function estimation problems. Literature shows its advantages in predicting non-linear data over other traditional time series methods, and even the widely used neural network analysis.

In this paper, we suggest a hybrid learning method which combines Stable Seasonal Pattern analysis and SVM. Its objective is to provide more accurate sales forecasts for an application in supply chain management.

139.2 Literature Review

139.2.1 Seasonality

Seasonality is defined as the tendency of time-series data to exhibit behavior that repeats itself every certain period [1]. A good understanding of seasonality helps the supply chain to provide accurate forecasts of demand trends. Thus it enables better production planning which must keep up with the pace of sales orders and marketing events. The Stable Seasonal Pattern will be introduced in Sect. 139.3.

139.2.2 Time Series Methods

Time series analysis has been used for forecasting purposes for a long time. Its range is from traditional time series methods to up-to-the-moment Artificial Neural Network. Large amounts of researches have been conducted in order to compare the accuracies and efficiencies of these forecasting models.

The Box-Jenkins ARIMA model is a combination of the AR and MA models. The most general Box-Jenkins model includes difference operators, autoregressive terms, moving average terms, seasonal difference operators, seasonal autoregressive terms and seasonal moving average terms [2]. Experiments showed that ARIMA is accurate for immediate and short-term forecasts [3].

Inspired by biological systems, particularly by research into the human brain, Artificial Neural Network (ANN) is a data-driven and self-adaptive method [4]. It is able to capture dynamic nonlinear pattern among the data. ANN outperforms traditional forecasting methods for quarterly and monthly data.

Based on the statistical learning theory developed by Vapnik (1998), Support Vector Machine is used in a number of classification and regression problems ranging from discrete manufacturing to bioinformatics. Its use for forecasting customer demand has also been validated [5].

In the following section, we will explain the details of Stable Seasonal Pattern and Support Vector Regression. A hybrid forecasting model will be experimented in Sect. 139.4 with real world data. The results will be compared with approved efficient ARIMA and ANN model in Sect. 139.5. Conclusion and further discussion will be given in the last section.

139.3 Hybrid Learning Method

139.3.1 Stable Seasonal Pattern (SSP)

With domain-specific knowledge, seasonality is already widely integrated within production planning, communication designing, marketing events scheduling, and other decision making processes in supply chain management

Thomas B. Fomby summarizes in his paper the implementation of SSP model in sales forecasting. Sales historical data is given by the matrix T[m, n], where T(i, j) (i = 1, 2,…, n; j = 1, 2,…, 12) represents the sales of a company in the jth month of the ith year.

Thus, the ith year's sales total is represented by

$$t_i^T = \sum_{j=1}^{12} t_{ij} \tag{139.1}$$

The seasonality index of the month j in the year i is given by

$$\mathrm{P_{ij}} = \mathrm{t_{ij}}/\mathrm{t_i^T},\ \mathrm{i} = 1, 2, \ldots, \mathrm{n};\ \mathrm{j} = 1, 2, \ldots, 12. \tag{139.2}$$

If the seasonality pattern of the historical data is stable over time, the average of the monthly seasonality index can give a more accurate representation of the seasonality

$$\bar{\mathrm{P}}_\mathrm{j} = \frac{1}{n}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{P_{ij}},\ \mathrm{j} = 1, 2\ldots, 12. \tag{139.3}$$

139.3.2 Support Vector Regression (SVR)

Based on Vapnik's statistical learning theory, Support Vector Machine uses a kernel-induced transformation from the original attribute space to a higher

dimensional space in order to maximize the margins between classes or minimize the error margin for regression. A well-chosen kernel function can transform the nonlinear problem into a linear model in a space of higher dimension.

In his book, M. Bishop illustrates SVM and SVR methods [6]. We will briefly go through the support vector method for regression. In simple linear regression

$$y(\mathbf{x}) = \mathbf{w}^{\mathbf{T}}\phi(\mathbf{x}) + b \tag{139.4}$$

We can then minimize a regularized error function given by

$$\frac{1}{2}\sum_{n=1}^{N}\{y_n - t_n\}^2 + \frac{\lambda}{2}\|w\|^2 \tag{139.5}$$

where the input vector **x** contains multiple independent variables, y is the predicted value, t is the actual value, and $\phi(\mathbf{x})$ denotes a fixed feature-space transformation.

Here we introduce a simple ϕ-insensitive error function to obtain a sparse solution

$$E_\epsilon(y(\mathbf{x}) - t) = \begin{cases} 0, & \text{if} |y(\mathbf{x}) - t| < \epsilon \\ |y(\mathbf{x}) - t| - \epsilon, & \text{otherwise} \end{cases} \tag{139.6}$$

Then we introduce two slack variables $\varepsilon_n \geq 0$ and $\hat{\varepsilon}_n \geq 0$, where $\varepsilon_n > 0$ corresponds to a point for which $t_n > y(\mathbf{x_n}) + \in$ and $\hat{\varepsilon}_n > 0$ corresponds to a point for which $t_n < y(\mathbf{x_n}) - \in$, as shown in Fig. 139.1.

Then the error function for support vector regression can be written as

$$C\sum_{n=1}^{N}(\varepsilon_n + \hat{\varepsilon}_n) + \frac{1}{2}\|\mathbf{w}\|^2 \tag{139.7}$$

Minimizing the error function above can be achieved by optimizing the Lagrangian below, where $a_n, \hat{a}_n, \mu_n, \hat{\mu}_n, \geq 0$,

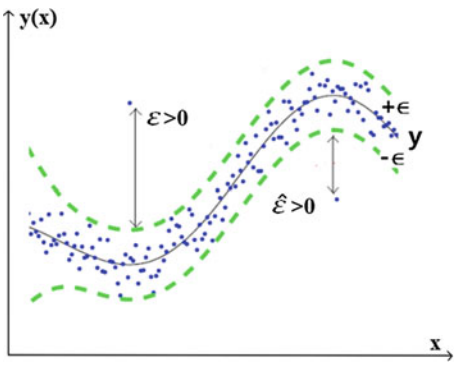


Fig. 139.1 SVR curve with the -tube

$$L = C\sum_{\mathrm{n}=1}^{\mathrm{N}}(\varepsilon_{\mathrm{n}} + \hat{\varepsilon}_n) + \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{\mathrm{n}=1}^{\mathrm{N}}(\mu_{\mathrm{n}}\varepsilon_{\mathrm{n}} + \hat{\mu}_{\mathrm{n}}\hat{\varepsilon}_n) - \sum_{n=1}^{N} a_n(\in + \varepsilon_{\mathrm{n}} + \mathrm{y_n} - \mathrm{t_n}) - \sum_{n=1}^{N} \hat{a}_n(\in + \hat{\varepsilon}_n - \mathrm{y_n} + \mathrm{t_n}) \tag{139.8}$$

Then we substitute using Eq. 139.4. By setting the derivatives of the Lagrangian with respect to $\mathbf{w}$, b, ε_n and $\hat{\varepsilon}_n$ to zero, we can get predictions expressed in terms of the kernel function $\mathrm{k}(\mathbf{x}, \mathbf{x_n})$

$$\mathrm{y}(\mathrm{x}) = \sum_{\mathrm{n}=1}^{\mathrm{N}}(a_n - \hat{a}_n)\mathrm{k}(\mathbf{x}, \mathbf{x_n}) + \mathrm{b} \tag{139.9}$$

With the corresponding Karush–Kuhn–Tucker (KKT) conditions

$$\begin{aligned} a_n(\in + \varepsilon_{\mathrm{n}} + \mathrm{y_n} - \mathrm{t_n}) &= 0 \\ \hat{a}_n(\in + \hat{\varepsilon}_n - \mathrm{y_n} + \mathrm{t_n}) &= 0 \\ (\mathrm{C} - a_n)\varepsilon_{\mathrm{n}} &= 0 \\ (\mathrm{C} - \hat{a}_n)\hat{\varepsilon}_n &= 0 \end{aligned} \tag{139.10}$$

139.3.3 Hybrid Sales Forecasting Method

In this paper, we incorporate SVR with SSP method. We use recurrent one-step and direct multiple-step SVR to predict future total annual sales. Then we can use the stable seasonal index to "distribute" the total sales among the various months of the next year.

139.4 Experiment

In this section, we conduct a forecast based historical sales data of a specific company from Jan 1965 to Dec 1971. Data are obtained from the Time Series Data Library and Data Market, complied by Australia. The last 24 observations are used as an out-of-sample test set.

In our sales history example, sales from 1965 to 1969 are used as training data. We can see in Fig. 139.2 that for each year, the seasonal peak begins in June, and then keeps increasing until the highest sales in November.

Stable seasonal pattern could be applied in this example to trace the sales trend of every year. The average seasonality indexes given in Table 139.1 are obtained by integrating Eqs. 139.2 and 139.3 from Sect. 139.3.

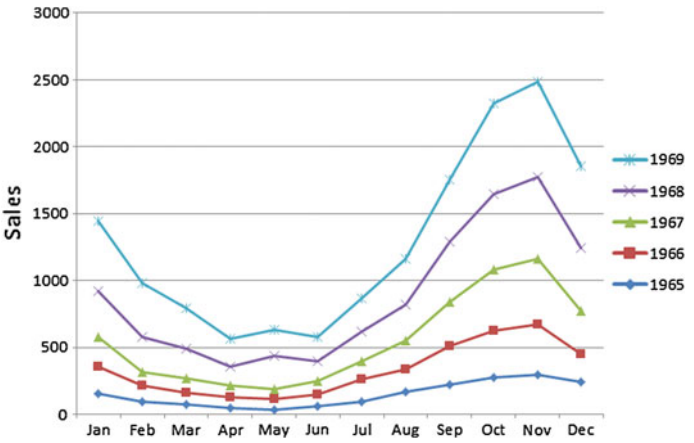


Fig. 139.2 Training data and seasonal trend

In our experiment, the SVR method illustrated in Sect. 139.3 is used with LS-SVMlab Toolbox [7]. The year number (1965–1969) is taken as input vector x, and y is the associated sales volume. Figure 139.3 gives the obtained regression curve, the kernel function that was used, and the related parameters.

When using the Support Vector Regression to predict sales of year 1970 and 1971, two methods, namely one-step and multi-step, are tested in order to compare forecasts accuracies. In one-step method, the first subsequent prediction is obtained and added to the input vector. The SVR function is retrained to predict the next period. In the multiple-step method, successive periods are predicted all at once [8]. The results of these two methods are compared using mean absolute percentage error (MAPE), which is a common error metric for quantifying the accuracy of predicted values.

$$\mathrm{MAPE} = \frac{100\,\%}{\mathrm{M}} \sum_{\mathrm{n=N+1}}^{\mathrm{N+M}} |(\mathrm{t_n} - \mathrm{y_n})/\mathrm{t_n}| \qquad (139.11)$$

where N is the number of training points, and M is the size of the forecasting horizon. Sales forecasts for the following 2 years are given in Table 139.2.

The annual sales predicted by the one-step SVM are retained due to their smaller MAPE. Monthly sales are obtained from SSP by multiplying annual sales by seasonality index. Figure 139.4 proves that our hybrid learning method can successfully capture seasonality trends and predict future sales.

Table 139.1 Average seasonal index (%)

Time	January	February	March	April	May	June	July	August	September	October	November	December
Index	9.33	6.37	5.14	3.65	4.10	3.76	5.62	7.54	11.35	15.06	16.09	11.99

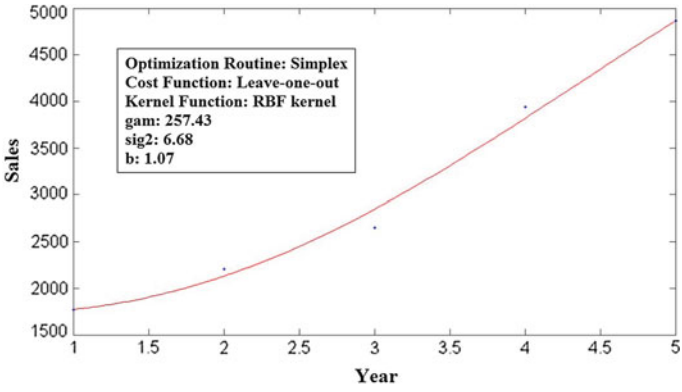


Fig. 139.3 SVR function estimate

Table 139.2 One-step SVM and multi-step SVM

Time	Actual sales	One-step SVM	Multi-step SVM
1970	5868	5798	5798
1971	6345	6442	6451
MAPE(%)	–	1.36	1.43

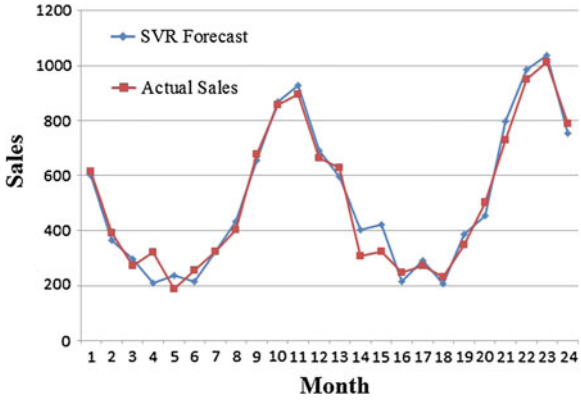
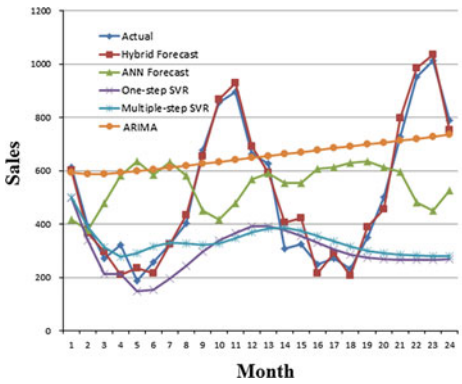


Fig. 139.4 Monthly sales forecast for year 1970 and 1971

139.5 Comparison with Other Forecast Methods

In this section, other widely used sales forecast methods have been tested to compare their prediction accuracy with the one of our hybrid learning methods. The performance of ARIMA, ANN, one-step SVR, and multiple-step SVR are tested with the same sales data as the one used in Sect. 139.4. By trial and test



Method	MAPE(%)	Rank
Hybrid	10.47%	1
ANN	70.69%	4
One-Step SVR	36.34%	2
Multiple-step SVR	38.55%	3
ARIMA	73.44%	5

Fig. 139.5 Comparison of forecast methods

processes, ARIMA (1, 1, 1) and ANN with one 5-neurons hidden layer are used in this comparison. Support Vector Regression is also tested by taking historical monthly sales directly as input vector to predict the following monthly sales (Fig. 139.5).

From the prediction results above, the superiority of the hybrid forecasting method seems quite clear. The prediction matches real sales data very closely. SVR and ANN outperform linear ARIMA model in capturing non-linear sales trend. Superiority of SVR over ANN is also shown in the above results. But in our specific industry example, where stable seasonal pattern is very obvious, a hybrid learning method, which combines the strengths of SSP and SVR methods, has a very strong advantage over other data-driven forecast methods.

For SVR training process, one-step SVR performs better than multiple-step SVR but the advantage is not significant.

139.6 Conclusion

In this paper, researchers proposed a hybrid learning method based on Stable Seasonal Pattern and Support Vector Regression to increase forecast accuracy in supply chain management when seasonality and trend are stable. SVR method is used to predict total sales value of the next forecast horizon, and SSP method is used to calculate monthly seasonality index.

This hybrid learning method was applied in a real sales series and was quite accurate in predicting future values. Compared with other time series forecast methods, the advantage of hybrid forecast is very obvious when capturing trend and seasonality. It means the absolute percentage error for monthly sales is lower than 11 % while the other methods are all above 36 % for the considered 2-year forecast.

The hybrid learning method is a more realistic approach for supply chain data processing. It can efficiently increase forecast accuracy by better integrating domain-specific knowledge and advanced learning theory.

References

1. Fomby, T.B.: Stable Seasonal Patter (SSP) model. Southern Methodist University. http://faculty.smu.edu/tfomby/eco5375/data/Notes (2010)
2. Prins, J.: Time series models. In: NIST/SEMATECH e-Handbook of Statistical Methods. http://www.itl.nist.gov/div898/handbook (2003)
3. Shahrabi, J., Mousavi, S.S., Heydar, M.: Supply chain demand forecasting: a comparison of machine learning techniques and traditional methods. J. Appl. Sci. **9**(3), 521–527 (2009)
4. Zhang, G., Eddy Patuwo, B., Hu, M.Y.: Forecasting with artificial neural networks: the state of the art. Int. J. Forecast. **14**, 35–62 (1998)
5. Levis, A.A., Parageorgiou, L.G.: Customer demand forecasting via support vector regression analysis. Chem. Eng. Res. Des. **83**(8), 1009–1018 (2005)
6. Bishop, C.M.: Pattern recognition and machine learning. Springer Science+Business Media, LLC., New York (2006)
7. De Brabanter, K., et al.: LS-SVMlab toolbox user's guide version 1.8. Internal Report: 10-146, ESAT-SISTA, Leuven, Belgium (2010)
8. Hamzacebi, C., Akay, D., Kutay, F.: Comparison of direct and iterative artificial neural network forecast approaches in multi-periodic time series forecasting. Expert Syst. Appl. **36**(2)Part 2, 3839–3844 (2009)

Chapter 140
Establishing a Project Management Model Based on CMMI: Experiences from Victory Soft Case Study

Xinmin Wang, Ling Liu and Yingjie Wei

Abstract To establish effective and efficient project management practices, the capability maturity model integrated (CMMI) is being developed to help companies improve project management. A CMMI model for small and medium scale software companies is established in this paper. The main stages of CMMI establishment are shown specifically. Both advantages and disadvantages are introduced. And then this model applies to Victory Soft Corporation, as an example. This model provides a reference template for Chinese small and medium scale software companies who are seeking appropriate software development project management model.

Keywords Project management · Capability maturity model integration (CMMI) · Software process improvement (SPI) · Victory soft corporation

140.1 Introduction

Many software projects fail finally for the lack of good project management. Since the 1980s, the rapid software expansion in scale followed by thorough project level risk has prompted people to "seek better ways to develop and evaluate software" [1]. Among them, the Capability Maturity Model Integration (CMMI) is becoming very popular worldwide as an industry standard of software process improvement (SPI) because of the level of details and breadth covered [2].

X. Wang (✉) · L. Liu · Y. Wei
School of Economics and Management, China University of Petroleum, Qingdao, China
e-mail: wxmwer@163.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_140,

In recent years, growing numbers of software development organizations are focusing on the guiding role of the CMMI in helping develop higher-quality software and reach their desired goals. However, many of them appear to have no idea on what are capability maturity models and how to follow the SPI initiatives based on process capability maturity models like CMMI [3].

In this paper, the basic theoretical knowledge of CMMI will be introduced briefly in Sect. 140.2. The establishment and application of a specific project management model based on the CMMI is proposed in Sects. 140.3 and 140.4, respectively, taking Victory Soft Corporation as a case. The conclusions are presented in Sect. 140.5.

140.2 CMMI Overview

CMMI is the successor to the capability maturity model (CMM), it is consistent with the international standard ISO/IEC 15504 [4]. The first version of CMM was developed in 1987 [5]. From then on, SEI has published several modules of CMM based on different architectures, contexts and methods. A model called Capability Maturity Model Integrated (CMMI) was published in 2000, and CMMI Version 1.1, 1.2 and 1.3 were published then. The major change in CMMI V1.3 is the support of Agile Software Development, improvements to high maturity practices and alignment of the representation (staged and continuous) [6].

CMMI is a fusion of multi-disciplinary, scalable product sets [7]. It provides a single integrated framework for improving an organization's process. The new integrated model framework eliminates the inconsistency of the various models, and it also increases transparency and understanding, and establishes an automated, extensible framework [8].

A Process Area (PA) is a cluster of related practices in an area. A process area is satisfied when company processes cover all of the generic and specific goals and practice for that process area. Selection priority of process areas is a big part of the establishment of CMMI [9]. All twenty-two of The CMMI Process Areas (PAs) can be grouped into the following four categories: process management, project management, engineering and supporting.

CMMI originated in the United States, widely promoted in Japan, Europe, Taiwan, India and other regions. Especially in India, the rate of IT companies adopted the CMMI even more than the United States. According to SEI statistics, most organizations assessed at CMMI Level 5 in the world are India's software enterprises. However, research and application of CMMI in China is still in its infancy. Less than 200 of Chinese IT companies are assessed at CMMI Level 2–5 respectively and among them, no more than 20 at CMM/CMMI Level 5, including Motorola, Huawei, Neusoft, UF, and BearingPoint. Five maturity levels of CMMI have developed as Fig. 140.1.

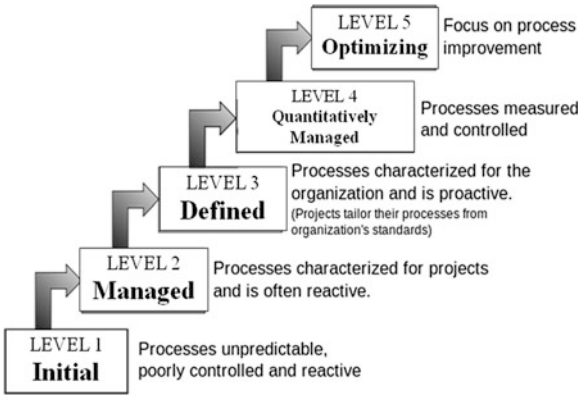


Fig. 140.1 The five maturity levels of CMMI

140.3 Establishment of CMMI Model

For small and medium scale software companies, level 3 of CMMI is sufficient. In this paper, a model of level 3 is established according to the practice of the Victory Soft.

In order to choose the suitable CMMI level for the company, the company appraisers should firstly reassess the existing progress of implementation, to identify whether the company has reached the key PAs goal of the CMMI. According to identified differences and problems during the investigation, counselors of company should improve every process of the KPAs respectively. The standard software processes can be regarded as a set of perfect process system documents by redefining, integrating, supplementing and improving the company's existing software process in accordance with the requirements of the CMMI model.

Each process area in the company's standard system documentation is corresponding to the CMMI standard procedure. Limited by the paper length, only one of the process areas which demands management process, is chosen to demonstrate how the processes support the process goals of the CMMI model standards.

Software requirements management activities run throughout the entire project life cycle. The purpose of the requirements definition is to form a clear, complete, consistent acceptance and testable requirements specification which is approved by project's stakeholders unanimously. The requirements definition process is shown in Fig. 140.2.

Requirements change control is used to deal with changes in demand, in accordance with the process of "Change Request–Approval–Change–Reconfirmation". It can prevent changes in demand from being out of control. Figure 140.3 shows the requirements change process.

Demand tracking is to ensure that products will be developed to establish and maintain the "requirements traceability matrix", it means to make the project

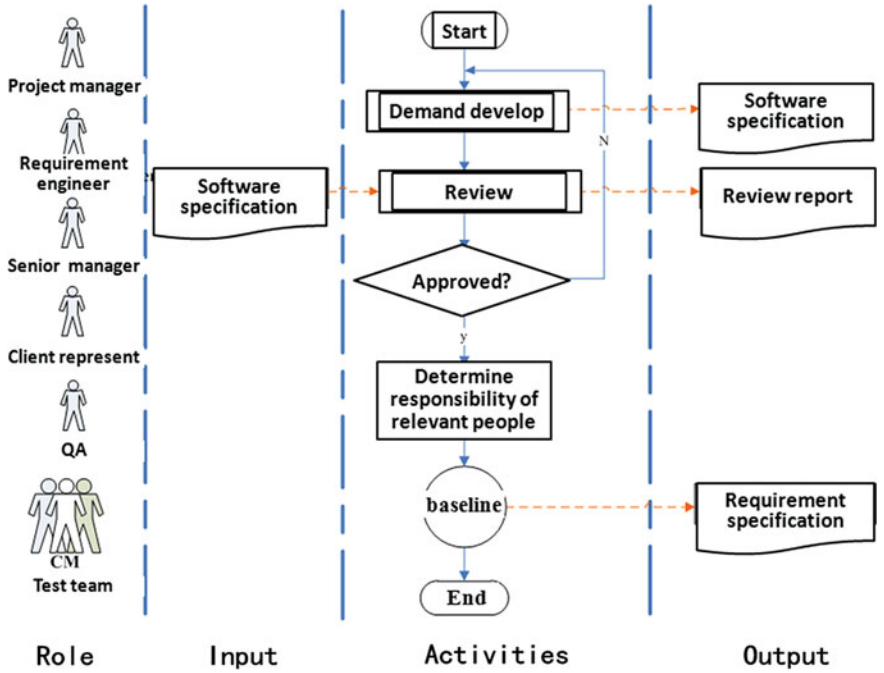


Fig. 140.2 Requirements definition process of company

benefit in audit, impact analysis, maintenance, tracking, redesigning, reusing, risk reducing, testing, etc. Figure 140.4 shows the demand tracking process.

140.4 Application Examples

140.4.1 Background

The Victory Soft Corporation is a medium scale IT company established in 2002, mainly oriented to the major domestic oil and petrochemical companies, this company mainly focus on product development and consulting services such as oil–gas exploration and development, petroleum engineering, production management, data management and professional software. In February 2007, Victory Soft implemented CMMI L3 appraisal and certification at the organizational level.

In this paper, the developing process of "Sinopec International Petroleum Company (SIPC) exploration information system" is chosen to show the application of CMMI L3 model established above. SIPC was built up as a set of integrated information system combined with information collection, storage and comprehensive application to meet the needs of business management such as

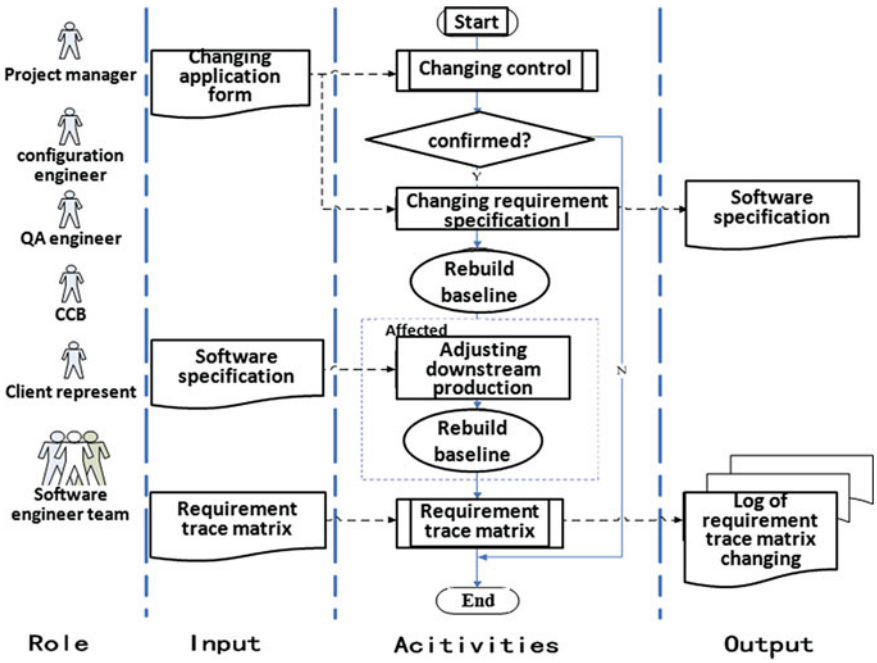


Fig. 140.3 Requirements change process of company

planning, implementation, controlling of oil field exploration projects. The development of the system involved Cross-sectional collaboration, multi-companies cooperation in different places, which is a large project with difficult management and much potential risk which is necessary to manage the risks of the project by CMMI model.

140.4.2 Risk Management Practices Based on CMMI L3

The purpose of risk management is to identify the potential problem so as to make plans on how to deal with risks at stages in whole life cycle of the project and take necessary measures to ease unfavorable influences.

Before the Victory Soft CMMI standard system document formed, the company seldom has any awareness about project risk management. And now, all processes about risk management were defined such as risk planning, risk identification and analysis, risk tracking and risk management measure. The risk influence (I), possibility (P) and risk factor (R) was defined by $R = P*I$. According to risk severity with schedule delays and budget overruns, the risk factors were categorized in 1–5 grades while risk possibility is divided respectively into 0–0.2,

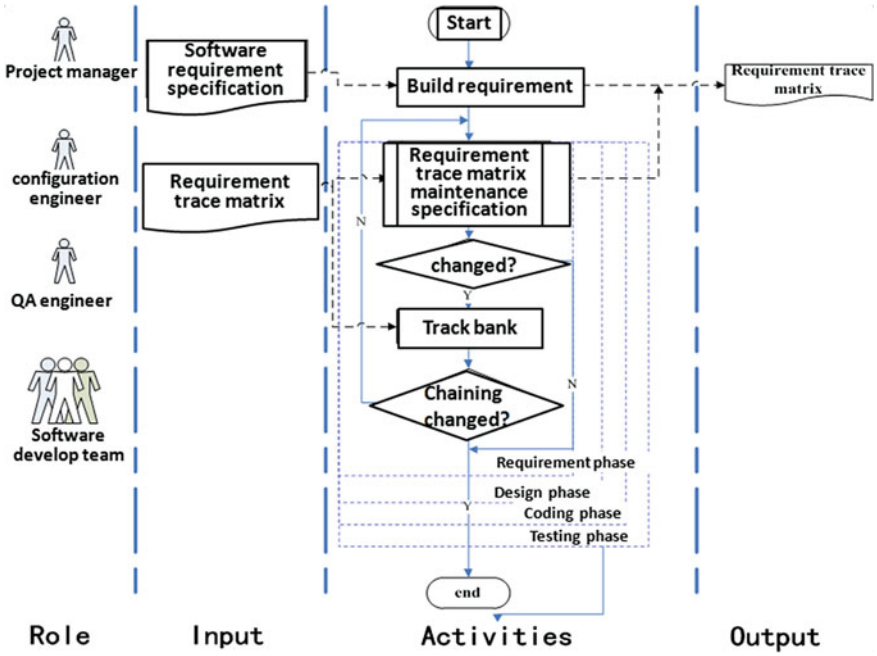


Fig. 140.4 Demand tracking processes of company

0.2–0.4, 0.4–0.6, 0.8–1. Based on the risk factors, the company formulated risk management strategies to avoid, transfer, slow, and accept.

In the implementation of SIPC, the detailed risk planning that the project team made according to the standard system documents of Victory Soft is shown in Table 140.1 as follows.

140.4.3 Analysis of Project Management Performance

With the various specified process management based on CMMI standard system and the risk management strengthened in SIPC exploration information system project, the risk factor reduced by 9.02 %, lower than previously expected, both the project schedule and costs were controlled in targeted range.

With the experience gained and social benefits perspectives, the establishment of a CMMI standard process system strengthened the staff's consciousness of software quality. It upgraded the project management level and software quality as well as the productivity, it constructed a good environment of software development with a set of project management methods and communication tools, it also improved customer satisfaction and enterprise images, all of these will finally bring to the business growth for the company.

Table 140.1 Risk management planning in SIPC

Sources of risk	Risk category	Risk description	Risk factor	Mitigation strategy
Internal risk	Technical risk	Whether it's new for project team	4	Acceptance
		Whether to gather all the quality metrics of software project	0.4	Avoidance
		Whether the software to be develop is connected to the software that developer provides	2.4	Acceptance
	Development environment risk	No available project management tools results in low efficiency of project management	1.6	Avoidance
		Project team members have training of each tools	1.6	Avoidance
	The number of personnel and experience-associated risk	Whether it has enough staff available	1.6	Avoidance
		Whether developers can participate in the whole project	2.4	Acceptance
		Whether there are appropriate personnel available	1.6	Avoidance
	Product scope risk	Without loc or fp to estimate product size result in the estimation without objective basis	2.8	Acceptance
	Business risk	Cost consumption by delay delivery	1.5	Transfer
	Particular project risk	Risk caused by inadequate collaborative development between the project team and technical personnel	2.4	Mitigation
		Database project cooperation	2	Acceptance
		Miscommunication in the combination with database project	3	Mitigation
External risk	Technical risk	Whether users need to create new algorithm or input and output technology	1.2	Transfer
	Particular project risk	Cooperation with other companies and the quality control of products provided	4	Acceptance

There were still some problems to be resolved in the implementation of CMMI as well. According to CMMI standards, most of the project status and necessary data should be recorded in documents when a project had been finished. In Victory Soft Corporation, a large number of daily document recording and report writing tasks according to CMMI requirements would take project manager's lots of time and energy, which discouraged in largely project managers to promote the implementation of the CMMI, it would result in the progress of CMMI implementation relatively slowly. Some records and reports on key process areas management eventually became a mere formality, which reduced the effects of CMMI to project management and software process improvement.

140.5 Conclusion

In this paper, a complete practical project management model based on the CMMI is proposed to demonstrate how to apply the standard system about software process improvement for small and medium scale software companies. CMMI helps software enterprises to manage projects and improve processes, to simplify the actions of key process improvement based on CMMI standards. With the increase of the company scale, the higher level of CMMI should be paid more attention to. Therefore, the lever 4 or level 5 of CMMI can be constructed on the basis of this paper, which is the further goal of this research.

Acknowledgments Funded by: MOE project of Humanities & Social Sciences (No.11YJCZH108); Fundamental Research Plan for the Central Universities (No.11CX04032B); Natural Funds Project of Shandong Prov.(No.ZR2011GQ004).

References

1. Shaw, M.: Writing good software engineering research papers: Minitutorial. In: Proceedings of the 25th International Conference of Software Engineering (ICSE 2003). IEEE Computer Society, Washington, pp. 726–736 (2003)
2. Yan, S., Xiaoqing, L.: Business-oriented software process improvement based on CMMI using QFD. Inf. Softw. Technol. **52**(1), 79–91 (2010)
3. Mahmood, N., Muhammad, A.B.: Identifying high perceived value practices of CMMI level 2: an empirical study. Inf. Softw. Technol. **51**(8), 1231–1243 (2009)
4. Mark, S., Mahmood, N., Ross, J., et al.: An exploratory study of why organizations do not adopt CMMI. J. Syst. Softw. **80**(6), 883–895 (2007)
5. Donald, J.R.: The CMMI: it's formidable. J. Syst. Softw. **50**(2), 97–98 (2000)
6. Minna, P., Annukka, M.: An approach for using CMMI in agile software development assessments: experiences from three case studies. IEEE Trans. J. Magn. **56**:584–599, Luxemburg (2006)
7. Lin, R.: The Integrated Software Development Management of CMMI. Publishing House of Electronics Industry, Beijing (2008)
8. Luo, Y.: Training Course Software Process and Capability Maturity Model Integration (CMMI). Tsinghua University Press, Beijing (2003)
9. Sun-Jen, H., Wen-Ming, H.: Selection priority of process areas based on CMMI continuous representation. Inf. Manag. **43**(3), 297–307 (2006)

Chapter 141
Retraction: Refining the Producer–Consumer Problem and Lamport Clocks

Yanchun Ma

Several conference proceedings have been infiltrated by fake submissions generated by the SCIgen computer program. Due to the fictional content the chapter "Refining the Producer–Consumer Problem and Lamport Clocks" by "Yanchun Ma" has been retracted by the publisher. Measures are being taken to avoid similar breaches in the future.

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_141,

Chapter 142
An Improved Risk Assessment Expert System for Elevator in Use

Yingjie Liu, Xingjun Wu, XinHua Wang, Weixiong Wang, Yuechao Song, Guojian Huang and Xinhua Wang

Abstract Traditional risk assessment method for elevator in use only focuses on objective factor, but many accidents in elevator are caused by subjective factors. This paper proposed an risk assessment method focusing on all the subjective and objective factors such as the design, manufacture, installation, maintenance, use, inspection elements, an weighted coefficient for considering the influence of different elements to the risk has been obtained by using the statistical experts opinions, then the risk level of the analyzed elevator can be obtained with the expert system, some suggestions are given according to the risk assessment results for improving the safety level of elevator in use. An example is given to illustrate how the proposed system is applied.

Keywords Elevator in use · Risk assessment · Compulsory inspection · Risk level

142.1 Introduction

The incidents of elevator have resulted in great loss in human life and aroused panic in our society, so more things should be done to prevent the incident and better improve the safety level of elevator. However, since the causes of incidents

Y. Liu (✉) · X. Wu · X. Wang · W. Wang · G. Huang
Guangzhou Academy of Special Equipment Inspection and Testing, Guangzhou, China
e-mail: Yingjieljd@163.com

X. Wang
School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou, China

Y. Song
State Key Laboratory of Fluid Power Transmission and Control, Zhejiang University, Hangzhou, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_142,

vary, it can be seen that most incidents resulted mainly from ignoring the safety rules by its users, and then is the manufacture and installation [1]. The cause of incident of elevator is influenced by both the objective and subjective factors. Risk assessment method can be employed to analyze those factors. Risk assessment which was firstly applied in process control system started in 1970s, now this technique are extended to other field such as transport system and management project [2]. The ISO standard of risk assessment was published in 2006 [3], which provides a general method to synthesize the risk level of the evaluated elevator and it can be seen that the analysis process in this standard is heavily dependent on the knowledge of risk assessors, and the considered factors are mainly on the objective factors of elevator. Shanghai Jiaotong university used safety checklist and database method to carry out risk assessment of elevator in use, in which the elevator system was divided into 11 subsystems, then the probability of the incident as well as the consequence of incident of each subsystem were analyzed, and finally the risk level of elevator can be obtained [4]. Risk assessment technique was used in safety function verification process by TUV [5]. Risk assessment system has been developed for facilitating the determining the inspection period of the elevator, the obtained results help to determine the inspection period for elevators [6]. A risk based method aimed at facilitating the maintenance work for elevator has been proposed, its considered factors were also focused on objective factors [7].

Up to now, a systematic risk assessment expert system for elevator in use which considers all subjective and objective factors has not been developed, this paper addresses to this issue aiming at quantifying the risk value of the assessed elevator for ensuring the safety of elevator users as well as elevator technicians.

142.2 The Structure and Subsystem of the Risk Assessment Expert System for Elevator in Use

142.2.1 The Structure of Risk Assessment System

The risk value of elevator is defined by combining the probability of incident with the severity of incidents which is shown in Fig. 142.1. The probability of incident is related to both subjective and objective factors such as the factors caused by design, manufacture, installation, maintenance, retrofit, and use of elevator, management of elevator as well as inspection of elevator. The severity of incidents is composed of three components such as trapped time in elevator, social impact, and emergency rescue effect.

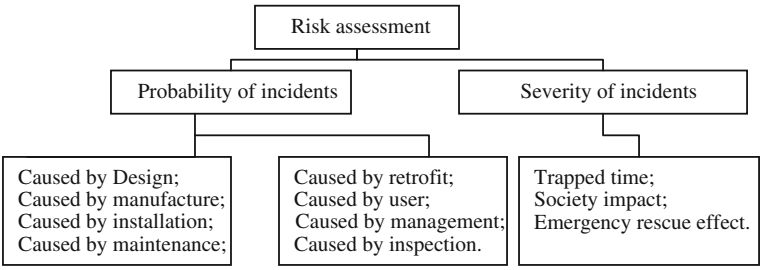


Fig. 142.1 The risk assessment method for elevator

142.2.2 The Subsystem Considering in Risk Assessment Expert System

The details for subsystems of the probability of incident are given in Tables 142.1, 142.2, 142.3, 142.4, 142.5, 142.6, 142.7, 142.8. Take the probability of incident caused by design subsystem for example, the elevator design quality contributes to the probability of incident, A four-component first layer subsystem is suggested in Table 142.1, and it is necessary to figure out how to chose the weighted coefficient of those components such as the level of design company u_{11}, the integrity of design document u_{12}, the evaluation result of design project u_{13}, the content of design document u_{14} by expert's opinions. To further investigate the influence of the content of design document, a three-component second layer subsystem is also developed, the characteristic value of those three considered components for a specific assessed elevator such as the calculation of traction ability, the verification of structure strength, the verification of safety component function are determined by risk assessors in accordance with an specified criterion, and its weighted coefficients are also obtained through expert's opinion. The rest subsystems are shown in Tables 142.2, 142.3, 142.4, 142.5, 142.6, 142.7, 142.8.

The subsystem of the severity of incident is given in Table 142.9.

Table 142.1 The influence of design to the probability of incident

Influenced factor	First layer subsystem	Second layer subsystem
Caused by design, u_1	The level of design company, u_{11}	
	The integrity of design document, u_{12}	
	The evaluation result of design project, u_{13}	
	The content of design document, u_{14}	The calculation of traction ability

Table 142.2 The influence of manufacture to the probability of incident

Influenced factor	First layer subsystem
Caused by manufacture, u_2	The level of manufacture company, u_{21}
	The integrity of manufacture document, u_{22}
	The quality of core components, u_{23}
	The results of certification test, u_{24}

Table 142.3 The influence of installation to the probability of incident

Influenced factor	First layer subsystem
Caused by installation, u_3	The level of installation company, u_{31}
	The quality of installation technician group, u_{32}
	The installation quality of core components in elevator, u_{33}
	The quality of self inspection, u_{34}
	The results of compulsory installation inspection, u_{35}

Table 142.4 The influence of maintenance to the probability of incident

Influenced factor	First layer subsystem
Caused by maintenance, u_4	The level of maintenance company, u_{41}
	The content of maintenance document, u_{42}
	The quality of maintenance technician, u_{43}
	The maintenance quality of elevator, u_{44}

Table 142.5 The influence of retrofit to the probability of incident

Influenced factor	First layer subsystem
Caused by retrofit, u_5	The analysis performed before retrofit, u_{51}
	The quality and performance of the retrofit company, u_{52}
	The performance of technicians, u_{53}
	The category of retrofit, u_{54}
	The results of self retrofit inspect, u_{55}
	The results of compulsory retrofit inspection, u_{56}

Table 142.6 The influence of use of elevator to the probability of incident

Influenced factor	First layer subsystem
Caused by use of elevator, u_6	The specification of elevator, u_{61}
	The condition of environment, u_{62}
	The condition of running state, u_{63}
	The previous incidents record, u_{64}

Table 142.7 The influence of management of elevator to the probability of incident

Influenced factor	First layer subsystem
Caused by management of elevator, u_7	The rule of elevator management, u_{71}
	The integrity of safety documents, u_{72}
	The capability of elevator keeper, u_{73}
	The set of safe signs, u_{74}

Table 142.8 The influence of inspection of elevator to the probability of incident

Influenced factor	First layer subsystem
Caused by inspection of elevator, u_8	The quality of Inspection technicians, u_{81}
	Inspection performance, u_{82}
	Inspection result, u_{83}

Table 142.9 The factors contribute to the severity of incident

Influenced factor	Subsystem
Trapped time, u_1'	The arrived time for rescue time, u_{11}'
	The real rescue time, u_{12}'
Social impact, u_2'	Flow rate of people using the elevator, u_{21}'
	Impacted zone when incident happened, u_{21}'
Emergency rescue, u_3'	

142.2.3 The Expert Weighted Factor for the Risk Assessment System

All the considered factors mentioned above in Tables 142.1, 142.2, 142.3, 142.4, 142.5, 142.6, 142.7, 142.8 are listed in a question form and sent to 25 experts in elevator field to obtain the weight coefficients which shown in Table 142.10. The collected data is processed with an average method given in Eq. 142.1,

$$w_i = \sum_{j=1}^{n} w_{ij} / \sum_{i=1}^{p} \left(\sum_{j=1}^{n} w_{ij} \right) \tag{142.1}$$

142.2.4 The Calculation Process for the Risk Value of Elevator in Use

To obtain the risk value of elevator in use R, the probability of incident P and the severity of incident S must be calculated first. The two parameters can be calculated through combining the elevator running characteristic matrix R and the elevator running impaction matrix M with the proposed expert weight co efficiency

Table 142.10 The obtained weight coefficient for risk assessment expert system

Weighted coefficient U for the probability of incident P					
Design u_1 (0.124)	u_{11}(0.210)	Installation u_3 (0.137)	u_{31}(0.160)	Use of elevator u_6 (0.120)	u_{61}(0.173)
	u_{12}(0.176)		u_{32}(0.237)		u_{62}(0.283)
	u_{13}(0.386)		u_{33}(0.215)		u_{63}(0.258)
	u_{14}(0.228)		u_{34}(0.183)		u_{64}(0.286)
Manufacture u_2 (0.118)	u_{21}(0.227)		u_{35}(0.205)	Management of elevator u_7 (0.121)	u_{71}(0.248)
	u_{22}(0.207)	Retrofit u_5 (0.126)	u_{51}(0.184)		u_{72}(0.218)
	u_{23}(0.328)		u_{52}(0.163)		u_{73}(0.264)
	u_{24}(0.238)		u_{53}(0.192)		u_{74}(0.270)
Maintenance u_4 (0.143)	u_{41}(0.237)		u_{54}(0.133)	Inspection of elevator u_8 (0.111)	u_{81}(0.340)
	u_{42}(0.201)		u_{55}(0.155)		u_{82}(0.332)
	u_{43}(0.279)		u_{56}(0.174)		u_{83}(0.328)
	u_{44}(0.283)				
Weighted coefficient U' for the severity of incident S					
u_1'(0.312)	u_{11}'(0.491)	u_2'(0.353)	u_{21}'(0.520)	u_3'(0.335)	u_{31}'(1.00)
	u_{12}'(0.509)		u_{22}'(0.480)		

matrix U, U'. Matrix R and M shown in Eq. 142.2 are given by the risk assessment technicians through checking and marking every subjective and objective aspects of the assessed elevator specified in the developed expert system.

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{pmatrix} \quad M = \begin{pmatrix} m_{11} & m_{12} & \cdots & m_{1n} \\ m_{21} & m_{22} & \cdots & m_{2n} \\ \vdots & \vdots & & \vdots \\ m_{m1} & m_{m2} & \cdots & m_{mn} \end{pmatrix} \tag{142.2}$$

The probability of incident of elevator P can be obtained by Eq. 142.3,

$$P = U \cdot R \tag{142.3}$$

The severity of incident of elevator S can be obtained by Eq. 142.4,

$$S = U^{'} \cdot M \tag{142.4}$$

The traditional risk value Rt is then calculated by Eq. 142.5.

$$R_t = 10^2 \cdot P \cdot S \tag{142.5}$$

Because an active on line state monitoring network including 1000 elevators has been built in our academy, it is reasonable to assume that on line monitoring date reflecting the risk level of elevator to some extent, it is a good complementary to the expert system. So an-on-line monitoring date based correction co efficiency f is used.

$$f = \sum_{i=1}^{8} r_i w_i \Big/ \sum_{i=1}^{8} r_i \tag{142.6}$$

Table 142.11 The criterion for classifying the residual risk level of elevator in use elevator in use

Risk level	Elevator state	Risk value
I	Good	[0, 10)
II	Acceptable	[10, 25)
III	Need minor improving	[25, 40)
IV	Bad	[40, 60)
V	worst	[60, 100]

where r_i is the occurrence rate of the individual incidents and w_i is the corresponding weight of the individual incidents.

The final modified residual risk value of elevator in use can be obtained by Eq. 142.7:

$$R_s = 10^2 \cdot f \cdot P \cdot S \tag{142.7}$$

The obtained residual risk value Rs for the monitoring elevator is then classified into five levels which are given in Table 142.11. Level I indicates that the assessed elevator is in good condition, there is no need to add extra safety measures for elevator, the level V shows the elevator is in the worst state, if the elevator is in the state, it is possible that some safety components is out of function, and/or the elevator is in bad management condition. The low value components of R and M indicate that the corresponding factors should be improved to prevent user from being injury by the possible incident happened in elevator. The user of elevator should take some measurements to reduce the risk level of elevator if the elevator is above level II.

142.3 An Example of Application of The Risk Assessment Expert System for Elevator in Use

Take a hotel elevator in Guangzhou as the risk assessed elevator. The floor of elevator is seven, the rated speed is 1.0 m/s, the rated capacity is 1000 kg. The on line monitoring statistical date is given in Table 142.12. The start number of this elevator is 485423 per year.

The correction co efficiency f is 0.175 calculated with Eq. 142.6:

Matrix R and M are given by the risk assessment technicians which are shown in Eqs. 142.10 and 142.11.

$$R = \begin{bmatrix} 0.1 & 0.2 & 0.1 & - & 0 & 0 \\ - & 0.1 & 0.3 & - & 0 & 0 \\ 0.2 & - & 0.2 & - & - & 0 \\ - & - & - & - & 0 & 0 \\ 0.1 & 0.15 & 0.2 & - & - & 0.2 \\ - & - & - & 0.1 & 0 & 0 \\ - & - & 0.2 & - & 0 & 0 \\ - & - & 0.15 & 0 & 0 & 0 \end{bmatrix} \quad M = \begin{bmatrix} 0.1 & 0.2 \\ 0.3 & 0.1 \end{bmatrix} \tag{142.8}$$

Table 142.12 The recent year statistical date of the assessed elevator

Incident type	Trapped in elevator	Safety circuit failure	Door locker failure	Door locker shortcut	Power failure	Brake failure	Elevator stopped by failure	Door vibration
Weighted coefficient	0.060	0.196	0.093	0.069	0.264	0.009	0.308	0.001
Incident number	6	0	18	3	7	1	15	0

Then one component of the probability of incident p1 given with Eq. 142.9,

$$P_1 = \sum_j u_{1j} \cdot (r_{1j} + \sum_k u_{1jk} \cdot r_{1jk}) = 0.13 \tag{142.9}$$

The total probability of incident can be obtained with Eq. 142.10,

$$P = \sum_i u_i \cdot P_i = 0.24 \tag{142.10}$$

The severity of the incident can also be obtained with Eq. 142.11,

$$S = \sum_i u_i' \cdot S_i = 0.35 \tag{142.11}$$

The modified residual risk value for monitoring elevator in use is shown in Eq. 142.12,

$$R_s = 10^2 \times 0.175 \cdot 0.24 \cdot 0.35 = 1.47 \tag{142.12}$$

The residual risk value is compared with the suggestion value in Table 142.11, it can be seen that the assessed elevator is at level I and in good condition, so there is no need to take extra safety measurement to enhance the safety level of the assessed elevator.

142.4 Conclusion

An improved risk assessment expert system for elevator in use is developed to quantify the residual risk level of elevator. Since all the subjective and objective factors involved in risk assessment work are considered in the proposed expert system, it is believed that the expert system can provide more information than the traditional one and the weighted coefficient of the risk assessment system have been obtained through investigating the statistical expert opinion, moreover, an correction coefficient for risk assessment obtained with the on line monitoring date of the assessed elevator is also introduced to modify the traditional risk value, an example of application of this risk assessment expert system is also present.

Acknowledgments This work is supported by the science and technology projects of General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China (2011QK321, 2010QK085, 2012QK065, 2010QK080), the projects of Guangdong Administration of Quality Supervision, Inspection and Quarantine (2011CT04, 2010CT04, 2010ZT02).

References

1. General Administration of Quality Supervision, Inspection of the People's Republic of China. Annual incidents report of special equipment. http://tzsbaqjcj.aqsiq.gov.cn/sgzl (2005–2011)
2. IEC 61508: Functional safety (1998)
3. ISO 14798:2006: Lifts, escalators and moving walks-risk assessment and reduction methodology (2006)
4. Gu, X.-Y., Zhu, C.-M., Zhang, P., et al.: Research of comprehensive safety assessment method for elevator system. China Saf. Sci. J. **18**(6), 146–151 (2008)
5. Haynl, M.: A look at IEC61508 the standard drives functional safety of machinery in the US and Europe. Mach. Control **12**(3), 21–25 (2010)
6. Mangalam, S., Foe, R.: Risk informed decision making by a public safety regulatory authority in Canada: a case study involving risk based scheduling of periodic inspections. WM'06 conference (2006)
7. Park, S.-T., Yang, B.-S.: An implementation of risk-based inspection for elevator maintenance. J. Mech. Sci. Technol. **24**(12), 2367–2376 (2010)

Chapter 143
Real-Time Service Integration Based on Business Process Execution Language

Le Zhao, Peng Xu and Ting Liu

Abstract In order to make the user-level service composition feasible, a real-time service integration based on Business Process Execution Language (BPEL) is put forward, which could execute a BPEL process by user's selection. By designing a system structure of real-time service integration based on BPEL, the difficulty waiting to figure out is summarized as the one-off recipient and keeping the business logic order unchanged. For illustration, the authors propose solutions on three aspects as message dependency, converting directed acyclic graph to workflow and the implicit message dependency. The results show that achieving one-off recipient should be accompanied by automatic processing all the data items entry inputting and listing the relevant recipients, together with real-time returning to the client after all the processes were completed. And if there is no data dependency between the processes selected, it should take the principle of parallel processing to reduce the overall process cycle; otherwise, it should be executed in order.

Keywords Real-time service integration · Business processes · BPEL

L. Zhao (✉)
Real Estate Archives of Hangzhou, Hangzhou, China
e-mail: zylele1204@163.com

P. Xu
Shandong Province Land Surveying and Planning Institute, Jinan, China

T. Liu
Institute of Remote Sensing and Earth Sciences, Hangzhou Normal University, Hangzhou, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_143,

143.1 Introduction

An important way of Web Service Composition is the workflow-based method [1]. Web Service could be seen as a business process and its composition as a combination of execution of the different business processes. The workflow-based Web Service Composition could be from the following two levels [2, 3], which are Web Service Choreography and Orchestration.

The standard of Web Services Choreography and Orchestration is Web Services Choreography Description Language [4] and Business Process Execution Language (BPEL) [5] respectively, that the former describes interactive behavior of Web Service Composition from a global perspective but the latter describes from a local perspective. BPEL has been enjoyed by the majority of people of all ages, with using an executable center process to cooperate with the service interaction. The overall goal, interrelated business and order of service invoking are all controlled by this center process. This centralized management enables services to add and delete with no understanding the interaction of each other, and allows compensating in the case of coming forth error or abnormity. The result can be seen as a new service, which can be executed by invoking other services [6]. As the same way of other orchestration, the design and publishing of BPEL needs technician's participation, and the user-level service composition is not yet feasible.

Therefore, a real-time service integration based on BPEL has been put forward, which could execute a BPEL process by user's selection. It explains how the system automates the generation of BPEL processes after users' selection as a real-time mode. We design a system structure of real-time service integration based on BPEL, which generates the BPEL by analyzing various business stakeholders' Web Service Description Language interface thereby determining the dependence of each business order, together with maintaining the existing business logic. This mode not only enhances the local parallel processing, but also makes the overall processing time reduced.

143.2 Problem Statement

Suppose that SH is a finite set of n-business stakeholders, which is composed by sh_1, sh_2, sh_3…,sh_n. Each stakeholder has input, output and interface with specific Web Service Description Language (WSDL). For any sh_i, here mark its input as SH_i^{in} and output as SH_i^{out} (which both or either is a set of standard data structure which is based on XML Schema).

Then the difficulty waiting to figure out by the real-time service integration based on BPEL can be summarized hereinafter.

(a) How to achieve the one-off recipient. It is the premise that the real-time service integration performs combination. Naturally it is noteworthy that the system how to achieve the one-off recipient and return the output material to the client after running.
(b) How to keep the business logic order unchanged. The new combined business logic order is from sh'_1 to sh'_k. But if the intersection set of SH_i^{out} and SH_{i+1}^{in} is an empty set, with the former being the output of sh'_i and the latter being the output of sh'_{i+1}, they could be parallel executing. And then the question of how to keep the business logic order unchanged and keep parallel executing at the same time is worth studying.

143.3 System Structure

Since all businesses are managed by one central service, and all stakeholders provide input and output interfaces, here using BPEL to control the integration of business processes is especially suitable. As shown in Fig. 143.1, there is the system structure.

It could be seen that the system open two interfaces up, and client interfaces receive business options to combine a new business. When system receives this composition sequence, the BPEL Generating Module could combine the Stakeholder. It is worth noticing that all the businesses combined are required to provide an interface for BPEL invoking since it is the active invoker. After the integration, system will publish it to the BPEL engine and open the interface for clients to invoke, and the latter will receive the Business Information Input. System will execute the BPEL after clients submit the entered information.

Take the business integration of existing property transaction and mortgage as an example. As shown in Fig. 143.2, there is a new business process integration.

At the acceptance stage, BPEL accept input from clients. When system execute to the transact business step, the input information of the 'transact' will be as parameters to invoke the distal end of 'Transaction Business'. Well, the latter will return result to BPEL after completing its execution. Similarly, the mortgage business will go through the same process. Finally, the system will output all material to return to the clients.

143.4 Solutions

The role of BPEL is to coordinate all the business processes, being able to truly reflect the user requirement at run time. Due to the current system supports users in a linear sequence of combined businesses, BPEL could also be simply design as serial implementation of the individual processes. This method is simple and effective, with no violation of the original intention of the users, but its execution

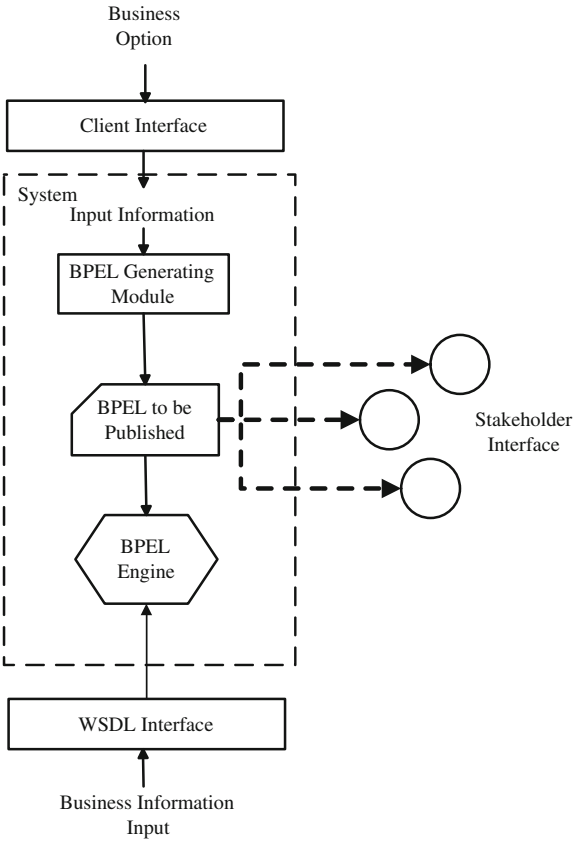


Fig. 143.1 System structure of custom service integration based on BPEL

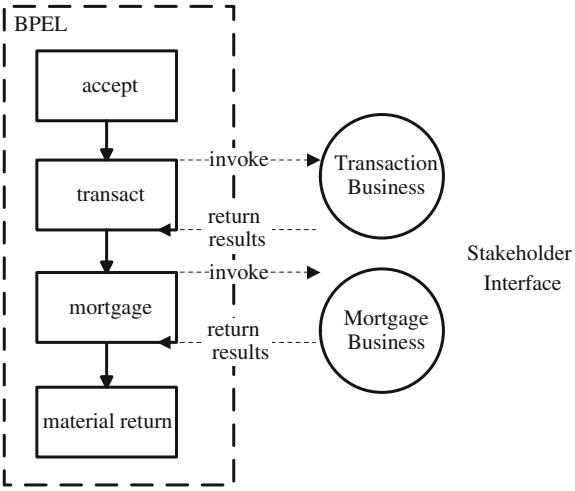


Fig. 143.2 Business process integration of existing property transaction and mortgage

efficiency is very low. Execution time of the business composition will be the sum of the separate process. Here, it is proposed that a solution makes the workflow design as serial basis but achieve as local parallel.

143.4.1 Message Dependency

Before introduce how to implement the local parallel, it is necessary to give some definition and hypothesis.

Definition 1: One business here is marked as sh_i, that has one input set and an output set, with the former marked as SH_i^{in} and the latter as SH_i^{out}. Both of each element in the set is called a message. When one business A's one input message needs business B's one output message as a parameter, it is called that A relies on B.

Definition 2: It is said that two messages are the same, if and only if the two message data structures (XML Schema) is exactly the same.

Hypothesis 1: It does not appear business loop dependence, in other words, there could be some service like sh_i, sh_{i+1},…,sh_j, existing shi depends on sh_{i+1}, sh_{i+1} depends on sh_{j+1}, …, sh_j depends on sh_i.

After the user selects all businesses require to be combine, the system can set up all the dependencies between those businesses through the analysis of the input and output messages. Because there is no business loop dependence existing, the dependency of all business is a Directed Acyclic Graph (DAG), as shown in Fig. 143.3. The 'source' means BPEL client input; accordingly 'destination' indicates its output. The nodes represent the business to be invoked, out-edges and in-edges represents the message output and input respectively.

143.4.2 Converting DAG to Workflow

Before introduce converting DAG to workflow, it is necessary to raise two theorems.

Theorem 1: DAG is DAG even if one or more nodes of its own are removed.

Theorem 2: There is at least one in-degree with node being zero in DAG, which is the sum of the number of in-edges.

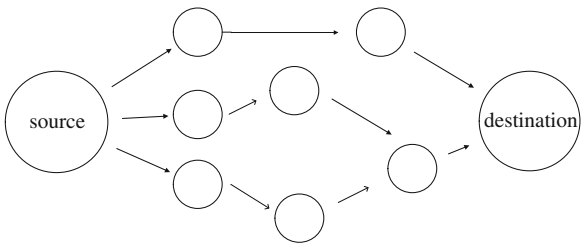


Fig. 143.3 Demonstration of DAG

When system establishes the dependencies of services, there will be a DAG taking shape. By the theorem 2, there is some in-degree with node being zero in DAG. Firstly, parallel execute those zero nodes, then delete them from DAG, and then get a graph that remains a Directed Acyclic Graph. Now just repeat the previous action, none of the nodes in the graph.

The following is the algorithm pseudo code.

```
Algorithm: Create_Linear_Process
Input: DAG
Output: BPEL
  BPEL =Empty BPEL
  While(DAG has vertices)
    BPEL = BPEL + ParallelExecution(vertices without in-edge)
    BPEL = BPEL + MsgAssignment(out-edges of vertices without in-edge)
    Remove_From_DAG(vertices without in-edge)
  End
```

This algorithm of Create_linear_process could successfully convert DAG to Workflows. Set the DAG as input and also BPEL as output. At first, assign the BPEL to empty; while the DAG has vertices, run its while loop. In accordance with this conversion, it could maximize the execution parallelism. For example, if Business 2′s input doesn't need the support of Business 1′s output, they can appear as such form in Fig. 143.4.

143.4.3 The Implicit Message Dependency

In some special cases, it cannot be judged the relationship between the nodes simply by the DAG. In other words, there may be some implicit dependencies

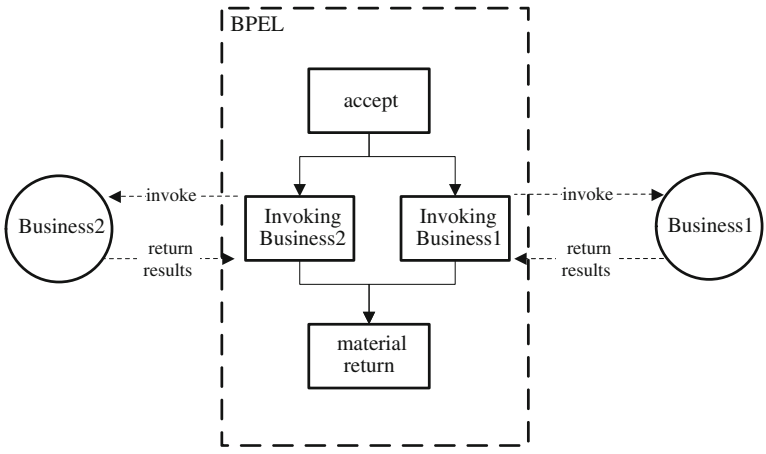


Fig. 143.4 An example of parallel business based on BPEL

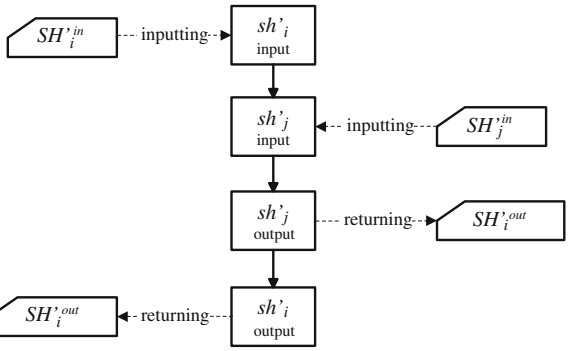


Fig. 143.5 An example of implicit message dependency

between the nodes and could not be described by the DAG. The following will give an example to illustrate.

Suppose that the user has selected sh'_1, sh'_2, sh'_3, …, sh'_k to combine, with sh'_i and sh'_j ($i < j$) not only among the selected items, but also having no dependence. However, if the business end of their own from one process, there may be an implicit dependency existing.

As shown in Fig. 143.5, there are two inputs and outputs respectively in one process, together with providing two services outward marked sh'_i and sh'_j. Set the input and output of sh'_i as SH'^{in}_i and SH'^{out}_i; similarly, the sh'_j's marked SH'^{in}_j and SH'^{out}_j. When users select sh'_i and sh'_j in the combined business, the system is likely to execute the sh'_i prior to sh'_j only relying on the input/output dependencies. In fact, successful invoking is needed between them, in other words, the former depend on the latter. It is an implicit dependency which cannot be represented in the DAG.

Possible result of this implicit dependency system will produce a deadlock; that is, when system executes the sh'_i prior to sh'_j, the returning value of the former needs the latter's input.

The system will never be able to invoke sh'_j's input for being waiting along.

For these problems, we propose the solution that, if there are a number of business processes from the same service interface in the business composition in question, these services allow only parallel execution instead of being sequential.

143.5 Conclusion

The authors propose solutions on three aspects as message dependency, converting directed acyclic graph to workflow and the implicit message dependency. It explains how real-time the system automates the generation of BPEL processes. The system generates the BPEL by analyzing various business stakeholders' Web Service Description Language interface thereby determining the dependence of each business order, together with maintaining the existing business logic.

It is shown that achieving one-off recipient should be accompanied by automatic by processing all the data items entry inputting and listing the relevant recipients, together with real-time returning to the client after all the processes were completed. And if there is no data dependency between the processes selected, it should take the principle of parallel processing to reduce the overall process cycle; otherwise, it should be executed in order. This mode not only enhances the local parallel processing, but also makes the overall processing time reduced.

References

1. Ko, J.M., Kim, C.O., Kwon, I.-H.: Quality-of-service oriented web service composition algorithm and planning architecture. J. Syst. Softw. **81**(11), 2079–2090 (2008)
2. Bucchiarone, A., Gnesi, S.: A survey on services composition languages and models. In: Proceedings of the International Workshop on Web Services Modeling and Testing, WS-MaTe, pp. 51–63. Palermo, Italy (2006)
3. Dijkman, R., Dumas, M.: Service-oriented design: a multi-view-point approach. Int. J. Coop. Inf. Syst. **13**(4), 337–368 (2004)
4. Avanesovm, T., Chevalier, Y., Mekki, M.A., et al.: Web services verification and prudent implementation. In: Proceedings of the DPM 2011 and SETOP 2011. LNCS 7122, pp. 173–189 (2012)
5. Mateo, J.A., Valero, V., Diaz, G.: An operational semantics of BPEL orchestrations integrating web services resource framework. In: Proceedings of the WS-FM 2011, LNCS 7176, pp. 79–94 (2012)
6. Fiammante, M.: Dynamic SOA and BPM: Best Practices for Business Process Management and SOA Agility. IBM Press, Upper Saddle River (2009)

Chapter 144
City Logistics Network Design and Optimization Under the Environment of Electronic Commerce

Yan Jiao, Dong Wang and Canquan Li

Abstract With the increasing consumer demand of online shopping, the necessity and significance of research on city logistics network are implicated. In this study, a model is developed to describe urban distribution network problem and minimize the costs of network logistics construction. Based on the order needs and distribution service level people proposed, the urban distribution system model is built, with taking consideration of the order density and distribution station service radius. The principle of the genetic algorithm combined with taboo search algorithm is described and proposed to find the solution to the model. The results of calculation indicate that the distribution system developed in this study is more economical and effective than the original one, and the validity of the proposed methods is demonstrated.

Keywords E-commerce · Urban distribution · Urban logistics · Genetic algorithm

144.1 Introduction

With the emergence and development of electronic commerce, the frequency of online shopping increase, the purchasing patterns of goods are more diverse and consumers demand better online shopping experience. Therefore, higher request on the service level of the urban logistics in the e-commerce environment is demanded, which creates new opportunities and challenges for city logistics. With the fact of the construction level of urban logistics network system in China, the current logistics infrastructure of some cities can only meet the low level, low

Y. Jiao · D. Wang (✉) · C. Li
Software College, Shanghai Jiao Tong University, Shanghai, China
e-mail: wangdong@cs.sjtu.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_144,

efficiency and small range of distribution service [1]. Many experts and scholars had proposed some solutions to solve the problem. For example, some Chinese scholars, such as Jinghui Tao, etc., consistently suggested that the improvement of both the city and regional logistics network services for distribution contributes to the whole logistics distribution system [2, 3]. Eiichi Taniguchi [4] and Gaetano Fusco [5] utilized the computer technology to develop a model to optimize and calculate the location planning of city logistics terminals. The reasonable and effective city logistics distribution system was presented in this paper and verified in a certain city.

144.2 City Logistics Network Model

144.2.1 City Logistics Network Problem Description

The urban area is divided into a number of city traffic zones by the urban main roads as boundaries. One or more city traffic zones share a distribution station, where citizens can take the packages themselves or request a door-to-door delivery. Direct delivery from the warehouse to plenty of distribution stations is not realized because that several depots is needed in which goods from the warehouse are further sorted and transshipped to distribution stations. The solution of the above problem has been made by designing a three-echelon freight distribution system. The design of a three-echelon freight distribution system includes the following decisions: location decisions; allocation decisions; routing decisions. Obviously, transportation tools used in the first echelon are different from that of the second echelon. With the designed transportation tools in the third layer, citizens may sometimes take the packages themselves. In Fig. 144.1 a schematic representation of a three-echelon freight distribution system is provided (Fig. 144.1).

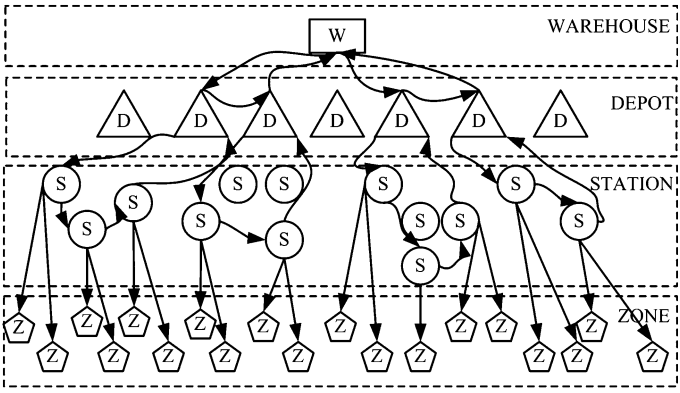


Fig. 144.1 An example of a three-echelon freight distribution system

144.2.2 City Logistics Network Model Design

-Sets:

W = {1,…, w}	set of the possible warehouse locations;
D = {1,…, d}	set of the possible depot locations;
S = {1,…, s}	set of station locations;
Z = {1,…, z}	set of zone;
G = {1,…, g}	set of first echelon vehicles, urban trucks;
V = {1,…, v}	set of second echelon vehicle, city freighters.

-Parameters:

CC_i^{depot}	fixed cost for opening a depot i, $i \in D$;
$CC_i^{station}$	fixed cost for opening a station i, $i \in S$;
CG	fixed cost for using an urban truck;
CV	fixed cost for using a city freighter;
ρ_j	order density of zone j, $j \in Z$;
P_j^{zone}	processing cost of each order in every zone j, $j \in Z$;
O_j^{zone}	forecast quantity of orders in every zone j, $j \in Z$;
$OC_i^{station}$	number of orders a station can operate once, $i \in S$;
$R_i^{station}$	the longest distance a station can serve, $i \in S$;
$CT_{ij}^{warehouse-depot}$	transshipment cost from i to j, $i \in W \cup D, j \in W \cup D$;
$CT_{ij}^{depot-station}$	transshipment cost from i to j, $i \in S \cup D, j \in S \cup D$;
$(x_i^{station}, y_i^{station})$	coordinate of station, $i \in S$;
$(x_j^{depot}, y_j^{depot})$	coordinate of depot, $j \in D$;
x_i^{depot}	0–1 variable, 1 if depot i is selected, 0 otherwise, $i \in D$;
$x_i^{station}$	0–1 variable, 1 if station i is selected, 0 otherwise, $i \in S$;
t^g	0–1 variable, 1 if urban truck i is used, 0 otherwise, $i \in G$;
t^v	0–1 variable, 1 if city freighter i is used, 0 otherwise, $i \in V$;
r_{ij}^g	0–1 variable, 1 if i is in front of j, 0 otherwise, $i \in W \cup D, j \in W \cup D$;
r_{ij}^v	0–1 variable, 1 if i is in front of j, 0 otherwise, $i \in S \cup D, j \in S \cup D$;
$x_{ij}^{station-zone}$	0–1 variable, 1 if zone j is served by station i, 0 otherwise, $i \in S, j \in Z$;
$x_{ij}^{depot-station}$	0–1variable, 1 if station j is served by depot i, 0 otherwise, $i \in D, j \in S$;

-Formulation:
Minimize

$$\sum_{i=1}^{s} CC_i^{station} x_i^{station} + \sum_{i=1}^{d} CC_i^{depot} x_i^{depot} + \sum_{g\in G} CGt^{g} + \sum_{v\in V} CVt^{v} + \sum_{i=1}^{s}\sum_{j=1}^{z} P_j^{zone} O_j^{zone} x_{ij}^{station-zone} x_i^{station} + \sum_{v\in V}\sum_{i\in D\cup S}\sum_{j\in D\cup S} CT_{ij}^{depot-station} r_{ij}^{v} + \sum_{g\in G}\sum_{i\in W\cup D}\sum_{j\in W\cup D} CT_{ij}^{warehouse-depot} r_{ij}^{g} \quad (144.1)$$

Subject to

$$(x_i^{station} - x_j^{zone})^2 + (y_i^{station} - y_j^{zone})^2 \mathrm{x}_{ij}^{station-zone} \le \mathrm{R}_i^{station} \quad i \in S, j \in Z \quad (144.2)$$

$$\sum_{j=1}^{z} x_{ij}^{station-zone} O_j^{zone} \le OC_i^{station} \quad i \in S \quad (144.3)$$

$$\sum_{i\in S}\sum_{j\in Z} x_{ij}^{station-zone} O_j^{zone} x_{li}^{depot-station} \le OC_l^{\mathrm{depot}} \quad l \in D \quad (144.4)$$

$$\sum_{i=1}^{s} x_{ij} = 1 \quad j \in D \quad (144.5)$$

$$\sum_{v\in V}\sum_{j\in D\cup S} x_{lj}^{v} = y_l \quad l \in S \quad (144.6)$$

$$\sum_{g\in G}\sum_{j\in \mathrm{W}\cup D} \mathrm{x}_{lj}^{g} = y_l \quad \forall l \in D \quad (144.7)$$

$$\sum_{l\in D\cup S} x_{lj}^{v} - \sum_{l\in D\cup S} x_{jl}^{v} = 0 \quad \forall j \in S\cup D, \forall v \in V \quad (144.8)$$

$$\sum_{l\in \mathrm{W}\cup D} r_{lh}^{g} - \sum_{l\in \mathrm{W}\cup D} r_{hl}^{g} = 0 \quad \forall h \in W\cup D, \forall g \in G \quad (144.9)$$

$$\sum_{l\in\Omega}\sum_{h\in\bar{\Omega}}\sum_{v\in V} x_{lh}^{v} \ge y_j \quad \forall j \in S, \forall \Omega \subset D\cup S, D \subseteq \Omega, \overline{\Omega} \cap \{j\} \ne \emptyset \quad (144.10)$$

$$\sum_{l\in\Omega}\sum_{h\in\bar{\Omega}}\sum_{g\in G} x_{lh}^{g} \ge y_j \quad \forall j \in D, \forall \Omega \subset W\cup D, W \subseteq \Omega, \overline{\Omega} \cap \{j\} \ne \emptyset \quad (144.11)$$

$$\sum_{l\in D\cup S}\sum_{j\in D} x_{lj}^{v} \le 1 \quad \forall v \in V \quad (144.12)$$

$$\sum_{l\in \mathrm{W}\cup D}\sum_{j\in \mathrm{W}} r_{lj}^{g} \le 1 \quad \forall g \in G \tag{144.13}$$

$$\sum_{h\in \mathrm{D}\cup \mathrm{S}} (x_{sh}^{v} + x_{sh}^{v}) - w_{ds} \le 1 \;\; \forall d \in D, \forall s \in S, \forall v \in V, w_{ds} = (0,1) \quad \forall d \in D, \forall s \in S \tag{144.14}$$

$$\sum_{h\in \mathrm{W}\cup \mathrm{D}} (x_{\mathrm{d}h}^{g} + x_{dh}^{g}) - w_{wd} \le 1 \quad \forall d \in D, \forall g \in G, \forall w \in W, w_{wd} = (0,1) \quad \forall d \in D, \forall w \in W \tag{144.15}$$

$$\sum_{g\in G} f_{wd}^{g} = \sum_{j\in Z} O_{j}^{zone} = \sum_{v\in V} f_{ds}^{v} \quad w \in W, d \in D, s \in S \tag{144.16}$$

$$f_{wd}^{g} \le \mathrm{UG} \sum_{h\in \mathrm{w}\cup \mathrm{D}} r_{dh}^{g} \quad \forall g \in G, \forall d \in D, \forall \mathrm{w} \in W \tag{144.17}$$

$$\mathrm{UG} \sum_{h\in \mathrm{w}\cup \mathrm{D}} r_{wh}^{g} - f_{wd}^{g} \ge 0 \quad \forall g \in G, \forall d \in D, \forall \mathrm{w} \in W \tag{144.18}$$

$$\sum_{l\in S}\sum_{k\in Z} O_{k}^{zone} x_{lk}^{station-zone} \sum_{j\in D\cup S} x_{sj}^{v} \le UVt^{v} \quad \forall v \in V \tag{144.19}$$

$$\sum_{\mathrm{w}\in W}\sum_{d\in D} f_{wd}^{g} \le UGt^{g} \quad \forall g \in G \tag{144.20}$$

The objective function (1) is the sum of four cost components: location cost for depots and stations, fixed cost for usage of urban trucks and city freighters, transportation cost on the first and the second echelons, the cost of the door-to-door deliveries. Constraints (2) impose that the distance from the zone center to the station that serves it can't be longer than the station's service radius. Constraints (3) impose that the sum of all zones' demands is not larger than the operation capacity of the station that serves them. Constraints (4) impose that the sum of all stations' demands is not larger than operation capacity of the depot that serves them. Constraints (5) impose that every zone must be served by one station. Constraints (6) impose that every station must be served by only one city freighter once. As the same, Constraints (7) impose that every depot must be served by only one urban truck once. Constraints (8) and (9) impose that urban trucks and city freighters enter into and exit from the same depot and station every time. Constraints (10) and (11) impose that urban trucks and city freighters must serve for more than one depot and station. Constraints (12) and (13) impose that urban trucks and city freighters can only be attributable to one depot and station. Constraints (14) impose that a station can be served by no more than one depot and a station links to no more than one depot in the route of city freighter

g. Constraints (15) impose that a depot can be served by no more than one warehouse and a depot links to no more than one warehouse in the route of urban truck v. Constraints (16) are the flow conservation constraints i.e. the amount of flow leaving the warehouse and the depots is to be equal to the total orders of the zones. Constraints (17) and (18) impose that city freighters are use in the second echelon and urban trucks are used in the first echelon. Constraints (19) and (20) impose that the demand assigned to a city freighter v and urban truck g has to be less than its own capacity, if the vehicle is used.

144.3 Algorithm Design of City Logistics Network

As a typical NP-Hard problem, the urban logistics network design can't be completely solved by a routine method. GA is an available method in network optimization, while taboo search algorithm is practicable in VRP. With the combination of GA and Taboo Search Algorithm, a new hybrid method is brought out to simplify bi-iterative complexity and settle the local convergence of GA. Figure 144.2 shows the frame of mixed Genetic Algorithm and Taboo Search Algorithm.

Algorithm in Pseudo Code Format is as following:

```
/*
* Pc: crossover probability
* Pm: mutation probability
* M: population size
* G: terminate evolution algebra G of GA
* Tf: fitness function, if any evolved individual's fitness function is more than Tf,
      then the evolution process terminates
* T: terminate evolution algebra of TS
* N: the times of searching solution's neighbourhood
*/
initialize Pm, Pc, M, G, Tf
generate first population Pop
   do
   {
      calculate each individual fitness F (i)
      initialize an empty population newPop
      do
      {
      select two individuals with probability from Pop according to the fitness
      if (random (0, 1) < Pc)
      {
         crossover the two individuals with Pc
      }
      if (random (0, 1) < Pm)
```

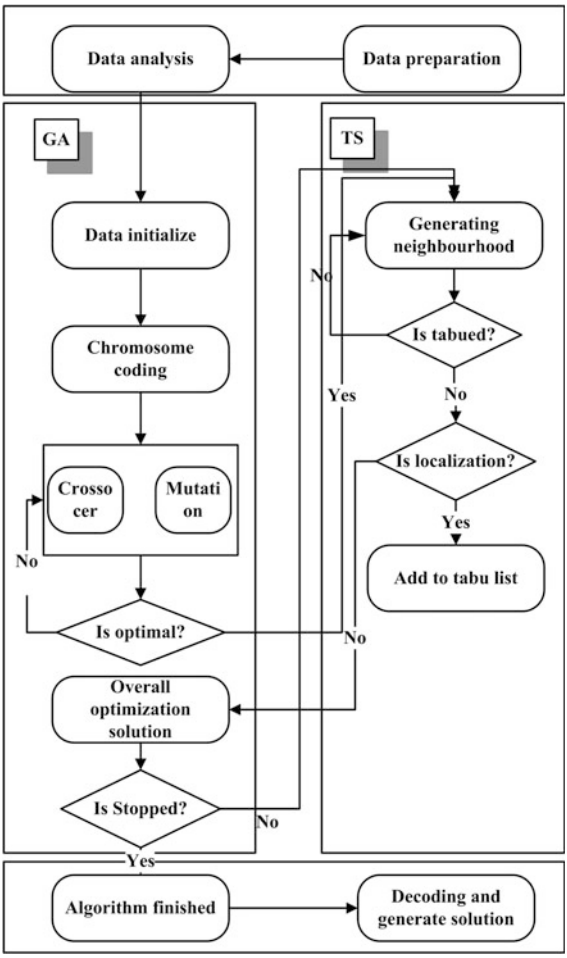


Fig. 144.2 Algorithm frame of Genetic Algorithm combined with taboo search algorithm

```
    {
       mutation the two individuals with Pc
    }
    add the two new individuals into newPop
    } until (M individuals is created)
    replace Pop with newPop
}until (Any chromosome outscore Tf, or breed algebra exceed G)
get the currently optimal solution Sg
initialize T, N
taboo list = null
length of tabu list = 1
Sbest = Sg
do
{
```

```
neighborhood numbers n = 0
calculate evaluation value E_best of S_best
do
{
   select a new string S' in the neighborhood of S
   if(S' is not in taboo list)
   {
      calculate evaluation value E_localbest of S'
      if (E_localbest < E_best)
      {
      put S' into taboo list
      }
      if (E_localbest >=E_best)
      {
      S = S_localbest
      E_best = E_localbest
      }
      n = n+1
      }
      }until (n >=N)
      S_best = S
      put S into taboo list
      t = t+1
      }until(t >=T)
      output S_best
```

Here are the important steps of algorithm realization:

- Chromosome unicodes
  A chromosome is made up of N natural numbers. Each chromosome contains not only locations and allocations of stations and depots, but also arrangements of transport vehicles and routs.
- Population regeneration.
  The crossover operator controls the speed of personal new-generation. It is used in both replacement of location and allocation of depots and stations and replacement of routs and arrangements of vehicles which includes single-point and multi-point crossover.
  The mutation operator is designed to control certain chromosomal genetic variants and improve the search capabilities of the algorithm. We use single-point and multi-point mutation in both vehicle and rout replacement and depot and station replacement.
- Termination condition
  Taboo search algorithm is used as a reference to termination condition of the algorithm. It avoids falling into local optimal with taboo table used to store the local optimum and find the best solution.

Table 144.1 Comparison diagram of optimization results

Optimization scheme	Iterations	Number of depots	Number of stations	Cost of logistics network (/day)	Distribution time of each order (h/order)	Distribution cost of each order (yuan/order)
Before optimize	_	20	140	132845.69	7–8	9–11
After optimize	500	5	67	879643.10	4.5–6	6.5–8
After GA	1000	10	92	107865.46	5–7	7–9

144.4 Instance of Verification

A city logistics distribution network was required to be built in Shanghai. Initially with the support of 20 depots and 140 stations, the fact of long distribution time and high distribution cost existed. Through calculation of mixed genetic algorithm and taboo search algorithm, the optimization results was concluded, which is shown in Table 144.1 compared with typical GA. As is shown in the table, not only the number of depots and stations and the cost of logistic network per day decrease, but also the distribution time of each order shortens and distribution cost of each order reduces after using the mixed algorithm. From Table 144.1 we concluded that GA mixed with TS is more efficient than typical GA, in addition GA mixed with TS is much more optimal after 500 iterations than that of GA after 1000 iterations.

144.5 Conclusion

A three-layer model is applied to establish the city logistics distribution network in this study. The model of one layer is based on traffic area density of stations. The model of other layer is derived from the establishment of stations which are based on the selected depots and optimized delivery distribution routing. At the meantime, optimal solution is obtained with using the genetic algorithm. Feasibility and correctness of this study were verified with real data.

References

1. Tao, J.: City Logistics Distribution System Research of Regional Center. Guangxi qinzhou bonded port area (2010)
2. Wen, H., Zhang, Z., Zeng, W.: Study on B to C E-business Distribution of Convenience Stores in China. Logistics Management, pp. 30–32 (2008)
3. Shi, L., Zhang, J.: Research on Distribution Mode By Logistics Nodes For E-commerce Websites. Southwest Jiao Tong University (2008)

4. Taniguchi, E., Noritake, M., Yamada, T.: Toru Izumitani. Optimal size and location planning of public logistics terminals. Transportation Research, pp. 207–222 (1999)
5. Fusco, G., Tatarelli, L., Valentini, M.P.: Last-Mile a procedure to Set-up an Opitimized Delivery Scheme. University di Roma “La Sapienza” (2009)

Chapter 145
The Evolution Process of Agri-Products Supply Chain System of "Company & Farmer"

Jiemei Li, Youjin Gu and Hao Wu

Abstract In order to study the evolution process of agri-products supply chain system of "company & farmer", a self-organization dynamic model is established on the basis of block growth model. Then the research analyzes the stability of the model and simulates the evolutionary process. The results show that cooperative and competitive effects have a close relationship with the system evolution directions. The system can be optimized by magnifying the cooperative effects. The farmer, as the initial smaller side in the system, will be going to die when the cooperative effect is negative. So in order to protect the farmers' interests, it is necessary for the government to encourage the organization of the farmers to amplify their scale.

Keywords "Company & Farmer" · Agri-products Supply Chain · Evolution · Simulation

145.1 Introduction

Agricultural industrialization is an important way to increase farmer's income and guide the transformation of agricultural growth mode [1] in China. During the period of the eleventh five-year-planning, various types of agricultural industrialization

J. Li (✉)
Faculty of Transportation Engineering, Kunming University of Science and Technology, Kunming, China
e-mail: lijiemei2@yahoo.com.cn

Y. Gu
Faculty of Management and Economics, Kunming University of Science and Technology, Kunming, China

H. Wu
Faculty of Architectural Engineering, Kunming University of Science and Technology, Kunming, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_145,

organizations in China came up to 250,000 and the farmer professional cooperatives reached to 370,000, the number of farmers driven by agricultural industrialization arrived 10.7 million [2]. At present the "company & farmer" is a popular industrialization mode in China. It has strong intrinsic rationality—it can effectively solve the problem of internal diseconomy of scale brought by the small farmers.

However in the practice of agriculture industrialization, the cooperative relationship between companies and farmers is extremely unstable. The breaching of contract phenomenon between farmers and companies is serious. Facing the plight of the farmers' interest losses, companies' operating risk increases, economies of scale forms difficultly.

Many scholars investigated the issues on the cooperation between company and farmer from different angles, such as the behavior choice of the company and farmer in participating in contract farming, and the measures of improving cooperation stability etc. However most of the existing researches only analyzed the superficial phenomenon of the problem. The research on the immanent evolution mechanism of "company & farmer" system hidden behind the phenomenon of interactive and disjointed relationship between company and farmer is lacked.

Since Prigogine put forward the Theory of Dissipative Structure and Haken founded Synergetic in 1970s, self-organization and synergetic theory has been widely applied in management science and system science [3]. Self-organization theory provides a new research angle for the analysis of the formation and evolution mechanism of "company & farmer" agricultural industrialization system. As a complex open system, "company & farmer" system is a typical dissipative system. Its development and evolution accord with the law of self-organization.

This paper attempts to establish a self-organization evolution model based on the analysis of the system self-organization evolution mechanism. It reveals the formation and evolution process and mechanism of the "company & farmer" system. It provides scientific basis of decision-making both for company and farmers.

145.2 The Self-Organization Evolution Mechanism of "Company & Farmer" System

Cooperation and competition is the natural attributes of "company & farmer" system. Their interactions are the dynamic sources of self-organization evolution of the system. The degree of the interactions determines the order and stability of the system [4]. The cooperation arising mainly from longitudinal relevance between the company and farmers based on different value creation segment. Mainly displays in two aspects: on the one hand, enterprises rely on alliance farmers to provide the primary products or raw materials, needn't to rent land and establish farms themselves. This can not only form a stable source of raw materials, but also save a lot of capital and cost. The company can focus on product development and

technological innovation, and then gain more profit. On the other hand alliance farmers rely on dragon-head enterprise to sale their primary products. The farmers not only have a stable market, but also have access to technical services, and can share the processing value-added profit of the agricultural products.

Competition effects are mainly derived from the development difference of the main body in the system. This kind of difference is mainly derived from two aspects: one is from the asymmetry distribution of the cooperative income, the other is from the development difference of main body in the system, such as comprehensive strength, brand reputation, learning ability and adaptability. The company is relatively large in scale, strong in economic strength and strong in information collection and application ability. So it often holds the interest distribution initiative. While the farmers are often minor in production scale, weak and dispersive in strength and lack of organization on their behalf. As an independent market main body, both the company and the farmers pursue the maximizing of their own interests. Under the environment of incomplete and asymmetric information, it has the possibility for both sides to breach the contract because of the driving of opportunistic. When the market price is higher than the contract price, the farmers have the strong motivation to sale their products to market. Conversely, when the market price is lower than the contract price, the company is inclined to buy agricultural products from the market.

In the cooperation and competition system of company and farmer, the main function of cooperation is to make interdependence between company and farmers, and change the extent of interdependence through the feedback of cooperation benefit distribution. The main role of competition is to bring the pressure of survival and development for the company and farmers. This kind of pressure can not only promote the collaborative development for the company and the farmers on the basis of different value creating segment, but also make intense contention for the income distribution between company and farmers.

145.3 Self-Organization Model of the System

In order to quantitative analysis the cooperation and competition influence on the scale development of the company and farmer and for researching the evolution process of the system, a evolution model is established on the basis of block growth model.

Block growth model (Logistic model): $\frac{dx}{dt} = rx(1 - \frac{x}{km})$, $x(0) = x0$, is put forward by Verhulst—a Holland biologist, based on the research on the population growth change law in the middle of the nineteenth Century. In the equation r represents the inherent growth rate of population; the factor rx reflects the population growth trend itself. Because of the block effect on the population growth from natural resources, environment etc., and the growth rate will drop after population growth to a certain number. And with the increase of population,

the block effect becomes more and more large. *km* represents the largest population that the natural resources and environmental conditions can accommodate, called population capacity. The factor $\left(1-\frac{x}{km}\right)$ reflects the block effects on population growth because of the limited resources. Block growth model can not only describe the population and many species variation law, but also has a wide range of applications in social economic fields [5].

The scale of company and farmer on the one hand expands constantly because of the inherent growth rate, on the other hand the growth rate slows down because of the constraints of resources and environment. And finally tends to a steady value. In addition, the company and the farmer can break their own maximum size limits and reach a higher state due to the cooperation. At the same time, the growth of both the scale are blocked as the mutual competition between company and farmer.

It selects the scale of company and the scale of farmer as the order parameters to describe the evolution process of "company & farmer" system.

Hypothesis:

(1) The growth rate of the company scale and farmer scale remains constant in a particular stage of development;
(2) Both the company scale and the farmer scale are continuous, differentiable function of time;
(3) The increase of the company scale and the farmer scale accords with logistic growth regularity;
(4) There are only a company and a farmer in the system.

The system evolution model is as follows:

$$\begin{cases} \dfrac{dx1}{dt} = z1\left(1-\dfrac{x1}{k1}-b12\dfrac{x2}{k2}+a12\dfrac{x2}{k2}\right)x1 \\ \dfrac{dx2}{dt} = z2\left(1-\dfrac{x2}{k2}-b21\dfrac{x1}{k1}+a21\dfrac{x1}{k1}\right)x2\ . \end{cases} \tag{145.1}$$

In Eq. (145.1):

$x1, x2$ The scale of company and the scale of farmers at t moment;
$z1, z2$ Constant growth rate of company and farmer which is able to achieved by relying only on their own ability;
$k1, k2$ The largest scale of company and farmer which can be formed under the restriction of the scarcity of economic resources;
$b12, b21$ Competitive effect coefficient, represents the impact on the company from farmer's competition and the impact on the farmer from the company's competition respectively;
$a12, a21$ Cooperative effect coefficient, represents the impact on the company from farmer's cooperation and the impact on the farmer from the company's cooperation respectively;

Factor $\left(1-\frac{x1}{k1}\right)$ and $\left(1-\frac{x2}{k2}\right)$ reflect block growth action on company and farmer respectively caused by the consumption of limited resource. *bij* and *aij* are competitive effect coefficient and the cooperative effect coefficient respectively, they indicate the effects on development scale of both sides caused by the cooperation and competition between company and farmer. It reflects the internal nonlinear interactions.

145.4 Model Simulations

According to Eq. (145.1), a system dynamics model is established by using the VensimPLE software (as shown in Fig. 145.1) [6].

The system evolution process will be simulated under the conditions of different parameters [7]. The initial values of parameters are as follows: $X_{10} = 3$, $X_{20} = 1$, $k_1 = 5$, $k_2 = 5$.

(1) There are both cooperation and competition between company and farmer at the same time, and the positive effect caused by cooperation is greater than the negative effect caused by competition. The growth rate which can be achieved by their own core competencies is same. Make $z_1 = 0.06$, $z_2 = 0.06$, $a_{12} = 0.8$, $b_{12} = 0.5$, $a_{21} = 0.8$, $b_{21} = 0.5$. The simulation results are shown in Figs. 145.2 and 145.3.

The results show when the benefits of cooperation outweigh the benefits of competition, both company scale and farmer scale exceed the maximum size which is attained only relying on their own ability under the limit of scarce resources.

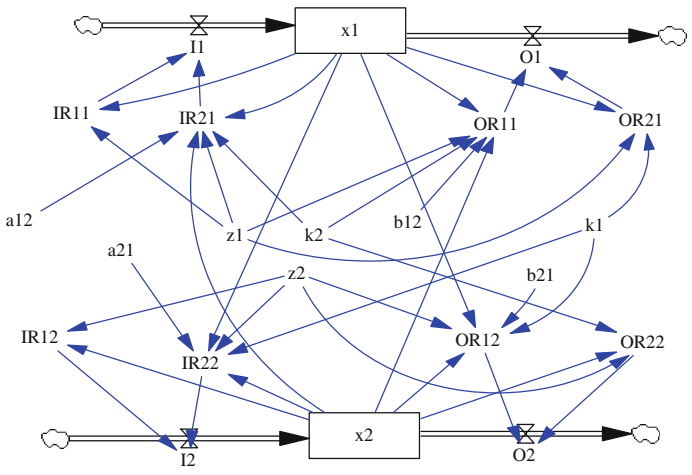


Fig. 145.1 "Company + farmer" cooperation and competition system dynamics model

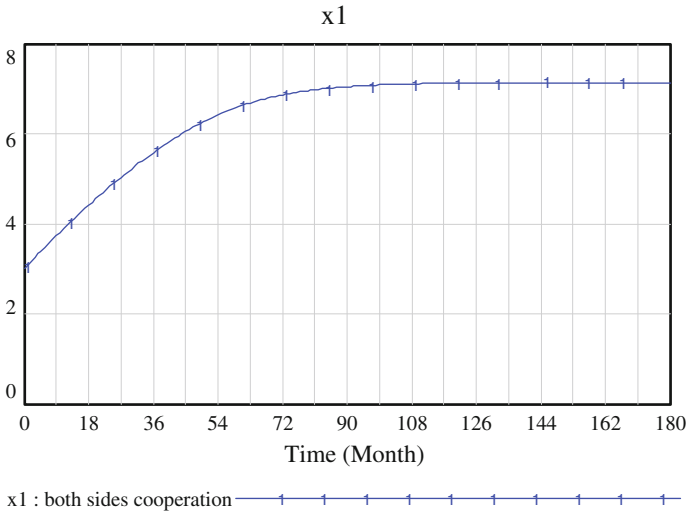


Fig. 145.2 The scale evolution process of company under effective cooperation

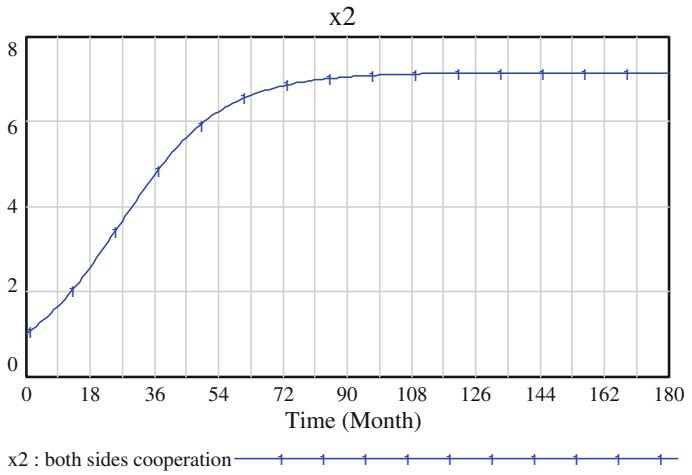


Fig. 145.3 The scale evolution process of farmer under effective cooperation

When the initial scale of the company and farmer is changed, make $X_{10} = 3$, $X_{20} = 0.5$, and other conditions keep the same, the system simulation results are not changed. This illustrates that the final running results of the system have no relationship with the initial size of company and farmer, and it is only related to the cooperative and competitive state between company and farmer.

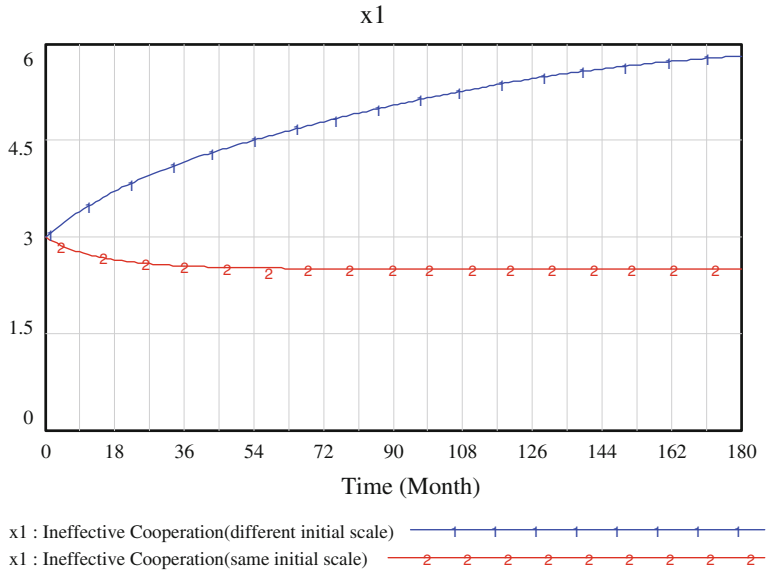


Fig. 145.4 The scale evolution process of company under ineffective cooperation

(2) When competition and cooperation exist between company and farmer at the same time, but cooperation brings negative effect, make $z_1 = 0.06$, $z_2 = 0.06$, $a_{12} = -0.8$, $b_{12} = 0.5$, $a_{21} = -0.8$, $b_{21} = 0.5$. Compare the two cases: the company and the farmer have the same initial size($X_{10} = 3$, $X_{20} = 3$); and the two parties have different initial size ($X_{10} = 3$, $X_{20} = 1$).The simulation results are shown in Figs. 145.4 and 145.5.

The results show that the initial size larger side can reach a maximum size, but the initial smaller side goes to die when the cooperation between the two sides brings negative effect and competition is intense. That is, the "company & farmer" system will break up when both the company and the farmer break a contract and there is a great discrepancy between the initial size of the two sides.

When the initial size of the two sides is identical, the operational efficiency is low and the final scale is smaller than the initial scale due to the negative effects of competition, but both sides can survive.

(3) There are competition and cooperation between company and farmer at the same time, but only one side is cooperative, and the cooperation effect from the other side is negative, make $z_1 = 0.06$, $z_2 = 0.06$, $a_{12} = -0.8$, $b_{12} = 0.5$, $a_{21} = 0.8$, $b_{21} = 0.5$ (farmer breaks contract and company is cooperative); *and* $z_1 = 0.06$, $z_2 = 0.06$, $a_{12} = -0.8$, $b_{12} = 0.5$, $a_{21} = 0.1$, $b_{21} = 0.5$ (farmer breaks contract and company is inefficient cooperative). The simulation results are shown in Figs. 145.6 and 145.7.

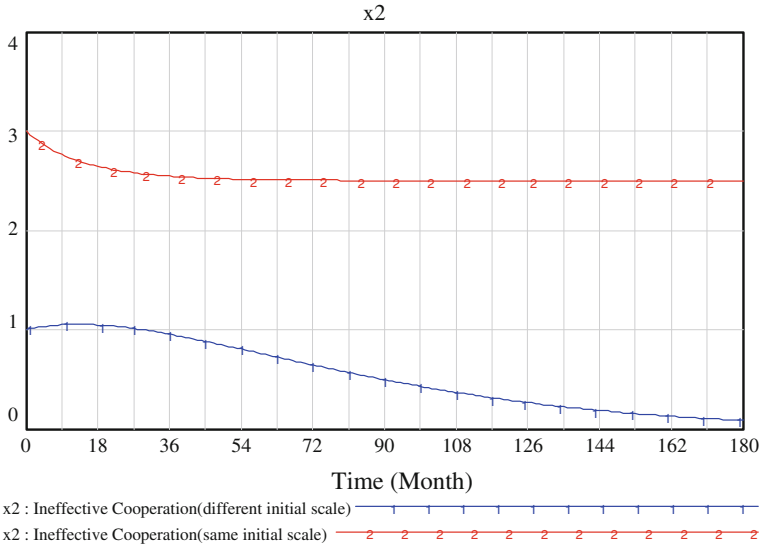


Fig. 145.5 The scale evolution process of farmer under ineffective cooperation

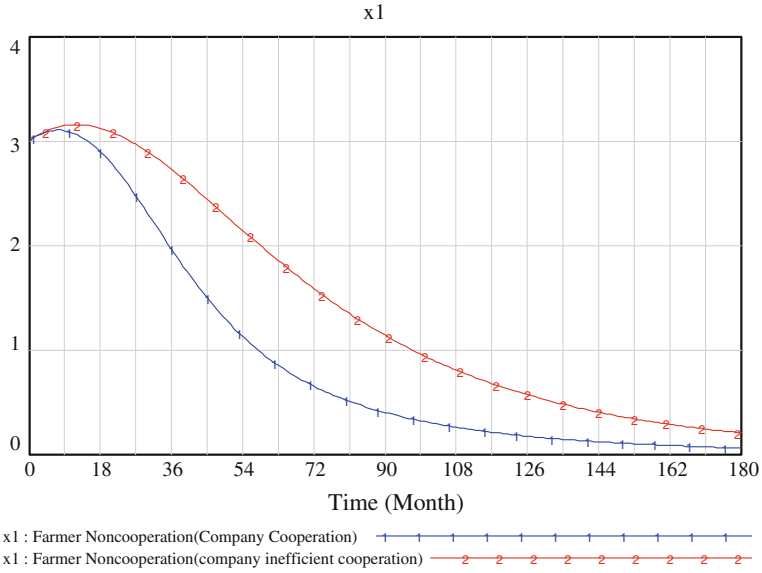


Fig. 145.6 The scale evolution process of company under farmer break contract

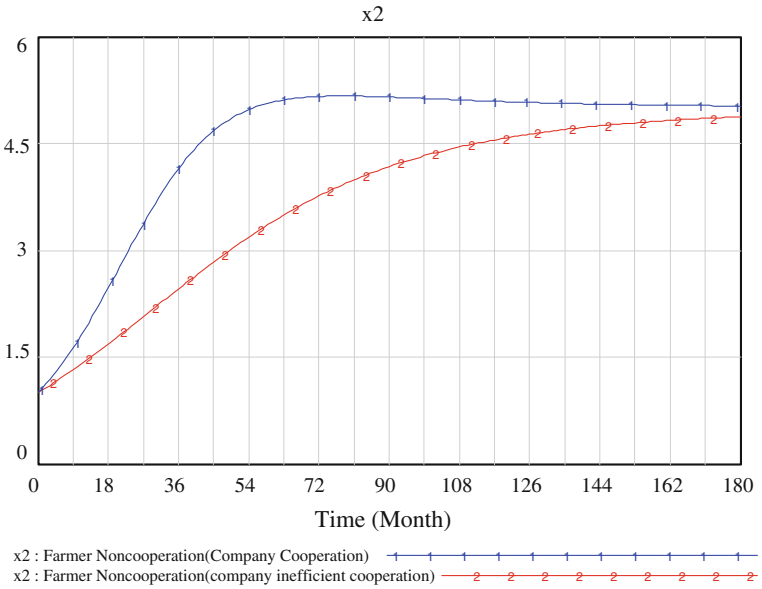


Fig. 145.7 The scale evolution process of farmer under farmer break contract

The results show that the company will go to die if the company continues to cooperate when the farmer breaks the contract. While the farmer will reach a certain scale, but can not achieve the optimal state. If the company takes negative cooperation ($a_{21} = 0.1$), the system would be in a more inefficiency state.

At the case of company breaches the contract and farmer is cooperative, make $z_1 = 0.06$, $z_2 = 0.06$, $a_{12} = 0.8$, $b_{12} = 0.5$, $a_{21} = -0.8$, $b_{21} = 0.5$, the system simulation results are similar: the farmer is eventually die.

This shows that as long as one side of company and farmer break the contract, the other party will go to die. And the breach side itself also can't achieve the optimal operational results. This causes destruction to both sides.

145.5 Conclusion

This paper establishes a system evolution model about cooperation and competition between company and farmer on the basis of analyzing the evolution dynamic mechanism of "company + farmer" system. It analyzes the stability of the model and simulates the evolutionary process. The following conclusions are obtained by simulation:

(1) The evolution results of the "company & farmer" system are closely related to the cooperation and competition effect. It's useful to amplify the cooperation

effect in the system, making the beneficial operation mode the leading role, and thus guide the system to the correct direction of evolution.

(2) Actively promote the organization of the farmers to expand the scale of agricultural production. In order to avoid the inherent defect of scale asymmetry of "company & household" mode, it's necessary to build new agricultural industrialization mode, such as "company + cooperation organization (the intermediary) + farmer" etc.

Acknowledgments This paper is supported by the Philosophy and Social Science Planning Project of Yunnan Province (NO.HZ201129).

References

1. Zhandong, Z., Mingshen, Z., et al.: On the cooperation mechanism of Company & Peasant household: practise and exploration, the transformation of institution and the trend of future. J. Manag. **23**(1), 62–67 (2010)
2. Data source: The twelfth five-year plan of agriculture and rural economic development of China
3. Simatupang, T.M., Sridharan, R.: Design for supply chain collaboration. Bus. Process Manag. J. **14**(3), 401–418 (2008)
4. Bengtsson, M., Kock, S.: Coopetition in business networks-to cooperate and compete simultaneously. Ind. Mark. Manage. **29**(5), 411–426 (2009)
5. Dongsheng, M.: Essentials of Systems Science, pp. 145–148. China Renmin University Press, Beijing (2010)
6. Yongguang, Z., Xiaojing, J.: System dynamics, pp. 97–103. Science Press, Beijing (2009)
7. Jiemei, L., Youjin, G., et al.: Combined evolution of self-organization and hetero-organization of company & household system. J. Kunming Univ. Sci. Technol. **12**(2), 80–86 (2012)

Chapter 146
Performance Evaluation of Distribution in the Distribution Center of Chain Supermarket Based on Unascertained and Analytic Hierarchical Model

Hongyu Li, Jichang Dong, Peng Gao and Xianyu Meng

Abstract In order to improve efficiency, meet diversified demands of stores; satisfy customers and perfect distribution activities, performance evaluation of distribution in the distribution center of chain supermarket is conducted. Firstly, the key factors which affect the performance of distribution center of chain supermarket are analyzed and evaluation indexes system is established in the paper. Then the index weights are determined by Analytic Hierarchical Model (AHM) and comprehensive evaluation is conducted with the help of unascertained measure theory. Finally, the method which is a probe into performance evolution of distribution center, has been proved applicable and reasonable with the example.

Keywords Chain supermarket · Distribution center · Performance evaluation · Unascertained theory · AHM

146.1 Introduction

Chain supermarkets whose industry form and management model are representative of retail industry have became an important part of chain operations. As a core technology of chain retail, logistic distribution technology is the key factor to achieve profits, improve efficiency and customer satisfaction, and create customer loyalty.

H. Li (✉) · J. Dong · P. Gao
School of Management, University of Chinese Academy of Sciences, Beijing, China
e-mail: hong_yu_li@163.com

X. Meng
Beijing Union University PingGu College, Beijing, China

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_146,

Many studies show that a company who enjoys a high level in the development and application of evaluation will get excellent performance: its productivity will increase from 14 to 20 % [1].

At present, most research on performance evaluation of distribution has taken the efficiency of logistics operations and services as evaluation indexes; little research has done on distribution costs, economic benefits and potential development of distribution. Most studies just took inner factors of logistics enterprise into account, showed less consideration to the influence of external factors and lack of a scientific evaluation indexes system on distribution performance. Furthermore, the methods of performance evaluation mainly focus on theoretical study, despite many mathematical, statistical methods and various evaluation methods which have been put into good use, such as: fuzzy Comprehensive Evaluation Method [2], efficiency coefficient method coordinates with comprehensive analysis and judgment method [3], data envelopment analysis (DEA) [4], comprehensive evaluation method of two-stage logistics system (DEA/AHP), Analytic Hierarchy Process method (AHP) [5], fuzzy clustering method [6], utility theory method [7]. Every method also bears some defects, for example, the application of DEA which needs lots of mathematical knowledge as a basis and its practicability is weak; it is subjective to fix standard weighing by using expert assessment method in fuzzy clustering method; the calculation process of utility theory method is very complicated and if unsuitable Benchmarking Enterprises are chosen, the final result will not be satisfied.

So the paper proposes systematic performance evaluation indexes of chain supermarket DC firstly which is established by analyzing the key factors affecting the performance of the chain supermarket DC. Secondly, using Attribute Hierarchy Model (AHM) which is simpler in operation than AHP to fix the standard weight so that the structured decision theory and unascertained measure have been integrated to comprehension evaluation, then a reasonable confidence identification criteria and sort marking criteria is obtained, and finally it achieves the combination of the structured decision theory and unstructured decision theory. All those make the evaluation result more clear and reasonable.

146.2 The Key Factors Which Affect the Performance of Distribution

The performance of distribution refers to a procedure of allocation, achievements of its result and efficiency situation. That is about how to achieve the fastest response speed, lowest cost of distribution, and the best delivery results under the condition of limited resources in DC. The paper mainly considers the following factors which influence the performance of distribution in the DC of chain supermarket [8–12].

(1) Distribution processes influence the performance of the distribution.
Distribution process is the most important direct factor in the efficiency of delivery, and the necessary condition of distribution activity. Particularly the operation of the key procedure plays an important role in the entire delivery activity.
(2) The information level influences the performance of the distribution.
Modern logistics information technology favors the control of stock, raises the efficiency of allocation, and reduces the cost of allocation.
(3) Equipments influence the performance of distribution
The choice and utilization of equipments in DC are playing decisive roles in mode of operation of DC and process of distribution, and they are important to the durative and continuity of the entire process of distribution.
(4) Management level influences the performance of distribution
The result that the distribution works will or won't achieve the given objective and done efficiently depends on the height of management level. It is reflected in how to coordinate the schedule of the whole distribution and leadership between departments and coordinating the relationship of exterior customers and whether to introduce new management ideas and concepts or not in development strategy, etc.
(5) The cost influences the performance of distribution
On the basis of meeting distribution needs of chain supermarket, the cost not only examines whether the process of distribution is reasonable and effective or not, but also reflects the degree of conformity and utilization of DC resources. Cost is the most direct economic reflection of distribution performance.
(6) The feedback of distribution effect influences the performance of distribution
The final comprehensive effect of the whole distribution can be obtained by the feedback of allocation effect which mainly displays in the level of customer service, such as these external targets: degree of customer satisfaction, rate of customer increment and market acquisition rate, etc.

146.3 The Establishment of Indexes System for Performance Evaluation of Distribution Center

Taking the distribution process and key factors into consideration, this paper combines the logistics activities of DC with material flow, information flow and capital flow of supply chain to establish the overall performance evaluation indexes system for DC of chain supermarket which is shown as Table 146.1.

The indexes system is built up by analyzing the whole business process objectively and giving attention to both low cost and customer service, which not only focuses on economic efficiency, but also pays attention to quality of service. It makes a holistic analysis of performance evaluation and avoids overlap and duplication of indicators.

Table 146.1 Indexes system for performance evaluation of DC

business process I_1					the level of information technology I_2			equipment I_3			cost I_4				management level I_5			feedback I_6			
order processing I_{11}	inventory I_{12}	sort I_{13}	processing I_{14}	dispatching I_{15}	timeliness I_{21}	accuracy I_{22}	level of MIS I_{23}	utilization ratio I_{31}	reliability I_{32}	advancement I_{33}	inventory cost I_{41}	transportation cost I_{42}	information processing cost I_{43}	administration cost I_{44}	planning situation I_{51}	inner coordination I_{52}	advance of method I_{53}	delivery consistency I_{61}	time restriction I_{62}	flexibility I_{63}	customer satisfaction I_{64}

146.4 Unascertained Measure Model

Supposed that $X = \{x_1, x2 \ldots xn\}$ means evaluation objects set; there are m indexes such as I_1, $I2$, *and Im* of $x_i (x_i \in X)$ that is $I = \{I_1, I2 \ldots Im\}$; x_{ij} is the observed value of object x_i under index I_j; $C = \{c_1, c_2 \ldots c_k\}$ is the evaluation space, thereinto c_k $(1 \leqq k \leqq K)$ *is* the k_{th} evaluation grade.

(1) Recognition of Single Index. On the condition that x_{ij} is given, for every single-factor index (attribute) $I_j (j = 1, 2 \ldots m)$, the value of μ_{ijk} which is the measure of observed value x_{ij} falls into grade $c_k (k = 1, 2 \ldots K)$ can be calculated.

Structuring the measure function $\mu_{ij}(x)$ and calculating the μ_{ijk} for every grade $k (k = 1, 2 \ldots K)$, the unascertained measure recognition matrix under the single index can be obtained as follows:

$$\mu_i = \begin{pmatrix} \mu_{i11} & \mu_{i12} & \cdots & \mu_{i1k} \\ \mu_{i21} & \mu_{i22} & \cdots & \mu_{i2k} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{im1} & \mu_{im2} & \cdots & \mu_{imk} \end{pmatrix} = \left(\mu_{ijk}\right)_{m \times n} \tag{146.1}$$

(2) Determination of Index Weight based on Attribute Hierarchy Model (AHM). AHM is a non-structure decision method, and it is improved from AHP. Compared with AHP, AHM is much easier to be applied since AHM does not need to calculate eigenvector and check the consistency, which only needs to make multiplication and addition operation.

Steps of calculating weight based on AHM are as follows [13].

(1) There are many influence factors of distribution performance in the DC of chain supermarket, so it needs many experts to participate in the evaluation. The basic idea is to evaluate the index's importance on each level separately

by the experts firstly. Then the experts calculate the arithmetic average of the index in corresponding level to get the synthetic evaluation result. In identical level, the various indexes get corresponding importance by comparing. Suppose that there are n factors b_1,b_2,…, b_n, if the importance of b_i is the same as the importance of b_j, then $b_{ij} = 1$; if b_i is slightly important than b_j, then $b_{ij} = 3$; if b_i is obviously important than b_j, then $b_{ij} = 5$; if b_i is more important than b_j, then $b_{ij} = 7$; if b_i is absolutely important than b_j, then $b_{ij} = 9$. Between them there are $b_{ij} = 2,4,6$ or 8. It is obvious that $b_{ij} = 1/b_{ji}$.

(2) Transforms 1–9 scale judgment matrix into AHM, and the transformation procedure is as follows:

$$\mu_{ij} = \begin{cases} \frac{2k}{2k+1} & a_{ij} = k \\ \frac{1}{2k+1} & a_{ij} = \frac{1}{k} \\ 0.5 & a_{ij} = 1 \;\; i \neq j \\ 0 & a_{ij} = 1 \;\; i = j \; k \text{ is positive integer and } k_i \geq 2 \end{cases} \tag{146.2}$$

It is Obvious that $\mu_{ij} = 0$, $\mu_{ij} \geq 0$, $\mu_{ji} \geq 0$, $\mu_{ij} + \mu_{ji} = 1(i \neq j)$, μ_{ij} is called the measure based on AHM. When $\mu_{ij} \geq \mu_{ji}$, it means that the plan P_i is better than the plan P_j.

(3) Make that

$$f_i = \mu_{i1} + \mu_{i2} + \ldots + \mu_{in} = \sum_{j=1}^{n} \mu_{ij} (i = 1, 2, \ldots, n) \tag{146.3}$$

$$c_i = 2f_i/(n^*(n-1)) \tag{146.4}$$

Thereinto, c_i expresses the score rate of μ_i, then $c = (c_1, c_2, \ldots, c_n)$ and $\sum_{i=1}^{n} c_i = 1$. According to the above, the importance order of each plan can be calculated.

(3) Recognition of multi-indexes.After the single index measured evaluation matrix of x_i and the index weights $w^{(i)}$ have been figured out, the unascertained measure recognition vector of x_i based on m indexes can be obtained as following.

$$\mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{ik})^T \tag{146.5}$$

thereinto,

$$\mu_{ik} = \sum_{j=1}^{m} w_j^{(i)} \cdot \mu_{ijk} (i = 1, 2, \ldots, m; \; k = 1, 2, \ldots, K) \tag{146.6}$$

μ_{ijk} is the measure of observed value x_{ij} falls into grade c_k. $w^{(i)}$ is index weight.

(4) Identified criterions. A confidence threshold is pre-determined called λ ($\lambda > 0.5$). According to the background and demands of the problem, λ is normally between 0.6 and 0.8.

Since evaluation space $\{c_1, c_2,\ldots,c_p\}$ is arranged in order.

$$k_0 = \min_k \left(k : \sum_{l=1}^{k} \mu_{il} \geq \lambda,\ 1 \leq k \leq K \right) \tag{146.7}$$

Sample x_i belongs to c_{k0} and the confidence is λ. The implication is that: the confidence that grade of x_i is not higher than c_{k0} is λ or the confidence of that grade of sample x_i is higher than C_{k0+1} is $1-\lambda$.

146.5 Empirical Analyses

There is a supermarket DC covering an area of more than 60,000 m^2. Average daily inflow and outflow of goods is 40,000 cases, refering food, non-staple food, household supplies, knitting, metal and other 40 categories 17,000 species.The DC is charged with responsibility to provide rich, full range, efficient and safe logistics services for 200 supermarkets under head office. In order to enhance the delivery capacity of DC and distribution of results, the paper conducts performance evaluation on the DC with the method of unascertained measured model and AHM.

146.5.1 Determination of Index Weight

(1) Selection of experts.Selecting relevant managers and experts from strategic,logistics, finance, customer relations, IT departments of the DC, then 20-member evaluation team is set up. Firstly, scoring every evaluation team by comparing between every two factors, then scores of each indicator are weighted average and rounding, data in judge matrix is gained finally.
(2) Calculation of weights.According to the index system above, the first level index vector is marked as $I_i = \{I_1, I_2, I_3, I_4, I_5, I_6\}$, thereinto, I_i ($i = 1,2,3,4,5,6$) represents the first i index of the level, and they are business process, the level of IT, equipment, cost, management and feedback correspondingly. Based on the above scoring method, the evaluation matrix is obtained. Similarly way is available for the second level evaluation matrix (Table 146.2).

By (146.2),it can be changed into judgment matrix of AHM which is shown as follows.

Table 146.2 Overall evaluations on the first-level index

I	I_1	I_2	I_3	I_4	I_5	I_6	w_i
I_1	1	3	5	3	1	1	w_1(0.2416)
I_2	1/3	1	3	1/3	1	1/3	w_2 (0.1190)
I_3	1/5	1/3	1	1/3	1/3	1/5	w_3 (0.0407)
I_4	1/3	3	3	1	3	1/3	w_4 (0.1905)
I_5	1	1	3	1/3	1	1/3	w_5 (0.1429)
I_6	1	3	5	3	3	1	w_6 (0.2654)

$$\begin{pmatrix} 0 & 0.8571 & 0.9091 & 0.8571 & 0.5000 & 0.5000 \\ 0.1429 & 0 & 0.8571 & 0.1429 & 0.5000 & 0.1429 \\ 0.0909 & 0.1429 & 0 & 0.1429 & 0.1429 & 0.0909 \\ 0.1429 & 0.8571 & 0.8571 & 0 & 0.8571 & 0.1429 \\ 0.5000 & 0.5000 & 0.8571 & 0.1429 & 0 & 0.1429 \\ 0.5000 & 0.8571 & 0.9091 & 0.8571 & 0.8571 & 0 \end{pmatrix}$$

By (146.3) and (146.4), it can conclude that

$$\mathrm{f} = (3.6234 \;\; 1.7857 \;\; 0.6104 \;\; 2.8571 \;\; 2.1429 \;\; 3.9805)$$

Weights are determined by normalization method and listed below.

$$\mathrm{c} = (0.2416 \;\; 0.1190 \;\; 0.0407 \;\; 0.1905 \;\; 0.1429 \;\; 0.2654)$$

Similarly way is available for the weights of the second level index.

$$\mathrm{c1} = (0.1663 \;\; 0.2766 \;\; 0.1611 \;\; 0.0402 \;\; 0.3558),\ \mathrm{c2} = (0.5524 \;\; 0.3333 \;\; 0.1143)$$

$$\mathrm{c3} = (0.5820 \;\; 0.3143 \;\; 0.1037), \qquad \mathrm{c4} = (0.4277 \;\; 0.3148 \;\; 0.2000 \;\; 0.0575)$$
$$\mathrm{c5} = (0.5887 \;\; 0.3333 \;\; 0.0779), \qquad \mathrm{c6} = (0.3148 \;\; 0.4425 \;\; 0.1757 \;\; 0.0670)$$

146.5.2 Comprehensive Evaluation of the Distribution Performance

(1) The Division of Evaluation Grade. In this paper, evaluation levels of distribution performance are divided into five grades: excellent, good, middle,poor, very poor.
(2) Evaluation on each sub-factor. This paper uses the jury scoring method to determine the degree of index I_i belonging to v_{ij}. Details are as follows. There are n experts in the judging panel, then the degree of a index I_i falling into the specific grade is defined as follows. r_{ij} = All the judges in a particular class is classified as V in the number of members for I_{ij}/n.

For the research, there are 20 experts and practitioners of the supermarket taking part in the questionnaire survey. The evaluation result is shown in Table 146.3.

$$\mu_1 = (0.1663\ \ 0.2766\ \ 0.1611\ \ 0.0402\ \ 0.3558)\begin{bmatrix} 0.2 & 0.3 & 0.5 & 0 & 0 \\ 0.2 & 0.4 & 0.4 & 0 & 0 \\ 0.2 & 0.3 & 0.4 & 0.1 & 0 \\ 0.4 & 0.4 & 0.1 & 0.1 & 0 \\ 0.5 & 0.3 & 0.2 & 0 & 0 \end{bmatrix}$$
$$= (0.3148\ \ 0.3317\ \ 0.3334\ \ 0.02010)$$

Similarly,

$$\mu 2 = (0.1886\ 0.4438\ 0.2676\ 0.1000\ 0),\ \mu 3 = (0.4375\ 0.3896\ 0.1311\ 0.0418\ 0)$$

$$\mu 4 = (0.1800\ \ 0.3370\ \ 0.4515\ \ 0.0315\ \ 0),\ \mu 5 = (0.3355\ \ 0.1744\ \ 0.3744\ \ 0.1078\ \ 0.0078)$$

$$\mu 6 = (0.4382\ 0.3690\ \ 0.1685\ 0.0243\ 0).$$

Then,

$$R = \begin{bmatrix} 0.3148 & 0.3317 & 0.3334 & 0.0201 & 0 \\ 0.1886 & 0.4438 & 0.2676 & 0.1000 & 0 \\ 0.4375 & 0.3896 & 0.1311 & 0.0418 & 0 \\ 0.1800 & 0.3370 & 0.4515 & 0.0315 & 0 \\ 0.3355 & 0.1744 & 0.3744 & 0.1078 & 0.0078 \\ 0.4382 & 0.3690 & 0.1685 & 0.0243 & 0 \end{bmatrix}$$

Weighs of the first level index is listed below.

$$c = (0.2416\ \ 0.1190\ \ 0.0407\ \ 0.1905\ \ 0.1429\ \ 0.2654)$$

Comprehensive evaluation vector of the DC performance is shown as below.

$$\mu = (0.2416\ 0.1190\ \ 0.0407\ 0.1905\ 0.1429\ 0.2654)^{*}\mathrm{R}$$
$$= (0.3148\ 0.3359\ 0.3020\ \ 0.0463\ 0.0011)$$

(3) Identification and sorting.Since the evaluation space is divided into five levels and arranged orderly, the rules of credible recognition is put to application. Set $\lambda = 0.7$, according to (146.7), it can be concluded that when $k_0 = 3$, then $0.3359 + 0.3020 = 0.9527 > 0.7$, which means the performance evaluation of the DC falls into the third grade: middle.

Table 146.3 Experts' evaluation results

First-level indexes	First-level weight	Second-level indexes	Weight	The corresponding values of experts reviews					The corresponding proportion of the experts				
				Excellent	Good	Middle	Poor	Very poor	Excellent	Good	Middle	Poor	Very poor
Business process	0.2416	Order processing	0.16630	4	6	10	0	0	0.2	0.3	0.5	0	0
		inventory	2766	4	8	8	0	0	0.2	0.4	0.4	0	0
		sort	0.16110	4	6	8	2	0	0.2	0.3	0.4	0.1	0
		processing	0402	8	8	2	2	0	0.4	0.4	0.1	0.1	0
		dispatching	0.3558	10	6	4	0	0	0.5	0.3	0.2	0	0
The level of IT	0.1190	Timeliness	0.5524	4	10	4	2	0	0.2	0.5	0.2	0.1	0
		accuracy	0.3333	4	8	6	2	0	0.2	0.4	0.3	0.1	0
		level of MIS	0.1143	2	6	10	2	0	0.1	0.3	0.5	0.1	0
Equipment	0.0407	Utilization ratio	0.5820	10	8	2	0	0	0.5	0.4	0.1	0	0
		reliability	0.3143	8	8	2	2	0	0.4	0.4	0.1	0.1	0
		advancement	0.1037	4	6	8	2	0	0.2	0.3	0.4	0.1	0
Cost	0.1905	Inventory cost	0.4277	4	8	8	0	0	0.2	0.4	0.4	0	0
		transportation cost	0.3148	4	6	10	0	0	0.2	0.3	0.5	0	0
		information processing	0.2000	2	6	10	2	0	0.1	0.3	0.5	0.1	0
		cost	0.0575	4	4	8	4	0	0.2	0.2	0.4	0.2	0
		administration cost											
Management level	0.1429	Planning	0.5887	8	2	8	2	0	0.4	0.1	0.4	0.1	0
		coordination	0.3333	6	6	6	2	0	0.3	0.3	0.3	0.1	0
		advance of method	0.0779	0	4	10	4	2	0	0.2	0.5	0.2	1
Feedback	0.2654	Delivery consistency	0.3148	10	8	2	0	0	0.5	0.4	0.1	0	0
		time restriction	0.4425	8	8	4	0	0	0.4	0.4	0.2	0	0
		flexibility	0.1757	8	6	4	2	0	0.4	0.3	0.2	0.1	0
		customer satisfaction	0.0670	10	4	4	2	0	0.5	0.2	0.2	0.1	0

146.6 Conclusion

This paper analyses the key factors affecting the performance of DC of chain supermarket, and conducts its performance evaluation by integrating unascertained and AHM theory. The unascertained measure method pays attention to "the ordered nature" of the evaluation space, and presents the reasonable confidence threshold and sorting criterion, which makes the evaluation result more clearer and more reasonable. Besides, the paper realizes the combination of qualitative analysis and quantitative analysis. Finally, this method has been proved applicable and reasonable with the example and is of reference value for the enterprises. Furthermore, the indexes system needs to be improved yet so as to meet the practice requirements. Thus more realistic evaluation results will be obtained.

References

1. Zhang, Y.: Internal performance evaluation of chain logistics enterprise based on grey system theory. Commun. Finance Account. **31**(2), 39–40 (2010)
2. Zhu, D., Zhang, X.: Study on distribution performance evaluation system for chain supermarket distribution centers. Logist. Technol. **31**(2), 124–126 (2012)
3. Sun, H.: Logistics Efficiency for Enterprises of Chain Operation. China Material Press, Beijing, pp. 1–26, 76–77 (2005)
4. He, M., Li, G.: Dynamic results evaluation of modern logistics management system. Math. Pract. Theory **33**(8), 56–58 (2003)
5. Wang, Y., Sun, L., Chen, H.: Logistics synthesis evaluating on DEA/AHP two-stage model. J. Xi'an Highw. Univ. **23**(3), 79–83 (2003)
6. Wei, X.: Applying the fuzzy classification method in evaluation of logistics performance. Logist. Technol. **23**(8), 29–32 (2003)
7. Ma, H., Zhang, G., Sheng, Y.: Application of utility theory at performance evaluation of logistic business enterprise. J. East China Shipbuild. Inst. (Nat. Sci. Edn.) **17**(6), 78–83 (2003)
8. Ding, J.: Research on the performance evaluation of distribution in the distribution center. Value Eng. **26**(6), 72–74 (2007)
9. Li, Y.: Design and application of the performance evaluation system in distribution center of chain supermarket. Master dissertation of Shanghai Jiao Tong University, Shanghai (2008)
10. Xuejin, S., Hu, F.: Status characteristic and pattern innovation of logistics distribution in chain supermarket. Storage Transp. Preserv. Commod. **27**(6), 17–18 (2005)
11. She, D.: Research on the performance evaluation of distribution in the Chain enterprise's distribution center. Master dissertation of Jinan University, Guangzhou (2009)
12. Zi, X., Ma, L.: Research on physical distribution of chain enterprises. Logist. Technol. **27**(9), 98–100 (2008)
13. Cheng, Q.: Analytic hierarchy process (AHP) and attribute hierarchical model (AHM). Syst. Eng. Theory Pract. **17**(11), 56–59 (1997)

Chapter 147
An Application of Radio-Frequency Identification Systems in Chinese Clothing Supply Chain

Luyang Liu

Abstract China, as a big clothing consumer and producer in the world, has made great progress in clothing field in the past several decades. However, Chinese clothing supply chain management is still at a low level. Because of several characteristics such as strong seasonal difference and fast changes in consumer demand, clothing industry needs a high efficient supply chain to satisfy these changes in demand. Radio-Frequency Identification (RFID) technology can provide technical support for this efficient supply chain. Radio-Frequency Identification (RFID) technology is a wireless automatic identification technology that uses radio-frequency electromagnetic fields to transfer data between a RFID tag and its reader without any contact. Applying RFID technology in clothing supply chain can not only improve its efficiency, but also achieve the goal of real-time tracking products and solve many tough problems. In this paper, a practical RFID system application in Chinese clothing supply chain is provided, including the application framework, selection of RFID tags and configuration of RFID system.

Keywords RFID · Clothing supply chain · Logistics network

147.1 Introduction

China, with a population of 1.3 billion, is one of the world's largest clothing consumers and producers. In recent years, with the increasing progress in the scale of productions and standards, China's clothing industry has made a great development. However, the backward management methods of Chinese clothing enterprises led to low efficiency of supply chain operations, high inventory and

L. Liu (✉)
School of Information and Communication Engineering, Beijing University of Posts and Telecommunications (BUPT), Beijing, China
e-mail: liuluyang.colin90@gmail.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_147,

other issues, which have a direct impact on the competitiveness of Chinese clothing brands.

The clothing industry has several characteristics: very strong seasonal difference, extremely fast changes in consumer demand and fast inventory turnover. To cater these characteristics, manufacturers have to distribute products to all retail stores in a short period of time after completing the production. This requires the formation of a rapid reaction supply chain between stores, companies and manufacturers, which could real-time share the information concerning store sale, the company's procurement, production and inventory situation. In this way, clothing manufactories could accurately acquire the directions and changes of staff, logistics, information and money on the entire supply chain.

Most of the domestic clothing enterprises is relatively backward in the technical support, and the goods spend lots of time on warehouse and picking processes resulting in a significant gap of inventory turnover time between international companies and domestic enterprises, which in a certain extant limit the development of local enterprises.

Clothing enterprises need the help of a powerful supply chain management tool to improve the efficiency of the supply chain logistics. RFID technology is a proper choice. Clothing attached with RFID tags could be tracked and analyzed in the whole clothing supply chain, in order to rationalize the logistics process, at the same time bring added value to customers by improving service levels. This application, as a result, enables enterprises to significantly increase market share and meet the needs of customers [1].

Radio-Frequency Identification (RFID) technology is a wireless automatic identification technology that uses radio-frequency electromagnetic fields to transfer data between a RFID tag and its reader without any contact [2]. RFID tags are small, cheap and uniquely identifiable devices offer enormous potential [1]. RFID tags could support a larger set of unique IDs than bar codes and can incorporate additional data such as manufacturer, product type, and even measure environmental factors such as temperature. Besides, RFID systems can discern many different tags located in the same general area without human assistance [3].

However, for a long time, RFID technology has been far from mainstream applications because of high cost of the whole application system. However, RFID technology has been gradually implemented in many fields. With the advantages of automatic identification and tracking objects in item level, RFID tags have been highly valued in many industries, especially in the area of stock and inventory control [1]. Commodities attached with RFID tags can be identified from a distance without requiring a line of sight [2]. Thus, RFID technology makes it possible to develop a real-time supply chain, which is entirely automatic [4].

Three major organizations are pioneers of large-scale adoption of RFID technology: Wal-Mart, Tesco, and the US Department of Defense who saw the tremendous benefits from RFID technology first. They mandated that their largest suppliers begin tagging all pallets of goods delivered to their warehouses with RFID tags and used RFID to lower their operational costs by streamlining the tracking of stock, sales, and orders [3, 5].

RFID technology in the clothing industry is mainly concentrated in manufacturers and retailers in developed countries such as Europe and the United States, including UK Marks & Spencer Group PLC, the U.S. GAP Co., Ltd., famous German clothing manufacturer Gardeur AG, etc. With the application of RFID technology to be more widely cognitive, Chinese market also began to explore the related RFID applications.

China has explored RFID applications in many areas: (1) Identification and Access Control; (2) Certification and Anti-Counterfeiting (3) Logistic (4) Animal Identification (5) Ticketing [5] Some Chinese clothing enterprises, including Li Ning, Semir, Xtep, Joeone, etc., also consider applying RFID technology to their supply chains to improve efficiency.

In summary, there are many applications of RFID technology in the field of European and American clothing industries, but Chinese clothing industry is still in the exploration stages. How to combine supply chain management status in Chinese clothing industries with RFID technology, how to improve the operational efficiency of the supply chain, and how to improve the level of supply chain management have become important issues which Chinese clothing enterprises are facing.

147.2 System Application in Clothing Distribution Supply Chain

147.2.1 Structure of Clothing Distribution Supply Chain System

Chinese clothing enterprises typically have three kinds of marketing patterns: agency, franchisee and direct sale. In the first two patterns, distribution supply chain must be multi-level. Because of the large territory of China, even though the latter is a direct selling pattern, its distribution supply chain might also be a multi-level one. As a result, multi-level distribution network is common in Chinese clothing industries. Figure 147.1 shows a typical framework of clothing supply chain logistics network.

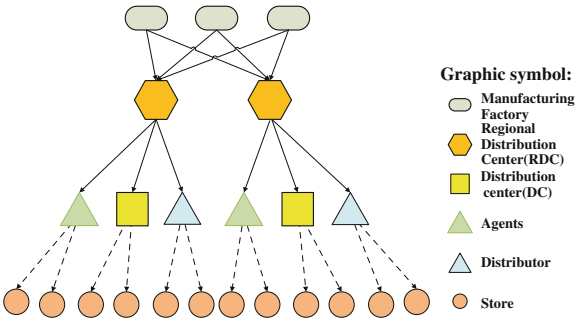


Fig. 147.1 The framework of clothing supply chain logistics network

In this framework, products produced by enterprises themselves or original equipment manufacturer (OEM) factories are sent to Regional Distribution Center (RDC) first, then these products will be sent to other subordinate distribution centers or agents' warehouses, who will accomplish the terminal transportation to stores.

147.2.2 Clothing Supply Chain RFID System Technology Framework

In the clothing distribution logistics process, products have to pass 4 nodes: factory, Regional Distribution Center (RDC), distribution center (or agents' warehouse), stores and 3 transport links: entering warehouse, distribution, terminal distribution at least in order to be sent to customers. Applying RFID technology in the clothing supply chain logistics system, RFID tags attached to clothes transport with them, thus it can be served as an information carrier which could real-time share information concerning products and its logistics process, as a result, enhance the competitiveness of the entire clothing supply chain. Figure 147.2 shows the RFID application in clothing supply chain.

147.2.3 RFID System Structure and Transmission Principle

RFID system consists of two parts: hardware part and software part. The hardware part consists of RFID tags, RFID readers, antennas, auxiliary modules and mainframe machines. The software part consists of RFID middleware and main database which could be connected through internet by mainframe machines at any time.

RFID tags consist of coupling components and chips, and each tag has a unique electronic code, identify the target object the tag attached. Under the control of mainframe machines and the RFID middleware, RFID reader can read information carried by RFID tag or write information to the RFID tag through its antenna. The

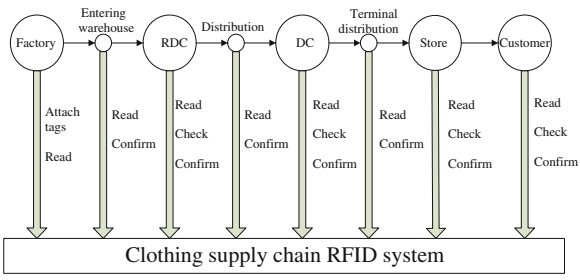


Fig. 147.2 RFID application in clothing supply chain

hardware part can be used to form a data collection area, which can automatically exchanging information with all RFID tags when pallets loaded with products pass though this area. All information can be aggregated to the real-time main database through internet. An organization called EPCglobal has created a worldwide standard for RFID and use Internet to share data via the EPCglobal Network. Thus enterprises can purchase the whole RFID systems complying EPCglobal standard.

147.2.4 RFID Tags Selection Application

RFID tags could be differed according to the criteria concerning the power source, reader interaction, and the computational capabilities [1]. RFID devices have 3 power source patterns, including passive way, semi-passive way, and active way, and four main communication bands of electro-magnetic field might be used. These bands are: (1) the low frequency (LF); (2) the high frequency (HF); (3) the ultra-high frequency (UHF); (4) a band at 2.4 GHz (only used in some active tags) [7]. A computational capability of RFID tag is also different in many ways. Besides these factors, whether to attach some sensor on RFID tags is also worth well deliberating. Therefore, how to select the most proper kind of RFID tag is vital significant.

147.2.4.1 Power Source Selection

RFID devices have 3 power source patterns, including passive way, semi-passive way, and active way. Table 147.1 shows comparison between these 3 power sources.

In sum, considering these 3 kinds of RFID devices, the most active one is obviously the most functional one. Active RFID tags can carry with different kinds of sensors to monitor the environment of transport and storage. However, for the retail trade, active tags are impractical. Passive tags, thanks to the fact that they don't carry with batteries or other power sources and low costs, are the most prospective choice for item-level use [1, 3].

Table 147.1 RFID tags power source comparison

Power source	Instruction	Characteristics	Applications
Active way	An on-board battery, always active	Large capacity, long range, expensive	Carry with sensor
Semi-passive way	A small battery, activate by RFID reader	High speed	RFID payment card
Passive way	No battery, use radio energy as its power source	Small, cheap	Item-level use

147.2.4.2 Communication Band Selection

In general, passive RFID tags work on 3 communication bands: LF, HF and UHF band. Table 147.2 shows comparison between different types of communication bands concerning their frequencies, characteristics and other remarks.

In sum, LF RFID tags are not proper in clothing industry. It seems to conclude that UHF RFID devices are the best choices in manufacturing line. However, HF RFID devices also have its own advantages in security performance and its mature technology research. Thus, both HF and UHF could be used in the RFID clothing distribution application. Here are some substantial methods of application about these different tags.

For UHF RFID tags, using RFID tags to replace traditional barcode is the most easy and convenient way. These tags could be embedded to cards and attached to their target clothes using plastic ropes. Because of its advantages in transmission range, exchanging speed and cost, UHF RFID tags could be the most suitable tags for item-level use in clothing supply chain. These tags could retrieve easily after their target clothes have been sold.

HF RFID tags, thanks to its good security performance, can combine planar-fashionable circuit board (P-FCB) technology to embed the whole circuit of RFID tags into clothes. P-FCB technology is a technique to make entire electronic circuits, like RFID tags and their antennas, on a fabric board. This technique could protect RFID tags from being destroyed unless their target clothes have been washed for more than 20 times, so that these tags can trace their clothes for longer time, record the process of sell and return about these clothes, and clearly understand the dynamic of clothes. These processes need a very good security level in order to prevent these tags from being distorted. HF RFID tag is a proper choice [6].

147.2.4.3 Contents and Capacity

All RFID tags must be given an electronic product code (EPC). In general, EPC have 3 types: 64-bit, 96-bit and 256-bit. Thus, HF and UHF RFID tags could be given EPCs to uniquely identify them. These EPCs and other information

Table 147.2 RFID tags communication band comparison [7, 8]

Communication band	Frequency	Characteristics	Remarks
Low frequency (LF)	135 kHz	Short range, low speed, high cost	Data collection
High frequency (HF)	13.56 MHz	Moderate range, moderate speed, moderate cost	Good security performance
Ultra high frequency (UHF)	900 MHz	Long range, high speed, low cost	Management on supply chain

concerning their target products can be recorded to main database. Some RFID tags offer additional read/write user memory that could be used for storage of additional information. For instance, an active RFID tag with a large capacity and a temperature censor can remember the changes of temperature for a long time.

147.3 Clothing RFID System Application

By applying RFID technology in clothing industries, goods could be tracked in real time at every stage of the supply chain. The company can exact learn which clothing is in the production process. Enterprises could check daily inventory counts to see which product backlog and which products require replenishment. This clothing RFID system application will be discussed in 4 parts in the following.

147.3.1 Clothing RFID Application in Manufacturing Plants

Firstly, in the terminal of manufacturing line, RFID tags could be attached to each product or each box depending on the circumstances. Then, each RFID tag will be given an EPC code, which must be encoded within the range given by EPCglobal. Massive data, including clothing band, category, color, SKU, storage form, marketing mode and etc., should be saved to main database and RFID tags if necessary. Depositing RFID systems in every node in the whole logistics system helps clothing enterprise collecting product information, writing logistics information and products status information of each RFID tag to database. Establishing a real-time database can also help enterprise updating logistics data in real time, tracking goods, preventing wrong products distribution. This RFID system helps clothing enterprises always grasp the situation of sale and at the same time facilitate production of next cycle.

147.3.2 Clothing RFID Application in LC

In logistics center, the RFID system can be applied to receive goods, put goods in storage, and pick goods according to order, outbound goods from storage and other segments. When clothes affixed with RFID tags enter the warehouse data collection area, the reader will automatically read tags to confirm whether the quantity, size, type of goods is consistent with the order and sent information to mainframes in LC. Mainframes will send receipt time and the extent of damage of goods to the main database, in order to complete clothing received process. Then, in the process of picking goods, Mainframes could automatically generate an application

according to the situation of inventory location of LC. After the storing process, technicians will use hand-held RFID readers to send the inventory information to mainframes through WLAN, and mainframes will change inventory information in the main database. In the outbound goods process, when pallets with clothes or packing boxes pass though data collection area at the exit of warehouse, all information in RFID tags will be read and send to system database to confirm whether error exist between these information and order of goods. Then, the data of these RFID tags in main database will be updated to the latest logistics information and products status information.

147.3.3 Clothing RFID Application in Transport Process

In transport links, RFID tags can be used to trace transport process. Transport conveyances carry with RFID system can collect codes of all RFID tags attached on clothes and send these codes to host system database. Each conveyance should possess a global positioning system (GPS). In this way, the host machine can only need to contact with conveyances to locate the position of each goods. Additionally, in some transport nodes like Goods collection stations and distribution centers, RFID systems should be deposited in order to real-time locate goods and change the logistics information and products status information in main database. RFID devices can also used to achieve automatic payment in toll stations or between distribution center and conveyance owner.

147.3.4 Clothing RFID Application in Retail Part

Applying RFID system in retail links will benefit both suppliers and retailers. Similar with the process in logistics center, RFID system can collect information of all RFID tags attached on them, and confirm data with order of goods when clothing retailers receive goods. Retailers can use RFID system to transmit data concerning point of sells (POS) to suppliers so that suppliers can keep abreast of the sales situation, and timely adjust production or procurement plans. On the other hand, suppliers use Electronic Data Interchange (EDI) system to sent orders to retailers before shipping goods so that the retailers can get ready for purchase.

At store, if clothes have not been placed back to their original places by customers, they are really tough to replace. However, by covering the entire retail store with RFID readers, those disordered clothes could be easily found and replaced by shop assistants. When customers take their favorite clothes walking through the RFID data collection area, systems in store will automatically add up price of all clothes. If a customer has a RFID payment card, RFID systems could directly deduct the cost from it without any artificial processes.

Besides, RFID technology can solve the replacement and warranty caused by the quality of clothing as well as fake and other issues. Because of the unique identification code, every RFID tags can accurately record detailed information about every aspect of the production, warehousing and sale. Thus customers can be better served and the cause of problems can be easily observed.

147.4 Conclusion

This paper systematically introduces characteristics of Chinese clothing enterprises supply chain and necessity of applying RFID technology into clothing industry. By putting forward the solutions and key technology issues concerning clothing supply chain system based on RFID technology, this paper pointed out the direction for the clothing enterprises of applying RFID technology. With more and more clothing enterprises are beginning to realize low-level information, processing capability has become the bottleneck of their development. And with a variety of information technologies and applications drive to maturity stage, the application of RFID technology in the field of clothing will be furthered as time goes by. Clothing supply chain based on RFID technology is of great significance to reduce inventory, improve work efficiency, shorten delivery time, reduce logistics costs and improve service capabilities.

References

1. Robshaw, M.J.B.: An overview of RFID tags and new cryptographic developments. Inf. Secur. Tech. Rep. **11**(2), 82–88 (2006)
2. Finkenzeller, K.: RFID Handbook, 2nd ed. Wiley, New York (2003)
3. Want, R.: An introduction to RFID technology. RFID Technology, January–March, pp. 25–33 (2006)
4. Bansal, R.: Coming soon to a Wal-Mart near you. IEEE Antennas Propag. Mag. **45**(6), 105–106 (2003)
5. Wu, N., Chang, Y.-S., Yu, H.-C.: The RFID Industry Development Strategies of Asian Countries. RFID Eurasia 2007, 1st Annual, pp. 1–6 (2007)
6. Kim, H., Kim, Y., Kim, B., Yoo, H.-J.: A wearable fabric computer by planar-fashionable circuit board technique. 2009 Body Sensor Networks, pp. 282–285 (2009)
7. Subramanian, V., Fréchet, J.M.J., Chang, P.C., Huang, D.C., Lee, J.B., Molesa, S.E., Murphy, A.R., Redinger, D.R., Volkman, S.K.: Progress toward development of all-printed RFID tags: materials, processes, and devices. In: Proceeding of the IEEE, vol. 93, no. 7, pp. 1330–1338 (2005)
8. Sen, D., Sen, P., Das, A.M.: RFID For Energy and Utility Industries, PennWell, ISBN 978-1-59370-105-5, pp. 1–48 (2009)

Author Index

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6,