

Xingui He
Ertian Hua
Yun Lin
Xiaozhu Liu *Editors*

Computer, Informatics, Cybernetics and Applications

Proceedings of the CICA 2011

Part 1

Lecture Notes in Electrical Engineering

Volume 107

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

Xingui He · Ertian Hua · Yun Lin
Xiaozhu Liu
Editors

Computer, Informatics, Cybernetics and Applications

Proceedings of the CICA 2011

Xingui He
Chinese Academy of Engineering
Beijing
People's Republic of China

Ertian Hua
Zhejiang Gongshang University
Hangzhou
People's Republic of China

Yun Lin
Zhejiang Gongshang University
Hangzhou
People's Republic of China

Xiaozhu Liu
School of Automation
Wuhan University of Technology
Luoshi Road 122
4300 Wuhan
People's Republic of China
email: randolbo@gmail.com

ISSN 1876-1100
ISBN 978-94-007-1838-8
DOI 10.1007/978-94-007-1839-5
Springer Dordrecht Heidelberg London New York

e-ISSN 1876-1119
e-ISBN 978-94-007-1839-5

Library of Congress Control Number: 2011940845

© Springer Science+Business Media B.V. 2012

No part of this work may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording or otherwise, without written permission from the Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work.

Printed on acid-free paper

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

Preface

Welcome to the proceedings of the International Conference on Computer, Informatics, Cybernetics and Applications 2011 (CICA 2011), which was held in September 13–16, 2011, in Hangzhou, China.

CICA 2011 will be a venue for leading academic and industrial researchers to exchange their views, ideas and research results on innovative technologies and sustainable solutions leading to Computer, Informatics, Cybernetics and Applications. The conference will feature keynote speakers, a panel discussion and paper presentations.

The objective of CICA 2011 is to facilitate an exchange of information on best practices for the latest research advances in the area of Computer, Informatics, Cybernetics and Applications, which mainly includes the intelligent control, or efficient management, or optimal design of access network infrastructures, home networks, terminal equipment, and etc. CICA 2011 will provide a forum for engineers and scientists in academia, industry, and government to address the most innovative research and development including technical challenges, social and economic issues, and to present and discuss their ideas, results, work in progress and experience on all aspects of Computer, Informatics, Cybernetics and Applications.

The CICA 2011 conference provided a forum for engineers and scientists in academia, industry, and government to address the most innovative research and development including technical challenges and social, legal, political, and economic issues, and to present and discuss their ideas, results, work in progress and experience on all aspects of Computer, Informatics, Cybernetics and Applications.

There was a very large number of paper submissions (721), representing 7 countries and regions, not only from Asia and the Pacific, but also from Europe, and North and South America. All submissions were reviewed by at least three Program or Technical Committee members or external reviewers. It was extremely difficult to select the presentations for the conference because there were so many excellent and interesting submissions. In order to allocate as many papers as possible and keep the high quality of the conference, we finally decided to accept 184 papers for presentations, reflecting a 25.5% acceptance rate. We believe that

all of these papers and topics not only provided novel ideas, new results, work in progress and state-of-the-art techniques in this field, but also stimulated the future research activities in the area of Computer, Informatics, Cybernetics and Applications.

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

July 2011

Xingui He
Ertian Hua
Yun Lin
Xiaozhu Liu
General and Program Chairs
CICA 2011

Contents

Part I Communication Technologies and Applications

1	Studies on Single Observer Passive Location Tracking Algorithm Based on LMS-PF	3
	Jing-bo He, Sheng-liang Hu and Zhong Liu	
2	Novel Methods for Extending Optical Code Set for Coherent OCDMA	11
	Chen Peng	
3	A Sequential Processing Method for Converted Measurement Kalman Filters Based on Orthogonal Transform	19
	Junlin Tian, Chengyu Fu and Tao Tang	
4	Improvement of an ID-Based Threshold Signcryption Scheme.	29
	Wei Yuan, Liang Hu, Xiaochun Cheng, Hongtu Li, Jianfeng Chu and Yuyu Sun	
5	Cryptanalysis of an Enhanced Event Signature Protocols for Peer-to-Peer Massively Multiplayer Online Games.	37
	Wei Yuan, Liang Hu, Hongtu Li and Jianfeng Chu	
6	Design of CAN Bus and Wireless Sensor Based Vehicle Tire Pressure Monitoring System	45
	Peng He-huan, Zheng Hong-ping and Ma Ze-yun	
7	Analysis and Test of the Exposure Synchronization of the Multi-Sensor Aerial Photogrammetric Camera System.	55
	Zhuo Shi, Li Yingcheng and Qu Lei	

8	Design and Application of Custom RS485 Communication Protocol	63
	Lei Zhou, Hu Liu and Quanyin Zhu	
9	Sensitivity of Neurons Exposed to AC Induction Electric Field	73
	Xiu Wang, Jiang Wang, Yanqiu Che, Chunxiao Han, Bin Deng and Xile Wei	
10	Spectra of Discrete Multi-Splitting Waveform Relaxation Methods to Determining Periodic Solutions of Linear Differential-Algebraic Equations	83
	Xiaolin Lin, Liming Liu, Hong Wei, Yuan Sang, Yumei Wang and Ronghui Lu	
11	A Low Power Limiting Amplifier Designed for the RSSI of a 5.8 GHz ETC Receiver	95
	Hang Yu, Yan Li, Wongchen Wei, Lai Jiang, Shengyue Lin and Zhen Ji	
12	A 3rd order Opamp-Based Tunable Low-Pass Filter Design for Data Demodulator of a 5.8 GHz ETC RF Receiver.	103
	Yan Li, Hang Yu, Shengyue Lin, Lai Jiang, Rongchen Wei and Zhen Ji	
13	Sensor Fusion Using Improvement of Resampling Algorithm Particle Filtering for Accurate Location of Mobile Robot.	111
	Xiang Gao and YaPing Fu	
 Part II Intelligence and Biometrics Technologies		
14	Sensor Placement Modes for Smartphone Based Pedestrian Dead Reckoning	123
	Shahid Ayub, Xiaowei Zhou, Soroush Honary, Alireza Bahraminasab and Bahram Honary	
15	A New Methodology of Judging the Observability of the System	133
	Xuehan Cheng and Yunxia Gao	
16	Criteria Conditions for Generalized Strictly Diagonally Dominant Matrices	141
	Yujing Liu and Li Guo	

17	Nonlinear Retarded Integral Inequalities for Discontinuous Functions and Its Applications	149
	Wu-Sheng Wang, Zizun Li and Anmin Tang	
18	Oscillatory and Asymptotic Behavior of a Second Order Nonlinear Differential Equation with Perturbation	161
	Li Gao, Quanxin Zhang and Xia Song	
19	A Expanded Binary Tree Method of Automatic River Coding and Algorithm	171
	Ji-qiu Deng, Huang-ling Gu and Xiao-qing Luo	
20	Domain-Specific Ontology Mapping Based on Common Property Collection	181
	Bo Jiang, Yunzhao Cheng and Jiale Wang	
21	Efficient Mobile Electronic Full Declaration System Based on Smart Device	191
	Xianyu Bao, Qing Lu, Bo Zhao, Weimin Zheng and Yang Wang	
22	The Predicate System Based on Schweizer–Sklar t-Norm and Its Completeness	201
	Li Qiao-Yan and Cheng Tao	
23	Providing Timely Response in Smart Car	211
	Jie Sun, Yongping Zhang, Bobo Wang and Kejia He	
24	A Study of Position Tracking Technology in Indoor Environments for Autonomous Wheelchair Driving	221
	Sung-Min Kim, Jae-Hoon Jeong, Jung-Hwan Lee and Sung-Yun Park	
25	Adaptive Regulation of a Class of Nonlinear Systems with Unknown Sinusoidal Disturbances	233
	Pengnian Chen	
26	Research on Improved Ant Colony Algorithm for TSP Problem	243
	Xiu-juan Qiu and Ting-gui Chen	

Part III Networks Systems and Web Technologies

27	Research on Internet of Things Technology	253
	Qinian Zhou, Lingling Chen, Ge Li and Zhenhao Zhang	
28	Hot Topic Detection Research of Internet Public Opinion Based on Affinity Propagation Clustering	261
	Hong Liu and Bi Wei Li	
29	TTMP: A Trust-Based Topology Management Protocol for Unstructured P2P Systems.	271
	Zhiping Liao, Song Liu, Gelan Yang and Jiancun Zhou	
30	Indoor Temperature and Humidity Monitoring System Based on WSN and Fuzzy Strategy	283
	Bo Chang and Xinrong Zhang	
31	Radioactive Target Detection Using Wireless Sensor Network	293
	Tonglin Zhang	
32	Optimization of Air Network Applied in the Express Based on Hub-and-Spoke Network: A Case Study of SF Express	303
	Peng Jianliang, Si Jiandong and Bao Fuguang	
33	Verification and Analysis for Ethernet Protocols with NuSMV	311
	Yujia Ge, Xiaofei Feng and Fangcheng Tang	
34	The Analysis and Modeling on Internet AS-Level Topology Based on K-Core Decomposition	321
	Jun Zhang, Hai Zhao and Bo Yang	
35	Based on Quasi-IP Address Model of Repeater Coordination	331
	Shaohong Yan, Lihui Zhou, Hong Wang, Yan Li, Lanqian Liu and Mengyuan Chen	
36	Fast Dynamic Mesh Moving Based on Background Grid Morphing.	339
	Tianjun Lin and Zhenqun Guan	
37	Research on Network Congestion Control System Based on Continuous Time Model.	349
	Huang Ze	

38 The Realization of OPNET and MATLAB Co-Simulation Based on HLA 357
 Yu Zhang, Zhen Yao and Xiangming Li

Part IV Data Modeling and Programming Languages

39 Design and Applications of Third Party Integrated Information Security Management System. 367
 Yuefeng Fang, Xiaoyong Wang and Fei Gao

40 A Protocol for a Message System for the Tiles of the Heptagrid, in the Hyperbolic Plane 377
 Maurice Margenstern

41 Monitoring Information System Quality: Between Reality and User Expectation at Bina Nusantara University. 387
 Anderes Gui, Cecilia Sabrina, Suryanto, Verina Kristian, Vina Margareta and Noviyanti Hadiwidjaja

42 Ordinal-Set Pair Analysis Prediction Model and Application in Liao River Basin. 395
 Jianxin Xu, Yindi Liu, Zezhong Zhang and Yanbin Li

43 Integrated RMS Layout and Flow Path Design: Modelling and a Heuristic method. 403
 Xianping Guan, Baijing Qiu and Hongbing Yang

44 Design and Implementation of Distributed Remote-Reading Water Meter Monitoring System Based on SaaS 413
 Junjie Li, Jianjiang Cui, Lilong Jiang, Zhijie Lu and Zhenyu Tan

45 Enabling the Traceability of Web Information Access in Laboratory Management via E-CARGO Model. 421
 Bin Xu

46 An Improved GVF Snake Model Using Magnetostatic Theory 431
 Bingyu Chen, Jianhui Zhao, Erqian Dong, Jun Chen, Yang Zhao and Zhiyong Yuan

47 Block Effect Reduction via Model Based Compressive Sensing. 441
 Jin Jianqiu, Zhang Zhiyong and Jiang Zhaoyi

48 The Design of Logistics Information Platform for the Yangtze River Delta 449
Peihua Fu and Xiaoli Gu

49 The Case Study for Three Kinds of Mobile Games 457
Quanyin Zhu, Suqun Cao, Rui Geng and Chuanchun Yu

50 New Methods of Specifying and Modeling Stochastic Concurrent Systems 471
Jingjing Liao, Mingzhe Wang and Fabin Guo

Part V Digital Image Processing

51 Digital Image Completion Techniques 481
Chao Huang, Huadong Hu, Chunxiao Liu and Caiyan Xie

52 Application of 2-D Wavelet Transform in Image Compression Based on Matlab. 491
Shi-Gang Hu, Xiao-Feng Wu and Zai-Feng Xi

53 Fingerprint Image Preprocessing Method Based on the Continuous Spectrum Analysis 501
Xiaosi Zhan and Zhaocai Sun

54 A Zero-Watermarking Algorithm for Digital Map Based on DWT Domain 513
Yun Ling, Chuan-Feng Lin and Zhi-yong Zhang

55 A Fault-Tolerance Shortest Routing Algorithm with PDPE on (n, k)-Star Graph for NoC 523
Jianguo Xing and Ming Feng

56 A Robust Zero-Watermarking Algorithm for 2D Vector Digital Maps. 533
W. Xun, H. Ding-jun and Z. Zhi-yong

57 An Efficient Non-Local Means for Image Denoising 543
Gang Yang, Qiegen Liu and Jianhua Luo

58 An Efficient Augmented Lagrangian Method for Impulse Noise Removal via Learned Dictionary 551
Qiegen Liu, Gang Yang, Jianhua Luo and Shanshan Wang

59 Motion Object Segmentation Using Regions Classification and Energy Model 561
 Xiaokun Zhang and Xuying Zhao

60 A Digital Image Scrambling Method Based on Hopfield Neural Network 569
 Li Tu and Xuehua Huang

61 Lattice Boltzmann Anisotropic Diffusion Model Based Image Segmentation 577
 Qian Wu, Yu Chen and Gang Teng

Part VI Optimization and Scheduling

62 Research on New Distributed Solution Method of Complex System Based on MAS 589
 Weijin Jiang, F. Guo Fan and Luo Zhong

63 An Order-Searching Genetic Algorithm for Multi-Dimensional Assignment Problem 597
 Yazhen Jiang, Xinming Zhang and Li Zhou

64 The Heuristic Algorithm of Stacking Layer for the Three-Dimensional Packing of Fixed-Size Cargoes 605
 Wang-sheng Liu, Hua-yi Yin and Mao-qing Li

65 Research on a Novel Hybrid Optimization Algorithm Based on Agent and Particle Swarm 613
 Weijin Jiang, Xingjun Jiang and Lina Yao

66 A Novel BCC Algorithm for Function Optimization. 623
 Jia-Ze Sun, Guo-Hua Geng, Ming-Quan Zhou and Wang Shu-Yan

67 Time-Space Hybrid Markov Model. 633
 Xin-yao Zou

68 Research on Turning Parameters Optimization Based on Genetic Algorithm 641
 Youde Wu and Jinchun Feng

69 A Power of a Meromorphic Function Sharing 1 IM with Its Derivative 649
Wang Gang and Gong Dianxuan

70 The Application of Particle Swarm Optimization in Stock Prediction and Analysis 659
Guorong Xiao

71 Research on a Novel Distributed Multi Agent System Plan Method 665
Weijin Jiang, Qing Jiang, Hong Li and Yan Su

72 A Method of Object Detection Based on Improved Gaussian Mixture Model 673
Jiajia Sun, Yanan Lian and Yumin Tian

73 A Novel Dynamic Resource Allocation Model Based on MAS Coordination 681
Weijin Jiang, Luo Zhong, Yuhui Xu and Yan Su

Part VII Education and Informatics

74 Research on Comprehensive Evaluation of Classroom Teaching Quality Based on Multi-Element Connection Number 691
Liao Xinfei

75 Program Design of DEA Based on Windows System 699
Ma Zhanxin, Ma Shengyun and Ma Zhanying

76 An Incremental and Autonomous Visual Learning Algorithm Based on Internally Motivated Q Learning 709
Qu Xinyu, Yao Minghai and Gu Qinlong

77 Design and Implement of Information System for Water Management Based on SaaS 721
Lilong Jiang, Junjie Chen, Zhijie Lu and Junjie Li

78 Pulse-Based Analysis on Teaching Quality Evaluation 729
Aijun Zhang, Yujia Ge and Biwei Li

79 French Learning Online Platform: French Livehand 739
Mei Ying Lu, Chun Lai Chai and Yu Shen

80 Application of Text Data Mining to Education in Long-Distance. 745
 Jianjie Song and Hean Liu

Part VIII Fuzzy System and Control

81 Adaptive Disturbance Rejection Control of Linear Time Varying System 755
 Dangjun Zhao, Zheng Wang, Yongji Wang and Weibing Hu

82 Nonlinear Predictive Functional Control Based on Support Vector Machine 763
 Zhang Xin-fang, Wang Cheng-li, Su Xian-hua and Wang Qiu-jin

83 Analysis and Circuit Simulation of a New Four-Dimensional Lorenz Time-Delay Chaotic System. 773
 Cui Zhiyong, Fan Zhongkui and Cao Panjing

84 The Design of Adaptive PI Speed Controller for Permanent Magnet Synchronous Motor Servo System. 783
 Yulin Gong and Yongyin Qu

85 Application of BP Neural Network in Faults Diagnosis of Process Valve Actuator. 793
 Fukai Deng, Erqing Zhang, Qunli Shang, Shan-en Yu and Yifang He

86 Optimal Detection of Distributed Target with Fluctuating Scatterers 803
 Tao Jian

87 Output Feedback Control for an Active Heave Compensation System 811
 Jia-Wang Li, Tong Ge and Xu-Yang Wang

88 Models for Multiple Attribute Decision Making with Intuitionistic Trapezoidal Information 821
 Liang Xiao-yi

89 Modeling the Risk Factors in Ergonomic Processes Using Fuzzy Logic 827
 Prafulla Kumar Manoharan, Sanjay Jha and Bijay Kumar Singh

90 Modified Projective Synchronization of Uncertain Chaotic Systems 833
 R. Z. Luo, Y. L. Wang and S. C. Deng

91 Multi-Agent Based Architecture Supporting Collaborative Product Lightweight Design 839
 Zheng Liu, Fuyuan Xu, Kai Pan, Xinjian Gu and Yongwei Zhang

92 Fuzzy Controller Design with Fault Diagnosis System Condition On-line Monitor Using Neural Network. 849
 Xiaochun Lou

Part IX Forensics, Recognition Technologies and Applications

93 Fingerprint Orientation Template Matching Based on Mutual Information 863
 Xuying Zhao, Xiaokun Zhang, Geng Zhao, Rong Qian, Xiaodong Li and Kejun Zhang

94 Human Action Recognition Algorithm Based on Minimum Spanning Tree 871
 Yi Ouyang and Jianguo Xing

95 Analysis of Selvi et al.’s Identity-Based Threshold Signcryption Scheme. 881
 Wei Yuan, Liang Hu, Hongtu Li and Jianfeng Chu

96 Analysis of an Authenticated 3-Round Identity-Based Group Key Agreement Protocol 889
 Wei Yuan, Liang Hu, Hongtu Li and Jianfeng Chu

97 Research on ECC Digital Certificate in ATN. 897
 Wu Zhijun, Lang Jihai and Huang Jun

98 A Robust Method Based on Static Hand Gesture Recognition for Human-Computer Interaction Under Complex Background. 907
 Xingbao Meng, Jing Lin and Yingchun Ding

99 Research on Privacy Preserving Based on K-Anonymity 915
 Xiao-ling Zhu and Ting-gui Chen

100 Existential Forgery Attack Against One Strong Signature Scheme 925
Zhengjun Cao and Qian Sha

101 Improved Min-Sum Decoding Algorithm for Moderate Length Low Density Parity Check Codes. 935
Waheed Ullah, Jiangtao and Yang FengFan

102 Cryptanalysis of an Authentication Protocol for Session Initiation Protocol. 945
Qi Xie, Mengjie Bao and Haijun Lu

103 A Cognitive Model in Biomimetic Pattern Recognition and Its Applications 953
Xiao Xiao, Xianbao Wang, Sunyuan Shen and Shoujue Wang

104 One Way to Enhance the Security of Mobile Payment System Based on Smart TF Card 961
Xiaoning Jiang, Xiping Wang and Yihao Huang

105 Localization on Discrete Grid Graphs 971
Anna Gorbenko, Vladimir Popov and Andrey Sheka

Part X Electronic Applications

106 The Research of Chinese Q&A System Based on Similarity Algorithm 981
Liu Zhen, Xiao Wenxian, Wan Wenlong and Yulan Li

107 A Model-Driven Method for Service-Oriented Modeling and Design Based on Domain Ontology 991
Ying Zhang, Xiaoming Liu, Zhixue Wang and Li Chen

108 Design of Three-Dimensional Garage Monitoring System Based on WinCC 999
Wenzhe Qi, Yanmei Yang, Lingping Shi,
Yafei Liu and Jianjun Meng

109 A Hybrid Modeling Method for Service-Oriented C4ISR Requirements Analysis 1007
Ying ZHANG, Xiaoming Liu, Zhixue Wang and Li Chen

110 Study on Normalized Operation of Internet Drugs Market. 1017
Feng Chen, Ying Lu, Bin Dun and Ming Lu

**111 Study on Accounting Shenanigan in Universities Using
Improved Game Theory** 1025
Jinfang Zhao

112 Improved Iterative Decoding Algorithm for Turbo Codes 1033
Zhenchuan Zhang and Jiang Nana

**113 Financial Evaluation of the Listed Companies Based
on Statistical Analysis Methods** 1043
Guofu zhang

114 Research of the Timer Granularity Based on Linux. 1053
Geng Qingtian, Jiang Nan, Zhao Hongwei and Liu Junling

**115 A New Software to Realize the Optimization of City
Sound-Planning** 1061
Ji Qing

Part XI Graphics and Visualizing

**116 Automatic Classification and Recognition of Particles
in Urinary Sediment Images** 1071
Yan Zhou and Houkui Zhou

**117 Ultrasound Strain and Strain Rate Imaging of the Early Sage
of Carotid Artery with Type 2 Diabetes Mellitus** 1079
Cun Liu, Yanling Zheng, Yuanliu He, Hongxia Xu, Juan Su,
Lili Zhang, Xiaohong Zhou and Changchun Liu

**118 Rate Control for Multi-View Video Coding Based on Statistical
Analysis and Frame Complexity Estimation.** 1089
Qiaoyan Zheng, Mei Yu, Gangyi Jiang, Feng Shao
and Zongju Peng

**119 Time-Consistent Preprocessing of Depth Maps for Depth
Coding in 3DTV** 1099
Fengfei Sun, Mei Yu, Gangyi Jiang, Feng Shao,
Zongju Peng and Fucui Li

120 A Mixture of Gaussian-Based Method for Detecting Foreground Object in Video Surveillance 1109
Xun Wang, Jie Sun and Hao-Yu Peng

121 Video Deformation Based on ASM 1119
Xiaofei Feng and Yujia Ge

122 An Enhanced Hybrid Image Watermarking Algorithm Using Chaotically Scrambled Technology 1129
Niansheng Cheng

123 A Digital Watermarking Technology Based on Wavelet Decomposition 1137
Song Yijun and Yang Gelan

124 An Improved Rate Control Algorithm of H.264/AVC Based on Human Visual System 1145
Meifeng Liu and Ling Lu

125 Contour Lines Extraction from Color Scanned Topographical Maps with Improved Snake Algorithm 1153
Huali Zheng and ZhouWei Guo

126 The Aberration Characteristics of Wave-Front Coding System for Extending the Depth of Field 1161
Zhang Rong-fu, Wang Liang-liang and Zhou Jinbo

Part XII Green Computing

127 Two Improved Nearest Neighbor Search Algorithms for SPH and Their Parallelization 1171
Zheng Dequn, Wu Pin, Shang Weilie, Cao Xiaopeng, Zhang Xia and Chen Wei

128 Comprehensive Evaluation of TPL Using Genetic Projection Pursuit Model with AHP in Supply Chain 1181
Fuguang Bao, Jiandong Si, Lei Zhang and Ertian Hua

129 A Modified PID Tunning Fitness Function Based on Evolutionary Algorithm 1191
Xiao Long Li and Dong Liu

130	Analysis of Convergence for Free Search Algorithm in Solving Complex Function Optimization Problems.	1201
	Lu Li, Zihou Zhang and Xingyu Wang	
131	Design of FH Sequences with Given Minimum Gap Based on Logistic Map 1.	1209
	Wan Youhong and Li Jungang	
132	Practical Criteria for Generalized Strictly Diagonally Dominant Matrices.	1221
	Yujing Liu, Jie Xu and Li Guo	
133	Research on New Rural Information Service Model of Agricultural Industrial Chain	1231
	Hai-han Yang, Xing-xia Shuai, Chun-hua Ju and Tie-zhu Zhang	
134	Reliability Evaluation Algorithm for Distribution Power System.	1241
	Jianghua Li and Juan Hu	
135	State-of-the-Art Line Drawing Techniques.	1249
	Lei Li, Yongqiang Zhou, Chunxiao Liu, Yingying Xu and Jie Fu	
136	The Cooperation and Competition Mechanism of Supply Chain Based on Evolutionary Game Theory	1259
	Peng Yang and Fu Pei-Hua	
137	A Method of Function Modeling Based on Extenics.	1267
	Jing Zhang, Ping Jiang, Jiaqi Wang, Runhua Tan and Yang Yang	
138	Application of MCR-ALS Computational Method for the Analysis of Interactions Between Copper Ion and Bovine Serum Albumin	1281
	Zhu Xin-feng, Wang Jian-dong and Li Bin	
139	Self-Learning Algorithm for Visual Recognition and Object Categorization for Autonomous Mobile Robots . . .	1289
	Anna Gorbenko and Vladimir Popov	
140	Time-Complexity of the Algorithm for Physical Ability Test	1297
	Feng Qiming	

Part XIII Data Management and Database System

141 A Distributed Keyword Search Algorithm in XML Databases Using MapReduce 1307
Yun Ling and Guangjian Xu

142 A Combined Data Mining Method and Its Application in Water Quality Trends Analysis 1317
Ji-Min Wang, Ding-Sheng Wan, Yue-Long Zhu and Shi-Jin Li

143 Model of Enterprise Innovation Based on Data Warehousing 1327
Jianwei Guo, Bingru Yang, Guangyuan Li and Nan Ma

144 Classifying Imbalanced Dataset Using Local Classifier Fusion 1337
Tong Liu, Weijian Ni and Yongquan Liang

145 Access Frequency Based Energy Efficiency Optimization in Data Centers 1347
Jie Wang, Yangang Cai, Chunxiao Liu, Jiabin Luo and Xinyi Ding

146 Application of Data Mining Technology in Jewelry Design 1357
Cen Qin

147 Logic Algebra Method for Solving Theoretic Problems of Relational Database 1367
Zhang YiShun and Ju ChunHua

148 Design History Knowledge Management System Based on Product Development Process Management and Its Implementation 1375
Peisi Zhong, Jiandong Song, Mei Liu, Xueyi Li and Shuhui Ding

149 The Application of Series Importance Points (SIP) Based Partition Method on Hydrological Data Processing 1385
Haixiong Chen

150 A Novel PSO k-Modes Algorithm for Clustering Categorical Data 1395
Lu Mei and Zhao Xiang-Jun

Part XIV E-Commerce and E-Government

151 An Analysis of Influential Factors of Human Resource Allocation in Local Taxation System and the Modeling Approaches 1405
Peiling Chen and Tian Yi

152 Classification Learning System Based on Multi-Objective GA and Frigid Weather Forecast 1417
Zhang Hongwei, Xu Jingxun and Zou Shurong

153 Empirical Analysis on Choice of Payment Terms in Foreign Capital Acquiring State-Owned Enterprise 1427
Ling Liu and Jingru Wu

154 An Integrated Application of Tourism Planning Based on Virtual Reality Technology and Indicator Assessment 1435
Yan Zhang and Xun Zhang

155 A Context Information Management Model for Tour Mobile E-Commerce 1443
Szemin Wong and Dongsheng Liu

156 Quantitative Quality Evaluation and Improvement in Incremental Financial Software Development 1453
Bin Xu, Meng Chen, Cun Liu, Juefeng Li, Qiwei Zhu and Aleksander J. Kavs

157 Study on Internet Drug Market Access Management 1463
Ying Lu, Bin Dun, Feng Chen and Ming Lu

158 Study on the Real-Name System Technology for the Ontology of Internet Medicine Market 1471
Bin Dun, Ying Lu, Feng Chen and Ming Lu

159 The Combined Stock Price Prediction Model based on BP Neural Network and Grey Theory. 1479
Yulian Fei, Junjun Xiao, Ying Chen and Weiwei Cao

Part XV Social and Economical Systems

160 Multivariate Curve Resolution with Elastic Net Regularization 1489
 Zhu Xin-feng, Wang Jian-dong and Li Bin

161 Research about Exoskeleton’s Reference Trajectory Generation Based on RBF Neural Network 1501
 Shaowu Dai, Shuangming Li and Zhiyong Yang

162 The Application of Digital Archives Classification with Progressive M-SVM to Wisdom School Building 1513
 Xiaobo Tao

163 Economic Growth Differential Model and Short-Term Economic Growth Momentum. 1521
 Fan Decheng and Liu Kan

164 A Study of Leading Industries Selection in Comprehensive Cities Based on Factor Analysis 1531
 Faming Zhang and Zhaoqing Ye

165 A Programme-Oriented Algorithm to Compute $Ord(f(x))$ 1539
 Lihua Liu and Zhengjun Cao

166 Finding the Academic Collaboration Chance in Open Research Community. 1549
 Bin Xu, Huanyu Zhou, Weigang Chen and Yujia Ge

167 Analysis of Competition and Cooperation of Ningbo-Zhoushan Port and Shanghai Port 1559
 Peihua Fu and Yangfei Chen

168 The Study About Long Memory and Volatility Persistence in China Stock Market Based on Fractal Theory and GARCH Model 1569
 Liu Cheng, Liu Jiankang and He Guozhu

169 Research on Home-Textile Enterprise-Oriented Comprehensive Management Integration Platform and Its Application 1575
 Penghui Zhan, Zailin Lu, Renwang Li, Zhigang Bao, Qian Yin and Yongxian Chen

170 Location-Aware Elderly Personal Safety Checking in Smart Home. 1583
 Bin Xu, Zepeng Hu, Peiyou Han and Yujia Ge

171 Research on Maximum Benefit of Tourist Enterprises Based on the Influence of Scenic Spot Ticket Discount Amount 1591
 Li Jian, Han Na and Shu Bo

Part XVI Web Service and Data Mining

172 A Fuzzy Multi-Criteria Group Decision Making Approach for Hotel Location Evaluation and Selection 1599
 Santoso Wibowo and Hepu Deng

173 Using Online Self-Adaptive Clustering to Group Web Documents 1609
 Noel Catterall

174 An Order-Based Taxonomy for Text Similarity 1617
 Yi Feng

175 Distributed Intrusion Detection System Using Autonomous Agents Based on DCOM 1625
 Yujia GE, Xiaofei Feng and Aijun Zhang

176 A Method for Uncertain Linguistic Multiple Attribute Decision Making and Its Application 1633
 Xiao-yi Liang

177 Web2.0 Environment of Personal Knowledge Management Applications 1639
 Jie Fang and Liqun Gong

178 Semantic Web and Its Applications 1647
 WenYing Guo

179 A Semantics-Based Web Service Matching Framework and Approach. 1655
 Bo Jiang, Tao He and Jiale Wang

180 Using Integrated Technology to Achieve Three-Dimensional WebGIS System in Park Planning 1665
Hu Lemin, Wu Qianhong, Yang Li and Cheng Chuanzhou

181 Hierarchical Base-k Chord Based on Semantic Networks 1673
Huayun Yan and Deqian Xue

182 Enterprises Application Integration Framework Based on Web Services and Its Interfaces 1685
Zhangbing Li, Wujiang Che and Jianxun Liu

183 The Automatic Classification 3D Point Clouds Based Associative Markov Network Using Context Information 1695
Gang Wang, Ming Li, TingTing Zhou and Longgang Chen

Author Index 1703

Organization

CICA 2011 was organized by Zhejiang Gongshang University, University of Washington, Lancaster University, Zhejiang University, Nanyang Technological University, Wuhan University of Technology, and sponsored by the National Science Foundation of China, Shanghai Jiao Tong University. It was held in cooperation with Lecture Notes in Electrical Engineering (LNEE) of Springer.

Executive Committee

General Chairs	Xingui He, Academician of Chinese Academy of Engineering, China Ertian Hua, Zhejiang Gongshang University, China Bahram Honary, Lancaster University, UK Juergen Bruess, AutoTxt, Germany
Program Chairs	Yun Lin, Zhejiang Gongshang University, China Ming Fan, University of Washington, USA Yuhang Yang, Shanghai Jiao Tong University, China
Local Arrangement Chairs	Wang Xun, Zhejiang Gongshang University, China Chunlai Chai, Zhejiang Gongshang University, China
Steering Committee	Qun Lin, Chinese Academy of Sciences, China MaodeMa, Nanyang Technological University, Singapore Nadia Nedjah, State University of Rio de Janeiro, Brazil Lorna Uden, Staffordshire University, UK Yiming Chen, Yanshan University, China

Aimin Yang, Hebei united University, China
 Chunying Zhang, Hebei United University, China
 Dechang Chen, Uniformed Services University
 of the Health Sciences, USA
 Mei-Ching Chen, Tatung University, Taiwan
 Rong-Chang Chen, National Taichung Institute
 of Technology, Taiwan
 Chi-Cheng Cheng, National Sun Yat-Sen
 University, Taiwan
 Ming Fan, Foster School of Business of
 University of Washington, USA
 Juergen Bruess, AutoTXT, Germany
 Bahram Honary, Lancaster University, UK
 Michael Darnell Warwick University, UK
 Plamen Angelov Lancaster University, UK
 Farideh Honary Lancaster University, UK
 T.R. Melia, Cisco Systems, Switzerland

Program/Technical Committee

Yuan Lin	Norwegian University of Science and Technology, Norwegian
Yajun Li	Shanghai Jiao Tong University, China
Yanliang Jin	Shanghai University, China
Mingyi Gao	National Institute of AIST, Japan
Yajun Guo	Huazhong Normal University, China
Haibing Yin	Peking University, China
Jianxin Chen	University of Vigo, Spain
Miche Rossi	University of Padova, Italy
Ven Prasad	Delft University of Technology, Netherlands
Mina Gui	Texas State University, USA
Nils Asc	University of Bonn, Germany
Ragip Kur	Nokia Research, USA
On Altintas	Toyota InfoTechnology Center, Japan
Suresh Subra	George Washington University, USA
Xiyin Wang	Hebei Polytechnic University, China
Dianxuan Gong	Hebei Polytechnic University, China
Chunxiao Yu	Yanshan University, China
Yanbin Sun	Beijing University of Posts and Telecommunications, China

Guofu Gui	CMC Corporation, China
Haiyong Bao	NTT Co., Ltd., Japan
Xiwen Hu	Wuhan University of Technology, China
Mengze Liao	Cisco China R&D Center, China
Yangwen Zou	Apple China Co., Ltd., China
Liang Zhou	ENSTA-ParisTech, France
Zhanguo Wei	Beijing Forestry University, China
Hao Chen	Hu'nan University, China
Lilei Wang	Beijing University of Posts and Telecommunications, China
Xilong Qu	Hunan Institute of Engineering, China
Duolin Liu	ShenYang Ligong University, China
Xiaozhu Liu	Wuhan University, China
Yanbing Sun	Beijing University of Posts and Telecommunications, China
Yiming Chen	Yanshan University, China
Hui Wang	University of Evry in France, France
Shuang Cong	University of Science and technology of China, China
Haining Wang	College of William and Marry, USA
Zengqiang Chen	Nankai University, China
Dumisa Wellington Ngwenya	Illinois State University, USA
Hu Changhua	Xi'an Research Insti. of Hi-Tech, China
Juntao Fei	Hohai University, China
Zhao-Hui Jiang	Hiroshima Institute of Technology, Japan
Michael Watts	Lincoln University, New Zealand
Tai-hon Kim	Defense Security Command, Korea
Muhammad Khan	Southwest Jiaotong University, China
Seong Kong	The University of Tennessee USA
Worap Kreesuradej	King Mongkuts Institute of Technology Ladkrabang, Thailand
Uwe Kuger	Queen's University of Belfast, UK
Xiao Li	CINVESTAV-IPN, Mexico
Stefa Lindstaedt	Division Manager Knowledge Management, Austria
Paolo Li	Polytechnic of Bari, Italy
Tashi Kuremoto	Yamaguchi University, Japan
Chun Lee	Howon University, Korea
Zheng Liu	Nagasaki Institute of Applied Science, Japan
Michiharu Kurume	National College of Technology, Japan
Sean McLoo	National University of Ireland, Ireland

R. McMenemy	Queens University Belfast, UK
Xiang Mei	The University of Leeds, UK
Cheol Moon	Gwangju University, Korea
Veli Mumcu	Technical University of Yildiz, Turkey
Nin Pang	Auckland University of Technology, New Zealand
Jian-Xin Peng	Queens University of Belfast, UK
Lui Piroddi	Technical University of Milan, Italy
Girij Prasad	University of Ulster, UK
Cent Leung	Victoria University of Technology, Australia
Jams Li	University of Birmingham, UK
Liang Li	University of Sheffield, UK
Hai Qi	University of Tennessee, USA
Wi Richert	University of Paderborn, Germany
Meh shafiei	Dalhousie University, Canada
Sa Sharma	University of Plymouth, UK
Dong Yue	Huazhong University of Science and Technology, China
YongSheng Ding	Donghua University, China
Yuezhi Zhou	Tsinghua University, China
Yongning Tang	Illinois State University, USA
Jun Cai	University of Manitoba, Canada
Sunil Maharaj Sentech	University of Pretoria, South Africa
Mei Yu	Simula Research Laboratory, Norway
Gui-Rong Xue	Shanghai Jiao Tong University, China
Zhichun Li	Northwestern University, China
Lisong Xu	University of Nebraska-Lincoln, USA
Wang Bin	Chinese Academy of Sciences, China
Yan Zhang	Simula Research Laboratory and University of Oslo, Norway
Ruichun Tang	Ocean University of China, China
Wenbin Jiang	Huazhong University of Science and Technology, China
Xingang Zhang	Nanyang Normal University, China
Qishi Wu	University of Memphis, USA
Jalel Ben-Othman	University of Versailles, France

Part I
Communication Technologies and
Applications

Chapter 1

Studies on Single Observer Passive Location Tracking Algorithm Based on LMS-PF

Jing-bo He, Sheng-liang Hu and Zhong Liu

Abstract As the emitter's velocity is given, it could be located by single observer. According to the tracking convergence fast specialty of the linear minimum mean-square error filter and the tracking accuracy specialty of the particle filter, a new passive location algorithm based on a LMS-PF is presented. It is compared with linear minimum mean-square error filtering and extended kalman filtering in passive location. It is proved that the location error by the algorithm is less than by other algorithms.

Keywords Particle filtering · Linear minimum mean-square error filtering · Extended Kalman filtering · Passive location

1.1 Introduction

As the emitter's velocity is given, it could be located by single observer. The algorithm of linear minimum mean-square error filter is often applied in single observer passive location as extended kalman filter. It has the specialty of tracking convergence fast but its location precision is low in the low precision of measuring azimuth. The extended kalman filter has the specialty of tracking location high precision but it is sensitivity for the first state estimation and the noise [1–4]. The particle filter is a sort of effective nonlinear filtering algorithm [5–7] and has the

J. He (✉) · S. Hu · Z. Liu
Electronics Engineering College, Naval University of Engineering,
Wuhan 430033, China
e-mail: jbh_1979@sina.com

Table 1.1 Comparison of LMS, EKF and PF: the single observer passive location tracking capability

Algorithm	Tracking convergence	Tracking precision
LMS	Fast	Low
EKF	Middle	Middle
PF	Slow	High

specialty of tracking high precision. And it needs not linearization process of state equation. The three algorithms are together compared, as illustrated in Table 1.1.

According to the tracking convergence fast specialty of the linear minimum mean-square error filter and the tracking accuracy specialty of the particle filter, a new passive location algorithm based on a LMS-PF is presented.

1.2 The Particle Filter Brief Introduction

Central for all navigation and tracking application is the motion model to which various kind of model based filters can be applied. Models that are linear in the state dynamics and nonlinear in the measurements are considered:

$$x_{t+1} = Ax_t + B_u u_t + B_f f_t \quad (1.1)$$

$$y_t = h(x_t) + e_t \quad (1.2)$$

Here

x_t state vector;

u_t measured inputs;

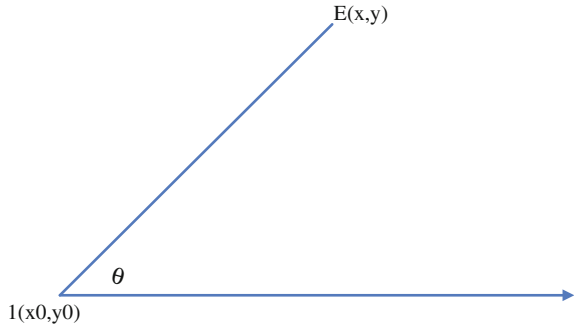
f_t unmeasured forces or faults;

y_t measurements;

e_t measurement error. A numerical approximation to particle filter is given in the following algorithm:

- (1) *Initialization*. Generate $x_0^i \sim p_{x_0}$, $i = 1, 2, \dots, N$. Each sample of the state vector is referred to as a particle.
- (2) *Measurement update*: update the weights by the likelihood (more generally, any importance function) $w_t^i = w_{t-1}^i p(y_t | x_t^i) = w_{t-1}^i p_{e_t}(y_t - h(x_t^i))$, $i = 1, 2, \dots, N$, and normalized. As an approximation, take $\hat{x}_t \approx \sum_{i=1}^N w_t^i x_t^i$.
- (3) *Resampling*. Many models are used, such as sampling importance resampling, residual resampling, MCMC, Rao-Blackwellised.
- (4) *Prediction*. Take a $f_t^i \sim p_{f_t}$, and simulate $x_{t+1}^i = Ax_t^i + B_u u_t + B_f f_t^i$.
- (5) Let $t = t + 1$, and iteration to item 2.

Fig. 1.1 Position relation between single observer and emitter



1.3 The Algorithm Based on LMS-PF for Signal Observer Passive Location Tracking

First, we make the model on the algorithm based on LMS for single observer passive location tracking; Second, the algorithm based on PF is given; Last, it is given that the algorithm based on LMS-PF. It is explained that we discuss the single observer passive location tracking problem in the two dimensions and three dimensions' can be discussed the same as these.

1.3.1 The Algorithm Based on LMS for Single Observer Passive Location Tracking

In Fig. 1.1, the position relation between single observer and emitter is indicated. According to the geometry

$$\operatorname{tg}\theta = \frac{y_t - y_0}{x_t - x_0} \tag{1.3}$$

This yields

$$y_t \cos \theta - x_t \sin \theta = y_0 \cos \theta - x_0 \sin \theta \tag{1.4}$$

And according to the emitter velocity (v_x, v_y) , we can get the prediction equation:

$$x_t = x_{t-1} + v_x \cdot T \tag{1.5}$$

$$y_t = y_{t-1} + v_y \cdot T \tag{1.6}$$

Here

- x_t, y_t state vector at time t ;
- x_{t-1}, y_{t-1} state vector at time $t - 1$;
- T sampling interval.

Depending on (1.4)–(1.6), we could get

$$r = HX \quad (1.7)$$

that $r = \begin{bmatrix} y_0 \cos \theta - x_0 \sin \theta \\ x_{t-1} + v_x \cdot T \\ y_{t-1} + v_y \cdot T \end{bmatrix}$, $H = \begin{bmatrix} -\sin \theta & \cos \theta \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$, $X = \begin{bmatrix} x_t \\ y_t \end{bmatrix}$. So the optimization state vector at time t are

$$\hat{X} = \begin{bmatrix} \hat{x}_t \\ \hat{y}_t \end{bmatrix} = (H^T H)^{-1} H^T r \quad (1.8)$$

1.3.2 The Algorithm Based on PF for Single Observer Passive Location Tracking

The particle filter is used in the single observer passive location tracking like II. But II estimates one state x and this need estimate two states (x, y) . So we improve the particle filter in: (1) Weights depend on the fault probability density function and need not iterate; (2) In the prediction, the particles are not used but optimization state vector; (3) For the (2), not resampling. A numerical approximation to particle filter in the single observer passive location tracking is given in the following algorithm:

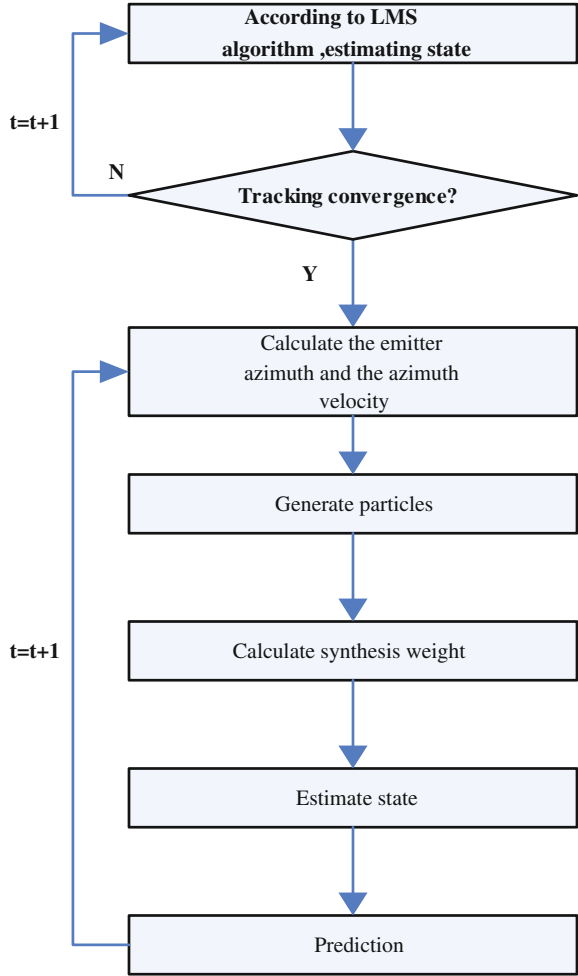
- (1) Depending on (1.1) and (1.2), the system state equation of single observer passive location tracking is

$$\begin{cases} x_t = x_{t-1} + v_x \cdot T + e_x \\ y_t = y_{t-1} + v_y \cdot T + e_y \end{cases} \quad (1.9)$$

$$\begin{cases} \theta = \arctg \frac{y_t - y_0}{x_t - x_0} + e_\theta \\ \dot{\theta} = \frac{v_x \cos \theta - v_y \sin \theta}{\sqrt{(x-x_0)^2 + (y-y_0)^2}} + e_{\dot{\theta}} \end{cases} \quad (1.10)$$

- (2) Depending on (10), get the emitter azimuth θ and azimuth velocity $\dot{\theta}$ currently.
(3) *Initialization*. Generate (x_t^i, y_t^i) depending on (9). Each sample of the state vector is referred to as a particle. And calculate azimuth θ_i and azimuth velocity $\dot{\theta}_i$ of each particle, $i = 1, 2, \dots, N$.
(4) *Weights Calculate*. Update the weights by the likelihood (more generally, any importance function) $w_\theta^i = p_{e_\theta}(\theta - \theta_i)$, $w_{\dot{\theta}}^i = p_{e_{\dot{\theta}}}(\dot{\theta} - \dot{\theta}_i)$, and normalized to $\bar{w}_\theta^i, \bar{w}_{\dot{\theta}}^i$, $i = 1, 2, \dots, N$.
(5) *Synthesis Weight*. $w_i = \bar{w}_\theta^i \cdot \bar{w}_{\dot{\theta}}^i$, and normalized to \bar{w}_i , $i = 1, 2, \dots, N$.
(6) *Measurement Update*. As an approximation, take $\hat{x}_t^\Lambda \approx \sum_{i=1}^N \bar{w}_i x_t^i$, $\hat{y}_t^\Lambda \approx \sum_{i=1}^N \bar{w}_i y_t^i$.

Fig. 1.2 The LMS-PF algorithm flow chart



(7) Prediction: Simulate
$$\begin{cases} x_{t+1}^i = x_t^i + v_x \cdot T + e_x \\ y_{t+1}^i = y_t^i + v_y \cdot T + e_y \end{cases}, \quad i = 1, 2, \dots, N.$$

(8) Let $t = t + 1$, and iteration to item 2).

1.3.3 The Algorithm Based on LMS-PF for Single Observer Passive Location Tracking

According to the tracking convergence fast specialty of the linear minimum mean-square error filter and the tracking accuracy specialty of the particle filter, a new passive location algorithm based on a LMS-PF is presented. The linear minimum

Fig. 1.3 The target tracking: trajectory for the LMS, EKF and LMS-PF

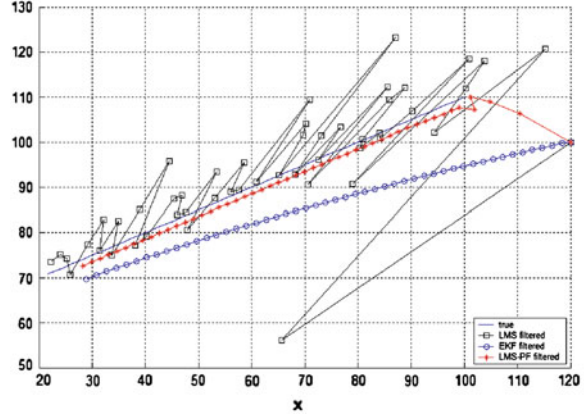


Table 1.2 Comparison of LMS, EKF AND PF: the single observer passive location tracking capability

Algorithm	Cost time (s)	Mean tracking error
LMS	0.0150	12.4296
EKF	0.0160	14.6275
PF	0.2350	7.4170

mean-square error filter A is used at the beginning of the single observer passive location tracking and at the tracking convergence we use the particle filter B. In Fig. 1.2 the algorithm flow chart is

1.4 The Signal Observer Passive Location Tracking Experiments

According to the algorithm based on LMS-PF, we make experiments comparing with LMS and EKF in the conditions: single observer position (10, 10), the emitter moving from (100, 110) to (20, 70) in mean-velocity, tracking sampling dots $L = 50$, azimuth-measurement error 0.30, the first tracking position dot (120, 100). The tracking trajectory is given in Fig. 1.3.

Defining the mean tracking error $\sigma = \frac{\sum_{i=1}^L \sqrt{(\hat{x}-x)^2 + (\hat{y}-y)^2}}{L}$, the tracking capability of algorithms is compared in Table 1.2.

So we can make the conclusion that the LMS-PF is higher than LMS and EKF in tracking precision but its cost time is much.

1.5 Conclusions and Discussion

We have discussed the algorithm based on LMS-PF for single observer passive location tracking and made comparison with the linear minimum mean-square error filter and extended kalman filter by making experiments. We have made the conclusion that the LMS-PF is better than others in tracking precision. But it depends on the azimuth error. Next, we would study high precision location algorithm based on multi-parameters, such as azimuth, frequency, and time of arrival.

References

1. Bar-Shalom Y (1993) Estimation and tracking, principles, techniques, and software. Artech House, Boston
2. Bell M, Cathey W (1993) The iterated Kalman filter update as a Gauss-Newton method. *IEEE Trans Autom Control* 38(2):294–297
3. Song TL, Speyer J (1985) A stochastic analysis of a modified gain extended Kalman filter with application to estimation with bearing only measurements. *IEEE Trans Autom Control* 30(10):940–949
4. Galkowski P, Islam M (1991) An alternative derivation of modified gain function of Song and Speyer. *IEEE Trans Autom Control* 36(11):1322–1326
5. Doucet A, Godsill SJ, Andrieu C (2000) On sequential simulation-based methods for Bayesian filtering. *Statist Comput* 10(3):197–208
6. Doucet A, de Freitas N, Gordon N (2001) Sequential Monte Carlo methods in practice. Springer-Verlag, New York
7. Gilks W, Richardson S, Spiegelhalter D (1996) Markov chain Monte Carlo in practice. Chapman & Hall, London

Chapter 2

Novel Methods for Extending Optical Code Set for Coherent OCDMA

Chen Peng

Abstract Compared with that used in incoherent system, the optical code (OC) for coherent OCDMA system is analyzed in definition, characters, correlation performances and construction method. Aperiodic correlation property of OC is newly defined, and made the criterion for the judgment of OC' performance in a coherent OCDMA system. A novel differential algorithm is firstly proposed for extending OC set with optimal aperiodic auto-correlation property. With such algorithm ten iterative shows that global aperiodic auto-correlation property and capacity are respectively twice and 3713 bigger than original 127-chip Gold sequence. In order to obtain targeted user number's codes in a finite code set with optimal aperiodic cross-correlation property of OC, a novel cross searching algorithm is firstly given. As an example, 49 codes with good aperiodic cross-correlation property (>22) and aperiodic auto-correlation property (>40) are selected from 129 Gold codes.

Keywords Optical code division multiple access • Aperiodic correlation • Differential algorithm • Optical code

2.1 Introduction

Optical code division multiple access (OCDMA) is considered as one of the most promising technologies for the next generation broad-band access network, OCDMA schemes are classified as either incoherent or coherent OCDMA based

C. Peng (✉)
PLA Commanding Communications Academy, 45 No. Jiefang Park Road, Wuhan,
China
e-mail: bigroc.chen@163.com

on their operation principle. The incoherent techniques, which work on an optical-power- intensity basis, process OCs in a unipolar (0, 1) manner, and coherent techniques, working on a field-amplitude basis, process OCs in a bipolar (+1, -1) manner all optically [1–3]. Incoherent OCDMA which Optical Orthogonal Code (OOC) is widely used in has been studied since 1990s [4, 5], and there are many constrainers and bounds on size of these codes in the literature [2]. Recently, coherent OCDMA using ultra-short optical pulse has been receiving increasing attention due to the advances in reliable and compact en/decoder devices and detection schemes [1, 6–8], however the OC set appropriate for coherent OCDMA has not yet been created. Gold codes as an effective alternative are widely used for coherent OCDMA at present [3, 9, 10], but Gold which is designed for the optimization of periodic correlation property cannot realize the optimization of aperiodic correlation property for coherent OCDMA [10], so an appropriate method to select OCs for coherent OCDMA is very necessary.

2.2 Definition

Codes for coherent OCDMA based on temporal phase en/decoding is bipolar, code (1, 0) is respectively relevant to (+1, -1). Sequential pulses within a bit interval are copies of a unitary optical pulse, and therefore they are coherent. However, pulses from different bit interval are incoherent since they come from different origin, so aperiodic correlation is appropriate for measuring the codes' property of coherent OCDMA. If code is denoted by (+1, -1), aperiodic correlation is defined as follows.

Aperiodic auto-correlation:

$$R_{xx}(m) = \sum_{i=1}^{N-|m|} x_i x_{i+|m|} \quad m = 0, \pm 1, \pm 2, \dots, \pm(N-1) \quad (2.1)$$

Aperiodic cross-correlation:

$$\begin{aligned} R_{xy}(m) &= \sum_{i=1}^{N-m} x_i y_{i+m} \quad m = 0, 1, 2, \dots, N-1 \\ R_{yx}(m) &= \sum_{i=1}^{N-|m|} y_i x_{i+|m|} \quad m = 0, -1, -2, \dots, -(N-1) \end{aligned} \quad (2.2)$$

Bipolar periodic correlation is given and followed as a contrast. Periodic auto-correlation:

$$R_{xx}(m) = \sum_{i=1}^N x_i x_{i+|m|} \quad m = 0, \pm 1, \pm 2, \dots, \pm(N-1) \quad (2.3)$$

Periodic cross-correlation:

$$R_{xy}(m) = \sum_{i=1}^N x_i y_{i+m} \quad m = 0, 1, 2, \dots, N-1$$

$$R_{xy}(m) = \sum_{i=1}^N y_i x_{i+|m|} \quad m = -1, -2, \dots, -(N-1)$$

$$x_i, y_i = 1 \text{ or } -1 \quad x_i = x_{i+N} \quad y_i = y_{i+N}$$
(2.4)

Aperiodic and periodic correlations are clearly distinguished from their definitions.

The ratio of aperiodic auto-correlation peak and the maximum of wing (P/W) and the ratio of aperiodic auto-correlation peak and the maximum of aperiodic cross-correlation (P/C) are the measurement of aperiodic auto- and cross-correlation properties respectively [3, 11].

$$P/W = \left(\frac{R_{xx}(0)}{\text{Max}(|R_{xx}(m)|)} \right)^2 \quad m = \pm 1, \pm 2, \dots, \pm(N-1) \quad (2.5)$$

$$P/C = \left(\frac{R_{xx}(0)}{\text{Max}(|R_{xy}(m)|)} \right)^2 \quad m = 0, \pm 1, \pm 2, \dots, \pm(N-1) \quad (2.6)$$

Analyzing the features of aperiodic auto- and cross-correlation, some basic rules are listed as follows:

Proposition 1:

Code (x_1, x_2, \dots, x_N) and $(x_N, x_{N-1}, \dots, x_1)$ have the same aperiodic auto-correlation property.

Proposition 2:

Code (x_1, x_2, \dots, x_N) and $(-x_1, -x_2, \dots, -x_N)$ have the same aperiodic auto-correlation property.

From Proposition 1 and Proposition 2, asymmetric code has the same aperiodic auto-correlation property as three other codes.

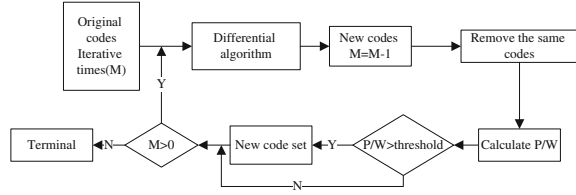
Proposition 3:

Code (x_1, x_2, \dots, x_N) and $(x_N, x_1, \dots, x_{N-1})$ have the different aperiodic auto-correlation property. That is the obvious difference between aperiodic and periodic correlation property.

2.3 Differential Algorithm

Codes based on the optimization of aperiodic correlation have not been created and it is impossible to judge aperiodic correlation property of all codes, so it is very urgent to discover an appropriate method for realizing the optimal aperiodic

Fig. 2.1 Flow chart of differential algorithm



correlation property of OCs. Differential algorithm is proposed to extend code set having good aperiodic auto-correlation property, its fundament is to obtain new codes from original codes using differential algorithm, and then decide whether to save these new codes through judging P/W one by one, so the number of new code set can be gradually grown under the control of iterative times (as Fig. 2.1).

The differential formula is defined as follows:

$$b_k = a_k \oplus b_{k-1} \quad b_k = 1 - a_k \oplus b_{k-1} \quad (2.7)$$

$$b_0 = 1 \text{ or } 0 \quad k = 1, 2, \dots, N$$

Where \oplus denotes modulo 2 addition, and a_k, b_k represent for the k th element of original code and new code respectively, so four new codes are generated from one code through the proposed differential algorithm.

Differential algorithm can promise good aperiodic auto-correlation property of new codes through setting appropriate threshold. As an example, 18 Gold codes with 127-chip which are randomly selected from a set of Gold codes are used as original codes, and the targeted threshold of aperiodic auto-correlation property is 50. Figure 2.2 shows that the obtained code number grows with the iterative times. After 10 times, we get 67162 codes whose aperiodic auto-correlation property are all beyond 50. Obviously, capacity is 3713 times bigger than original codes and can be farther grown by adding iterative times. To some extent the number of obtained codes is divergent for this threshold which is receivable for coherent OCDMA system, the result shows that differential algorithm is appropriate to extend code set. Figure 2.3 shows the comparison of P/W between a set of Gold codes which is used for selecting original codes and obtained codes by differential algorithm. Bold line and thin line give the P/W frequency of new codes and the set of Gold codes whose capacity is 127×129 because of proposition 3. Figure 2.4 also shows that P/W of that set of Gold codes and new codes is concentrated in 40 and 80 respectively, so global P/W of OC set is distinctly improved by differential algorithm.

Each user is assigned a unique signature OC which can distinguish itself from other users in the OCDMA system [12], every user has a unique couple of encoder and decoder designed by the assigned code, encoders are used for encoding the optical pulse from every user, and then decoders which are matched for different encoders are used for decoding the encoding signal, so original pulse signal from different users are recovered respectively. In the circumstance of ignoring noise, interrupts among users' decoding signal are dominant for recovering signal pulse,

Fig. 2.2 Capacity increased with iterative times

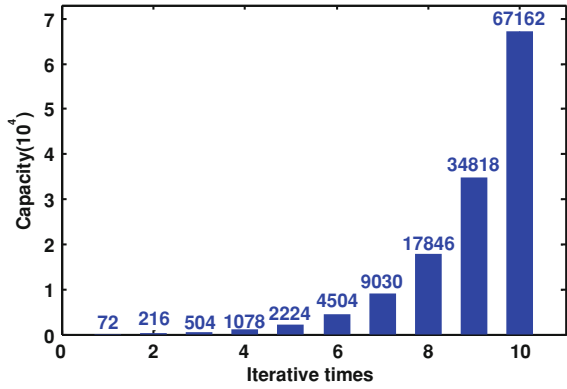


Fig. 2.3 Comparison between a set of gold and the result of differential algorithm

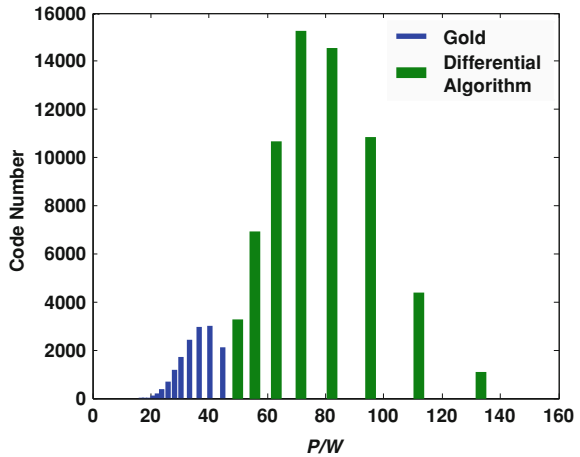
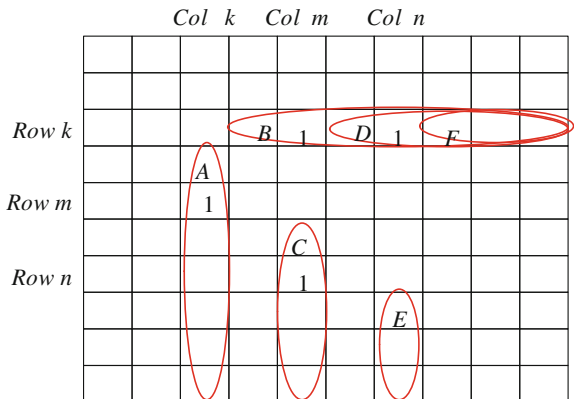


Fig. 2.4 Diagram of cross searching algorithm



and P/C of different user codes are the measurement of interrupts in coherent OCDMA system, so it is critical for coherent OCDMA to search codes having big P/C of each other.

2.4 Cross Searching Algorithm

In essence, the process of searching codes for OCDMA is to drag a sub-set which has enough big P/C of each other from code set having big P/W . In this paper, cross searching algorithm is given to do this work. Now we have M codes which P/W satisfy the demand, the flow of choosing N codes whose $P/C > V$ from M codes as follows:

- (1) Built a $M \times M$ zero-matrix, the value of (i, j) is prepared for the P/C of code i and code j , $k = 1$;
- (2) Calculate the values of P/C between code k and code $k + 1, k + 2, \dots, M$, through judging if $P/C > V$, respectively put 1 or 0 to the relevant position of matrix $(k, k + 1), (k, k + 2), \dots, (k, M)$ (region B), region A is the same as B, calculating $K(1)$ which denotes the number of nonzero in k -th row and saving nonzero column number (m, n, \dots) . If $K(1) > N - 1$ continue next step, else $k = k + 1$, and return to (2);
- (3) Calculate the values of P/C between code m and code $m+1, m+2, \dots, M$, through judging if $P/C > V$, respectively put 1 or 0 to the relevant position of matrix $(m + 1, m), (m + 2, m), \dots, (M, m)$ (region C), calculating $K(2)$ which denotes the number of the same element whose value is 1 between $(k, m + 1), (k, m + 2), \dots, (k, M)$ (region D) and $(m + 1, m), (m + 2, m), \dots, (M, m)$ (region C) and saving the same element (n, l, \dots) . If $K(2) > N - 2$, so two codes are chosen, continue next step, else $k = k + 1$, and return to (2);
- (4) Calculate the values of P/C between code n and code $n + 1, n + 2, \dots, M$, through judging if $P/C > V$, respectively put 1 or 0 to the relevant position of matrix $(n + 1, n), (n + 2, n), \dots, (M, n)$, calculating $K(3)$ which denotes the number of the same element whose value is 1 between $(k, n + 1), (k, n + 2), \dots, (k, M)$ (region F) and $(n + 1, n), (n + 2, n), \dots, (M, n)$ (region E) and saving the same element (l, \dots) . If $K(3) > N - 3$, so three codes are chosen, continue next step, else $k = k + 1$, and return to (2);
- (5) Next step is the same as 3 or 4, and terminate until N codes are chosen.

Figure 2.4 shows the process of searching, ellipses show the region of codes needed to be filled with appropriate value 1 or 0 in every step.

The algorithm reduces effectively the unnecessary calculating time by judging the possibility of finding enough optical codes in next steps when every new code is added to the selected code set, and promises that P/C of the code set is beyond threshold by judging the P/C between every new code and the chosen codes.

As an example, 49 OCs $P/W > 40$ and $P/C > 22$ are obtained by cross searching algorithm from 129 Gold codes. Figure 2.5 shows the aperiodic

Fig. 2.5 Aperiodic correlation properties of 49 OC

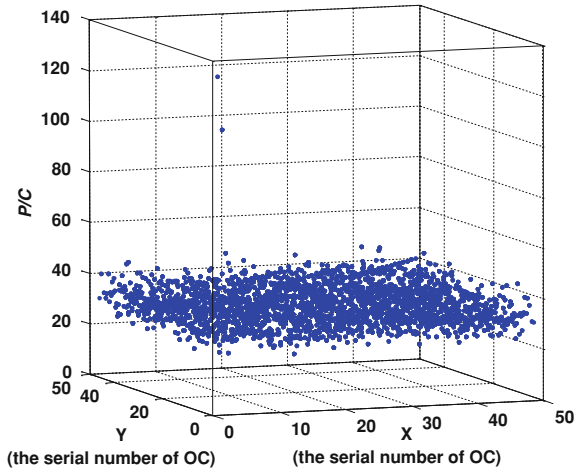


Fig. 2.6 OC subset used widely in OCDMA system

P/C >	17.9				20.5			
P/W >	17.9	30	35	38	42	35	38	42
Number of codes	129	110	82	48	29	33	22	17

correlation, and the value of diagonal denotes P/W of codes, and the value of (x, y) denotes the P/C between code x and code y . We can take a comparison between above result and OC set (as Fig. 2.6) which is widely used in OCDMA system [11], there are only 17 OC that satisfy the $P/W > 42$ and $P/C > 20.5$. Our result has the obvious advantage comparing to that, so the differential and cross searching algorithm we proposed are very useful.

2.5 Conclusion

Coherent OCDMA system based on Super Structured Fiber Bragg Gratings (SSFBG) and Planar Lightwave Circuits (PLC) has been the most promising technologies because of the advantages of all optical processing, soft capacity on demand, protocol transparency. An important factor to realize OCDMA system is the OCs' optimization of aperiodic correlation property. The paper obtains three fundamental rules from the definition of aperiodic correlation, and proposes a novel differential algorithm for extending the code set and improving the global aperiodic auto-correlation property, and cross searching algorithm is firstly given for the optimization of aperiodic cross-correlation property. The result shows that

the OC set obtained by our methods is obviously better than the used OC set in capacity and aperiodic correlation property. This improvement means that we establish good base for next work to realize multi-user coherent OCDMA system.

References

1. Taro H, Xu W, Naoya W (2005) Ten-user truly asynchronous gigabit OCDMA transmission experiment with a 511-chip SSFBG en decoder. *J Light Technol* 24:95
2. Reza O, Kumar P (2006) Spreading sequences for asynchronous spectrally phase encoded optical CDMA. ISIT, Seattle
3. Xu W, Matsushima K, Kitayama K (2005) High-performance optical code generation and recognition by use of a 511-chip 640-Gchip/s phase-shifted super structured fiber Bragg grating. *Opt Lett* 30:355
4. Zhang J (1998) Application of strict optical orthogonal codes to asynchronous all-optical CDMA networks, ICCT'98
5. Zhang J (1999) Flexible optical fiber CDMA networks using strict optical orthogonal codes for multimedia broadcasting and distribution applications. *Trans Broadcasting* 45:106
6. Xu W, Kitayama K (2004) Analysis of beat noise in coherent and incoherent time-spreading OCDMA. *J Light Technol* 22:2226
7. Grunnet-Jepsen A, Johnson AE (1999) Demonstration of all-fiber sparse lightwave CDMA based on temporal phase encoding. *Photonics Technol Lett* 11:1283
8. Kitayama K, Xu W, Naoya W (2006) OCDMA over WDM PON-solution path to Gigabit-symmetric FTTH. *J Light Technol* 22:1654
9. Teh PC, Petropoulos P, Ibsen M, Richardson DJ (2001) A comparative study of the performance of seven- and 63-chip optical code-division multiple-access encoders and decoders based on super structured fiber Bragg gratings. *J Light Technol* 9:1352
10. Koji M, Xu W, Kutsuzawa S et al (2004) Experimental demonstration of performance improvement of 127-Chip SSFBG En/Decoder using apodization technique. *Photonics Technol Lett* 16:2192
11. Xu W, Koji M, Akihiko N et al (2004) High reflectivity superstructured FBG for coherent optical code generation and recognition. *Optical Express* 12:5457
12. Ma W, Chao Z, Hongtu P, Lin J (2002) Performance analysis on phase encoded OCDMA communication system. *J Light Tech* 20:798

Chapter 3

A Sequential Processing Method for Converted Measurement Kalman Filters Based on Orthogonal Transform

Junlin Tian, Chengyu Fu and Tao Tang

Abstract To effectively reduce computational requirements of converted measurement Kalman filters (CMKF), a sequential processing method was presented in this paper. Firstly, the spherical measurement of the target position was converted into Cartesian coordinate domain, resulting in converted measurement error and the corresponding statistics (mean and covariance). Secondly, a sequential processing method which sequentially treated scalar components of the converted measurement vector, was derived based on orthogonal transform. Finally, the proposed sequential processing method was utilized to implement the CMKF algorithm for a target tracking scenario. The simulation results show that the proposed method can effectively reduce the computational requirements of CMKF algorithm while achieving desired tracking performance.

Keywords Target tracking · Kalman filter · Converted measurement · Orthogonal transform · Sequential processing

J. Tian (✉) · C. Fu · T. Tang
Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu 610209,
China
e-mail: tianjunlin22@163.com

C. Fu
e-mail: cyfu@ioe.ac.cn

T. Tang
e-mail: prettang@gmail.com

J. Tian
Graduate School of the Chinese Academy of Sciences, Beijing 100039, China

3.1 Introduction

As a linear, unbiased, and minimum mean square error estimation method, Kalman filtering algorithm has been widely used in the problem of target state estimation, such as IR, optoelectronic, and radar target tracking [1–5]. In practice, the target dynamics are usually modeled in Cartesian coordinate frame, whereas the measurements of the target position are obtained in sensor coordinates, for example, the spherical coordinate frame. Thus this kind of tracking problem is connected with nonlinear estimation [2–4, 6, 7]. Basically there are two approaches to solve this nonlinearity problem. One is the extended Kalman filter (EKF) in which the nonlinear measurement equation is linearized through Taylor series expansion technique [8–10]. The other is the converted measurement Kalman filter (CMKF) which converts the spherical measurements into Cartesian coordinate domain and then utilizes the linear version of Kalman filtering algorithm to estimate the target state [11–16].

In many practical applications, reducing computational requirements while achieving desired tracking performance is an important issue in the design of tracking filters. The proposed sequential converted measurement Kalman filter (SCMKF), which based on orthogonal transform, sequentially performs measurement updates with each scalar component of the converted measurement vector, and thus provides an effective way to reduce the filter’s computational requirements. Simulation results demonstrate the effectiveness of the SCMKF algorithm.

The organization of this chapter is as follows. [Section 3.1](#) briefly describes the nonlinearity problem in target tracking and presents the basic block CMKF algorithm. In [Sect. 3.2](#) we theoretically derive the proposed sequential processing method for CMKF algorithm based on orthogonal transform. Simulation is carried out for a target tracking scenario in [Sect. 3.3](#). Besides, performance comparison between the proposed method and the block CMKF algorithm is also carried out. [Section 3.4](#) gives the conclusion.

3.2 Block CMKF Algorithm

Consider the following discrete-time linear dynamic model:

$$\mathbf{X}_k = \Phi_{k|k-1}\mathbf{X}_{k-1} + \Gamma_{k|k-1}\mathbf{W}_{k-1}, \mathbf{X}_k|_{k=0} = \mathbf{X}_0 \quad (3.1)$$

where the Cartesian state vector \mathbf{X}_k consists of position and velocity of the target, i.e., $\mathbf{X} = (x, \dot{x}, y, \dot{y}, z, \dot{z})^T$ and the process noise \mathbf{W}_{k-1} is assumed to be zero-mean white Gaussian with covariance \mathbf{Q}_{k-1} . The system matrices $\Phi_{k|k-1}$ and

$$\Phi_{k|k-1} = \text{diag}(F_2, F_2, F_2), F_2 = \begin{pmatrix} 1 & T \\ 0 & 1 \end{pmatrix}$$

$\Gamma_{k|k-1}$ can be expressed as:

$$\Gamma_{k|k-1} = \text{diag}(G_2, G_2, G_2), G_2 = \begin{pmatrix} T^2/2 \\ T \end{pmatrix},$$

where T is the sampling interval.

It is assumed that the sensor measures the target range r , azimuth angle θ , and elevation angle ϕ . As a result, the measurement equation is described by the following discrete-time nonlinear equation:

$$\begin{aligned} Y_k &= h(X_k) + V_k = \begin{pmatrix} r_k \\ \theta_k \\ \phi_k \end{pmatrix} + \begin{pmatrix} \tilde{r}_k \\ \tilde{\theta}_k \\ \tilde{\phi}_k \end{pmatrix} \\ &= \begin{pmatrix} (x_k^2 + y_k^2 + z_k^2)^{1/2} \\ \arccos [x_k / (x_k^2 + y_k^2)^{1/2}] \\ \arccos [(x_k^2 + y_k^2)^{1/2} / (x_k^2 + y_k^2 + z_k^2)^{1/2}] \end{pmatrix} + \begin{pmatrix} \tilde{r}_k \\ \tilde{\theta}_k \\ \tilde{\phi}_k \end{pmatrix} \end{aligned} \quad (3.2)$$

where the measurement noises \tilde{r}_k , $\tilde{\theta}_k$, $\tilde{\phi}_k$ are assumed mutually uncorrelated and zero-mean white Gaussian with variance matrix expressed by

$$R_k = \text{diag}\{\sigma_r^2, \sigma_\theta^2, \sigma_\phi^2\}$$

In the implementation of CMKF algorithms, the spherical measurement vector $(r_m, \theta_m, j_m)^T$ is converted to Cartesian coordinate frame via the following transformation:

$$Y_k^c = \begin{pmatrix} x_{mk} \\ y_{mk} \\ z_{mk} \end{pmatrix} = \begin{pmatrix} r_{mk} \cos\theta_{mk} \cos\phi_{mk} \\ r_{mk} \sin\theta_{mk} \cos\phi_{mk} \\ r_{mk} \sin\phi_{mk} \end{pmatrix} \quad (3.3)$$

where x_{mk} , y_{mk} , and z_{mk} are scalar components of the Cartesian coordinate measurement vector Y_k^c . From (3.3), the converted measurement error vector in Cartesian coordinate frame can be represented as

$$V_k^c = \begin{pmatrix} \tilde{x}_{mk} \\ \tilde{y}_{mk} \\ \tilde{z}_{mk} \end{pmatrix} = \begin{pmatrix} x_{mk} - \hat{x}_k \\ y_{mk} - \hat{y}_k \\ z_{mk} - \hat{z}_k \end{pmatrix} = \mu_k^c + \tilde{V}_k^c \quad (3.4)$$

where μ_k^c is the mean vector of converted measurement error and \tilde{V}_k^c is the debiased converted measurement error vector. The first two moments (i.e., mean

and covariance) can be obtained in terms of target's true range r , azimuth angle θ , and elevation angle ϕ , as given below:

$$\begin{aligned}\mu_k^c(r, \theta, \phi) &= E(\mathbf{V}_k^c | r, \theta, \phi) \\ \mathbf{R}_k^c(r, \theta, \phi) &= \text{Var}(\mathbf{V}_k^c | r, \theta, \phi) = E(\tilde{\mathbf{V}}_k^c \tilde{\mathbf{V}}_k^{cT} | r, \theta, \phi)\end{aligned}\quad (3.5)$$

The above expressions of mean and covariance are unrealizable since the target's true position is not available in practice. To solve this problem, the measurement-conditioned method can be adopted to approximate the true mean and covariance. Details concerning this issue can be found in [17], and the measurement-conditioned moments can be expressed as

$$\begin{aligned}\mu_k^{c*} &= E[\mu_k^c(r, \theta, \phi) | r_m, \theta_m, \phi_m] = \mu_k^{c*}(r_m, \theta_m, \phi_m) \\ \mathbf{R}_k^{c*} &= E[\mathbf{R}_k^c(r, \theta, \phi) | r_m, \theta_m, \phi_m] = \mathbf{R}_k^{c*}(r_m, \theta_m, \phi_m)\end{aligned}\quad (3.6)$$

With the aforementioned converted measurement and the corresponding first two moments, the CMKF algorithm is listed below. Time Propagation:

$$\begin{aligned}\hat{\mathbf{X}}_{k|k-1} &= \Phi_{k|k-1} \hat{\mathbf{X}}_{k-1|k-1} \\ \mathbf{P}_{k|k-1} &= \Phi_{k|k-1} \mathbf{P}_{k-1|k-1} \Phi_{k|k-1}^T + \Gamma_{k|k-1} \mathbf{Q}_{k-1} \Gamma_{k|k-1}^T\end{aligned}\quad (3.7)$$

Measurement Update:

$$\begin{aligned}\mathbf{K}_k &= \mathbf{P}_{k|k-1} \mathbf{H}_k^T (\mathbf{H}_k \mathbf{P}_{k|k-1} \mathbf{H}_k^T + \mathbf{R}_k^{c*})^{-1} \\ \hat{\mathbf{X}}_{k|k} &= \hat{\mathbf{X}}_{k|k-1} + \mathbf{K}_k (\mathbf{Y}_k^c - \mathbf{H}_k \hat{\mathbf{X}}_{k|k-1} - \mu_k^{c*}) \\ \mathbf{P}_{k|k} &= (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k) \mathbf{P}_{k|k-1}\end{aligned}\quad (3.8)$$

with observation matrix:

$$\mathbf{H}_k = \begin{pmatrix} 100000 \\ 001000 \\ 000010 \end{pmatrix}$$

3.3 Sequential Algorithm Based on Orthogonal Transform

Although it is assumed that the measurement errors are mutually uncorrelated among spherical coordinates, the converted measurement errors in Cartesian coordinates are coupled with each other, resulting in nondiagonal covariance matrix. In this way, the sequential processing method cannot be applied.

It is reasonable to assumed that the covariance matrix \mathbf{R}_k^{c*} of converted measurement error is real symmetric and positive definite. Then there exists an orthogonal matrix M such that the following holds

$$\mathbf{R}_k^{c*} = \mathbf{M}^T \text{diag}(\lambda_1, \lambda_2, \lambda_3) \mathbf{M} \quad (3.9)$$

where the orthogonal matrix \mathbf{M} is composed of eigenvectors $\mathbf{e}_i (i = 1, 2, 3)$ associated with eigenvalues $\lambda_i (i = 1, 2, 3)$, i.e., $\mathbf{M} = (\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$. Note that both the eigenvalues $\lambda_i (i = 1, 2, 3)$ and orthogonal matrix \mathbf{M} depend on spherical measurements $r_m, \theta_m,$ and ϕ_m , thus can be expressed as $\lambda_i = \lambda_i(r_m, \theta_m, \phi_m) (i = 1, 2, 3)$, $\mathbf{M} = \mathbf{M}(r_m, \theta_m, \phi_m)$.

From (3.3) and (3.6), the debiased converted measurement equation can be expressed as

$$\mathbf{Y}_k^c - \mu_k^{c*} = \mathbf{H}_k \mathbf{X}_k + \tilde{\mathbf{V}}_k^c$$

Since the covariance matrix \mathbf{R}_k^{c*} of converted measurement error can be diagonalized through orthogonal transform, resulting in orthogonal matrix \mathbf{M} and eigenvalues $\lambda_i (i = 1, 2, 3)$, the above debiased converted measurement equation can be transformed as

$$\mathbf{M}(\mathbf{Y}_k^c - \mu_k^{c*}) = \mathbf{M}\mathbf{H}_k \mathbf{X}_k + \mathbf{M}\tilde{\mathbf{V}}_k^c$$

Let $\tilde{\mathbf{V}}_k^{c*}$ denote the transformed converted measurement error term $\mathbf{M}\tilde{\mathbf{V}}_k^c$. Using the identity $\mathbf{M}^T \mathbf{M} = \mathbf{M}^{-1} \mathbf{M} = \mathbf{I}$, the mean and covariance can be calculated as follows:

$$\begin{aligned} \mathbf{E}(\tilde{\mathbf{V}}_k^{c*}) &= \mathbf{E}(\mathbf{M}\tilde{\mathbf{V}}_k^c) = \mathbf{M}\mathbf{E}(\tilde{\mathbf{V}}_k^c) = \mathbf{0} \\ \text{Var}(\tilde{\mathbf{V}}_k^{c*}) &= \text{Var}(\mathbf{M}\tilde{\mathbf{V}}_k^c) = \mathbf{M}\text{Var}(\tilde{\mathbf{V}}_k^c)\mathbf{M}^T = \mathbf{M}\mathbf{R}_k^{c*}\mathbf{M}^T \\ &= \mathbf{M}\mathbf{M}^T \text{diag}(\lambda_1, \lambda_2, \lambda_3) \mathbf{M}\mathbf{M}^T = \text{diag}(\lambda_1, \lambda_2, \lambda_3) \end{aligned} \quad (3.10)$$

Then the debiased converted measurement equation after orthogonal transformation can be expressed as

$$\mathbf{Y}_k^{c*} = \mathbf{H}_k^* \mathbf{X}_k + \tilde{\mathbf{V}}_k^{c*} \quad (3.11)$$

where

$$\begin{aligned} \mathbf{Y}_k^{c*} &= \mathbf{M}(\mathbf{Y}_k^c - \mu_k^{c*}), \mathbf{H}_k^* = \mathbf{M}\mathbf{H}_k, \tilde{\mathbf{V}}_k^{c*} = \mathbf{M}\tilde{\mathbf{V}}_k^c \\ \mathbf{E}(\tilde{\mathbf{V}}_k^{c*}) &= \mathbf{0}, \text{Var}(\tilde{\mathbf{V}}_k^{c*}) = \text{diag}(\lambda_1, \lambda_2, \lambda_3) \end{aligned}$$

Because the covariance matrix of debiased converted measurement error after orthogonal transformation is diagonal, we can sequentially process the scalar components of the converted measurement vector. Let us define

$$\mathbf{Y}_k^{c*} = \begin{pmatrix} \mathbf{Y}_{k,1}^{c*} \\ \mathbf{Y}_{k,2}^{c*} \\ \mathbf{Y}_{k,3}^{c*} \end{pmatrix}, \mathbf{H}_k^* = \begin{pmatrix} \mathbf{H}_{k,1}^* \\ \mathbf{H}_{k,2}^* \\ \mathbf{H}_{k,3}^* \end{pmatrix}$$

Then the SCMKF algorithm can be listed as follows. Time Propagation:

$$\begin{aligned}\hat{\mathbf{X}}_{k|k-1} &= \Phi_{k|k-1} \hat{\mathbf{X}}_{k-1|k-1} \\ \mathbf{P}_{k|k-1} &= \Phi_{k|k-1} \mathbf{P}_{k-1|k-1} \Phi_{k|k-1}^T + \Gamma_{k|k-1} \mathbf{Q}_{k-1} \Gamma_{k|k-1}^T\end{aligned}\quad (3.12)$$

Measurement Update:First Step:

$$\begin{aligned}s_{k,1} &= \mathbf{H}_{k,1}^* \mathbf{P}_{k|k-1} \mathbf{H}_{k,1}^{*T} + \lambda_1, \mathbf{K}_{k,1} = \frac{\mathbf{P}_{k|k-1} \mathbf{H}_{k,1}^{*T}}{s_{k,1}} \\ \hat{\mathbf{X}}_{k|k,1} &= \hat{\mathbf{X}}_{k|k-1} + \mathbf{K}_{k,1} \left(\mathbf{Y}_{k,1}^{c*} - \mathbf{H}_{k,1}^* \hat{\mathbf{X}}_{k|k-1} \right) \\ \mathbf{P}_{k|k,1} &= \left(\mathbf{I} - \mathbf{K}_{k,1} \mathbf{H}_{k,1}^* \right) \mathbf{P}_{k|k-1}\end{aligned}\quad (3.13)$$

Second Step:

$$\begin{aligned}s_{k,2} &= \mathbf{H}_{k,2}^* \mathbf{P}_{k|k,1} \mathbf{H}_{k,2}^{*T} + \lambda_2, \mathbf{K}_{k,2} = \frac{\mathbf{P}_{k|k,1} \mathbf{H}_{k,2}^{*T}}{s_{k,2}} \\ \hat{\mathbf{X}}_{k|k,2} &= \hat{\mathbf{X}}_{k|k,1} + \mathbf{K}_{k,2} \left(\mathbf{Y}_{k,2}^{c*} - \mathbf{H}_{k,2}^* \hat{\mathbf{X}}_{k|k,1} \right) \\ \mathbf{P}_{k|k,2} &= \left(\mathbf{I} - \mathbf{K}_{k,2} \mathbf{H}_{k,2}^* \right) \mathbf{P}_{k|k,1}\end{aligned}\quad (3.14)$$

Third Step:

$$\begin{aligned}s_{k,3} &= \mathbf{H}_{k,3}^* \mathbf{P}_{k|k,2} \mathbf{H}_{k,3}^{*T} + \lambda_3, \mathbf{K}_{k,3} = \frac{\mathbf{P}_{k|k,2} \mathbf{H}_{k,3}^{*T}}{s_{k,3}} \\ \hat{\mathbf{X}}_{k|k} &= \hat{\mathbf{X}}_{k|k,2} + \mathbf{K}_{k,3} \left(\mathbf{Y}_{k,3}^{c*} - \mathbf{H}_{k,3}^* \hat{\mathbf{X}}_{k|k,2} \right) \\ \mathbf{P}_{k|k} &= \left(\mathbf{I} - \mathbf{K}_{k,3} \mathbf{H}_{k,3}^* \right) \mathbf{P}_{k|k,2}\end{aligned}\quad (3.15)$$

3.4 Simulation Results and Analysis

In order to demonstrate the effectiveness of the proposed SCMKF algorithm, we consider a target tracking problem. The target moves straightly with a nearly constant speed. The initial conditions are $(-6800, 1000$ and 1000 m) for position and $(680$ m/s, $0, 0)$ for velocity. The standard deviation of the acceleration errors is set at $(5 \times 10^{-2}$ m/s², 5×10^{-2} m/s², 5×10^{-2} m/s²). The sensor provides spherical position measurement of the target with sampling intervals 0.02 s.

The block CMKF algorithm and SCMKF algorithm are applied to estimate the spherical position of the target. These two algorithms are implemented under the same simulation environment (CPU 2.00 GHz, Memory 1.00 GB) to compare their performances in terms of time consumption and estimation accuracy. A Monte Carlo simulation of 100 runs is carried out to obtain the averaged estimation errors in spherical position. Because of space limit, only the simulation

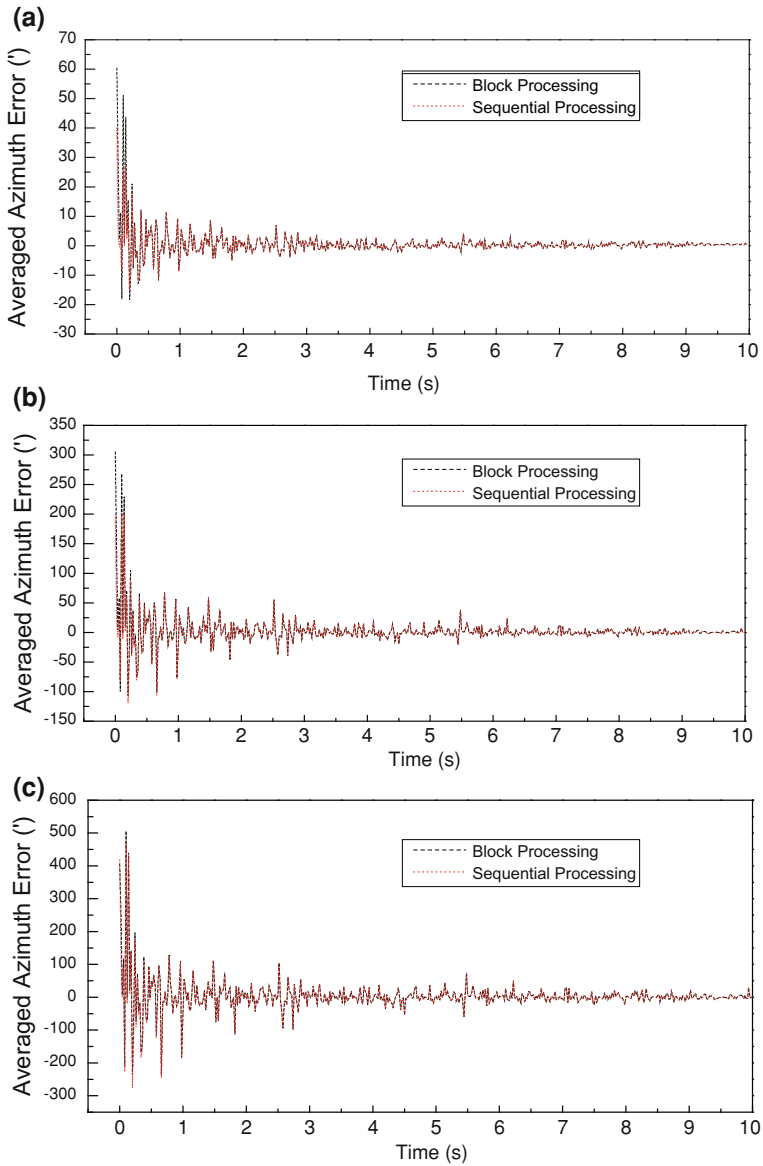


Fig. 3.1 Comparison of averaged azimuth errors between block processing and sequential processing methods: (a) $\sigma_r = 10$ m, $\sigma_\theta = \sigma_\varphi = 1^\circ$; (b) $\sigma_r = 50$ m, $\sigma_\theta = \sigma_\varphi = 5^\circ$; (c) $\sigma_r = 100$ m, $\sigma_\theta = \sigma_\varphi = 10^\circ$

results of azimuth component are presented here. Results for range and elevation components are similar to these of azimuth component.

The time consumption is 1.282 s for block CMKF algorithm, while 0.969 s for the SCMKF algorithm, both for 500 steps simulation (10 s). The superiority of the proposed SCMKF algorithm in time consumption is an important issue, especially for real time applications in which time delay can significantly undermine the tracking performance.

As shown in Fig. 3.1, the estimation accuracy of SCMKF algorithm is comparable with that of block CMKF algorithm. Similar results are obtained at different levels of spherical measurement error. These simulation results demonstrate the validity of the proposed SCMKF algorithm. Furthermore, the eigenvalues of covariance matrix of converted measurement error can be considered as the equivalent variances of the converted measurement error after orthogonal transformation. Thus the measurement update order, in which the scalar components of converted measurement vector are sequentially processed, will be investigated in future study.

References

1. Sun Z-G (2004) Infrared maneuver target tracking using angle-only measurements. *Laser & Infrared* 34:449–451
2. Miller KS, Leskiw DM (1982) Nonlinear estimation with radar observations. *IEEE Trans Aerosp Electron Syst* 18:192–200
3. Park S-T, Lee JG (2001) Improved Kalman filter design for three-dimensional radar tracking. *IEEE Trans Aerosp Electron Syst* 37:727–739
4. Park S-T, Lee JG (1998) Design of a practical tracking algorithm with radar measurements. *IEEE Trans Aerosp Electron Syst* 34:1337–1344
5. McEver MA, Kimbrell JE (2006) Tracking filter algorithm for automatic video tracking. Acquisition, tracking, and pointing XX. In: *Proceedings of SPIE*, vol 6238. pp 1–7
6. Clemons III TM, Chang KC (2008) Effect of sensor bias on space-based bearing-only tracker. Signal processing, sensor fusion, and target recognition XVII. In: *Proceedings of SPIE*, vol 6968. pp 1–9
7. Hardiman DF, Kerce JC, Brown GC (2006) Nonlinear estimation techniques for impact point prediction of ballistic targets. Signal and data processing of small targets 2006. In: *Proceedings of SPIE*, vol 6236. pp 1–12
8. Balaji B, Ding Z (2009) A performance comparison of nonlinear filtering techniques based on recorded radar datasets. Signal and data processing of small targets 2009. In: *Proceedings of SPIE*, vol 7445. pp 1–12
9. Yuan M, Chen J, Liang X (2008) An improved adaptive tracking algorithm based on “current” statistical model. Seventh international symposium on instrumentation and control technology: measurement theory and systems and aeronautical equipment. In: *Proceedings of SPIE*, vol 7128 pp 1–7
10. Tian X, Bar-Shalom Y (2007) Robust tracking for very long range radars, part I: algorithm comparisons. Signal and data processing of small targets. In: *Proceedings of SPIE*, vol 6699. pp 1–11
11. Lerro D, Bar-Shalom Y (1993) Tracking with debiased consistent converted measurements versus EKF. *IEEE Trans Aerosp Electron Syst* 29:1015–1022

12. Zhao Z, Li XR, Jilkov VP (2004) Best linear unbiased filtering with nonlinear measurements for target tracking. *IEEE Trans Aerosp Electron Syst* 40:1324–1336
13. Gray JE, Alouani AT (2007) Characterization of polar to Cartesian coordinates transformation and its effect on target track quality. Acquisition, tracking, pointing, and laser systems technologies XXI. In: *Proceedings of SPIE*, vol 6569. pp 1–14
14. Subramaniam M, Tharmarasa R, McDonald M, Kirubarajan T (2008) Passive tracking with sensors of opportunity using passive coherent location. *Signal and data processing of small targets 2008*. In: *Proceedings of SPIE*, vol 6969. pp 1–12
15. Ritter LJ, Weir B (2007) The effect of various filters on covariance consistency in the presence of a nonlinear tracking problem. *Signal and data processing of small targets 2007*. In: *Proceedings of SPIE*, vol 6699. pp 1–12
16. Mei W, Bar-Shalom Y (2009) Unbiased Kalman filter using converted measurements: revisit. *Signal and data processing of small targets 2009*. In: *Proceedings of SPIE*, vol 7445. pp 1–9
17. Suchomski P (1999) Explicit expressions for debiased statistics of 3D converted measurements. *IEEE Trans Aerosp Electron Syst* 35:368–370

Chapter 4

Improvement of an ID-Based Threshold Signcryption Scheme

Wei Yuan, Liang Hu, Xiaochun Cheng, Hongtu Li,
Jianfeng Chu and Yuyu Sun

Abstract Signcryption can realize the function of encryption and signature in a reasonable logic step, which can lower computational costs and communication overheads. In 2008, Fagen Li et al. proposed an efficient secure id-based threshold signcryption scheme. The authors declared that their scheme had the attributes of confidentiality and unforgeability in the random oracle model. However, our previous analysis shows that scheme is insecure against malicious attackers. Further, we propose a probably-secure improved scheme to correct the vulnerable and give the unforgeability and confidentiality of our improved scheme under the existing security assumption.

Keywords Identity-based · Signcryption · Bilinear pairing · Cryptanalysis

W. Yuan · L. Hu · H. Li · J. Chu (✉) · Y. Sun

Department of Computer Science and Technology, Jilin University, Changchun, China
e-mail: chujf@jlu.edu.cn

W. Yuan

e-mail: yuanwei1@126.com

L. Hu

e-mail: hul@mails.jlu.edu.cn

H. Li

e-mail: li_hongtu@hotmail.com

Y. Sun

e-mail: sunyy@ccu.edu.cn

X. Cheng

The School of Computing Science, Middlesex University, London, UK

e-mail: xiaochun.cheng@gmail.com

Y. Sun

Software Institute, Changchun University, Changchun, China

4.1 Introduction

Encryption and signature are the two basic cryptographic tools offered by public key cryptography for achieving confidentiality and authentication. Signcryption can realize the function of encryption and signature in a reasonable logic step which is proposed by Zheng [1]. Comparing to the traditional way of signature then encryption or encryption then signature, signcryption can lower the computational costs and communication overheads. As a result, a number of signcryption schemes [2–8] were proposed following Zheng’s work. The security notion for signcryption was first formally defined in 2002 by Baek et al. [9] against adaptive chosen ciphertext attack and adaptive chosen message attack. The same as signature and encryption, signcryption meets the attributes of confidentiality and unforgeability as well. In 1984, Shamir [10] introduced identity-based public key cryptosystem, in which a user’s public key can be calculated from his identity and defined hash function, while the user’s private key can be calculated by a trusted party called Private Key Generator (PKG). The identity can be any binary string, such as an email address and needn’t to be authenticated by the certification authentication. As a result, the identity-based public key cryptosystem simplifies the program of key management to the conventional public key infrastructure. In 2001, Boneh and Franklin [11] found bilinear pairings positive in cryptography and proposed the first practical identity-based encryption protocol using bilinear pairings. Soon, many identity-based [12–18] schemes were proposed and the bilinear pairings became important tools in constructing identity-based protocols. Group-oriented cryptography [19] was introduced by Desmedt in 1987. Elaborating on this concept, Desmedt and Frankel [20] proposed a (t, n) threshold signature scheme based RSA system [21]. In such a (t, n) threshold signature scheme, any t out of n signers in the group can collaboratively sign messages on behalf of the group for sharing the signing capability. Identity-based signcryption schemes combine the advantages of identity-based public key cryptosystem and Signcryption. The first identity-based threshold signature scheme was proposed by Baek and Zheng [22]. Then Duan et al. [23] proposed an identity-based threshold signcryption scheme in the same year by combining the concepts of identity based threshold signature and encryption together. However, in Duan et al.’s scheme, the master-key of the PKG is distributed to a number of other PKGs, which creates a bottleneck on the PKGs. In 2005, Peng and Li proposed an identity-based threshold signcryption scheme [24] based on Libert and Quisquater’s identity-based signcryption scheme [25]. However, Peng and Li’s scheme does not provide the forward security. In 2008, another scheme was proposed by Fagen Li et al. [26], which is more efficient comparing to previous scheme.

In this chapter, we show that the threshold signcryption scheme of Fagen Li et al. is vulnerable if the attacker can replace the group public key or even the attacker can intercept the intermediate messages. Further, we propose a probably-secure improved scheme to correct the vulnerable and give the unforgeability and confidentiality of our improved scheme under the existing security assumption.

4.2 The Improvement of Fagen Li et al.' Scheme

The scheme involves four roles: the PKG, a trust dealer, a sender group $U_A = \{M_1, M_2, \dots, M_n\}$ with identity ID_A and a receiver Bob with identity ID_B .

Setup: given a security parameter k , the PKG chooses groups G_1 and G_2 of prime order q (with G_1 additive and G_2 multiplicative), a generator P of G_1 , a bilinear map $e : G_1 \times G_1 \rightarrow G_2$, a secure symmetric cipher (E,D) and hash functions $H_1 : \{0, 1\}^* \rightarrow G_1$, $H_2 : G_2 \rightarrow \{0, 1\}^{n_1}$, $H_3 : \{0, 1\}^* \times G_1 \times \{0, 1\}^* \times G_1 \rightarrow Z_q^*$. The PKG chooses a master-key $s \in {}_R Z_q^*$ and computes $P_{pub} = sP$. The PKG publishes system parameters $\{G_1, G_2, n_1, e, P, P_{pub}, E, D, H_1, H_2, H_3\}$ and keeps the master-key s secret. **Extract:** Given an identity ID , the PKG computes $Q_{ID} = H_1(ID)$ and the private key $S_{ID} = sQ_{ID}$. Then PKG sends the private key to its owner in a secure way.

Keydis: suppose that a threshold t and n satisfy $1 \leq t \leq n < q$. To share the private key S_{ID_A} among the group U_A , the trusted dealer performs the steps below.

- 1) Choose F_1, \dots, F_{t-1} uniformly at random from G_1^* , construct a polynomial $F(x) = S_{ID_A} + xF_1 + \dots + x^{t-1}F_{t-1}$
- 2) Compute $S_i = F(i)$ for $i = 0, \dots, n$. ($S_0 = S_{ID_A}$). Send S_i to member M_i for $i = 1, \dots, n$ secretly.
- 3) Broadcast $y_0 = e(S_{ID_A}, P)$ and $y_j = e(F_j, P)$ for $j = 1, \dots, t-1$.
- 4) Each M_i then checks whether his share S_i is valid by computing $e(S_i, P) = \prod_{j=0}^{t-1} y_j^{i^j}$. If S_i is not valid, M_i broadcasts an error and requests a valid one.

Signcrypt: let M_1, \dots, M_t are the t members who want to cooperate to signcrypt a message m on behalf of the group U_A .

- 1) Each M_i chooses $x_i \in {}_R Z_q^*$, computes $R_{1i} = x_i P$, $R_{2i} = x_i P_{pub}$, $\tau_i = e(R_{2i}, Q_{ID_B})$ and sends (R_{1i}, τ) to the clerk C.
- 2) The clerk C (one among the t cooperating players) computes $R_1 = \prod_{i=1}^t R_{1i}$, $\tau = \prod_{i=1}^t \tau_i$, $k = H_2(\tau)$, $c = E_k(m)$, and $h = H_3(m, R_1, k, Q_{ID_A})$.
- 3) Then the clerk C sends h to M_i for $i = 0, \dots, t$.
- 4) Each M_i computes the partial signature $W_i = x_i P_{pub} + h \eta_i S_i$ and sends it to the clerk C, where $\eta = \prod_{j=1, j \neq i}^t -j(i-j)^{-1} \bmod q$.
- 5) Clerk C verifies the correctness of partial signatures by checking if the following equation holds: $e(P, W_i) = e(R_{1i}, P_{pub}) (\prod_{j=0}^{t-1} y_j^{i^j})^{h \eta_i}$

If all partial signatures are verified to be legal, the clerk C computes $W = \sum_{i=1}^t W_i$; otherwise rejects it and requests a valid one.

- 6) The final threshold signcryption is $\sigma = (c, R_1, W)$.

Unsigncrypt: when receiving σ , Bob follows the steps below.

- 1) Compute $\tau = e(R_1, S_{ID_B})$ and $k = H_2(\tau)$.

- 2) Recover $m = D_k(c)$
- 3) Compute $h = H_3(m, R_1, k, Q_{ID_A})$ and accept σ if and only if the following equation holds: $e(P, W) = e(P_{pub}, R_1 + hQ_{ID_A})$

4.3 Security Analysis of Our Improved Scheme

In this section, we will give a formal proof on Unforgeability and Confidentiality of our scheme under CDH problem and DBDH problem.

Theorem 1 (Unforgeability) *Our improved scheme is secure against chosen message attack under the random oracle model if CDH problem is hard.*

Proof Suppose the challenger C wants to solve the CDH problem. That is, given (aP, bP) C should compute abP .

C chooses system parameters $\{G_1, G_2, n_1, e, P, P_{pub}, E, D, H_1, H_2, H_3\}$, sets $P_{pub} = aP$, and sends parameters to the adversary E (the hash functions H_1, H_2, H_3 are random oracles).

H_1 query: C maintains a list L_1 to record H_1 queries. L_1 has the form of $(ID, \alpha, Q_{ID}, S_{ID})$. Suppose the adversary Eve can make H_1 queries less than q_{H_1} times. C selects a random number $j \in [1, q_{H_1}]$. If C receives the j -th query, he will return $Q_{ID_j} = bP$ to Eve and sets $(ID_j, \perp, Q_{ID_j} = bP, \perp)$ on L_1 . Else C selects $\alpha_i \in Z_q^*$ computes $Q_{ID_i} = \alpha_i P$, $S_{ID_i} = \alpha_i P_{pub}$, returns Q_{ID_i} to E and sets $(ID_i, \alpha_i, Q_i, S_i)$ on L_1 .

H_2 query: C maintains a list L_2 to record H_2 queries. L_2 has the form of (τ, k) . If C receives a query about τ_i , selects $k_i \in Z_q^*$, returns k_i to E, and sets (τ_i, k_i) on L_2 .

H_3 query: C maintains a list L_3 to record H_3 queries. L_3 has the form of (m, R, k, Q, h) . If C receives a query about $(m_i, R_{1i}, k_i, Q_{ID_i})$, selects $h_i \in Z_q^*$, returns h_i to Eve, and sets $(m_i, R_{1i}, k_i, Q_{ID_i}, h_i)$ on L_3 .

Signcrypt query: if C receives a query about Signcrypt with message m_i , identity ID_i

1. Select $x_i \in Z_q^*$, $W_i \in G_1$
2. Look-up L_1, L_2 , set $Q_{ID_i} = \alpha_i P$ in L_1 , $k_i = k_i$ in L_2 , and compute $R_i = x_i Q_{ID_i}$
3. Set $h_i = H_3(m_i, R_i, k_i, Q_{ID_i})$.
4. Return (h_i, W_i) to Eve.

Finally, Eve output a forged signcrypton (m, h_i, W_i, Q_{ID_i}) . If $Q_{ID_i} \neq Q_{ID_j}$, Eve fails. Else, if $Q_{ID_i} = Q_{ID_j}$, Eve succeeds in forging a signcrypton.

As a result, C gains two signcrypton ciphertexts which meet:

$$e(P, W_i) = e(P_{pub}, R_i + h_i Q_{ID_i})$$

$$e(P, W_j) = e(P_{pub}, R_j + h_j Q_{ID_j})$$

Thus,

$$e(P, (W_i - W_j)) = e(P_{pub}, (R_i + h_i Q_{ID_i}) - (R_j + h_j Q_{ID_j})) \quad (4.1)$$

Note $Q = Q_{ID_i} = Q_{ID_j}$, (4.1) can be expressed as

$$e(P, (W_i - W_j)) = e(P_{pub}, (R_i - R_j) + (h_i - h_j)Q) \quad (4.2)$$

$$\because P_{pub} = aP, Q_{ID_j} = bP$$

(4.2) can be expressed as $e(P, (W_i - W_j)) = e(aP, ((\alpha_i - \alpha_j) + (h_i - h_j))bP)$

$$\therefore W_i - W_j = ((\alpha_i - \alpha_j) + (h_i - h_j))abP$$

Hence, the CDH problem $abP = \frac{W_i - W_j}{(\alpha_i - \alpha_j) + (h_i - h_j)}$ can be computed by C with aP and bP .

Theorem 2 (Confidentiality) *Our improved scheme is secure against adaptive chosen ciphertext and identity attack under the random oracle model if DBDH problem is hard.*

Proof Suppose the challenger C wants to solve the DBDH problem. That is, given (P, aP, bP, cP, τ) , C should decide whether $\tau = e(P, P)^{abc}$ or not. If there exists an adaptive chosen ciphertext and identity attacker for our improved scheme, C can solve the DBDHP.

C chooses system parameters $\{G_1, G_2, n_1, e, P, P_{pub}, E, D, H_1, H_2, H_3\}$, sets $P_{pub} = aP$, and sends parameters to the adversary E (the hash functions H_1, H_2, H_3 are random oracles).

H_1 query: C maintains a list L_1 to record H_1 queries. L_1 has the form of $(ID, \alpha, Q_{ID}, S_{ID})$. Suppose the adversary Eve can make H_1 queries less than q_{H_1} times. C selects a random number $j \in [1, q_{H_1}]$. If C receives the j -th query, he will return $Q_{ID_j} = bP$ to Eve and sets $(ID_j, \perp, Q_{ID_j} = bP, \perp)$ on L_1 . Else C selects $\alpha_i \in Z_q^*$ computes $Q_{ID_i} = \alpha_i P$, $S_{ID_i} = \alpha_i P_{pub}$, returns Q_{ID_i} to E and sets $(ID_i, \alpha_i, Q_i, S_i)$ on L_1 .

H_2 query: C maintains a list L_2 to record H_2 queries. L_2 has the form of (τ, k) . If C receives a query about τ_i , selects $k_i \in Z_q^*$, returns k_i to E, and sets (τ_i, k_i) on L_2 .

H_3 query: C maintains a list L_3 to record H_3 queries. L_3 has the form of (m, R, k, Q, h) . If C receives a query about $(m_i, R_{1i}, k_i, Q_{ID_i})$, selects $h_i \in Z_q^*$, returns h_i to Eve, and sets $(m_i, R_{1i}, k_i, Q_{ID_i}, h_i)$ on L_3 .

Signcrypt query: if C receives a query about Signcrypt with message m_i , identity ID_i

1. Select $c_i \in Z_q^*$, $W_i \in G_1$
2. Look-up L_1, L_2 , set $Q_{ID_i} = \alpha_i P$ in L_1 , $k_i = k_i$ in L_2 . Compute $R_i = c_i P$, if $ID_i \neq ID_j$. Else, if $ID_i = ID_j$, compute $R_i = cP$

3. Set $h_i = H_3(m_i, R_i, k_i, Q_{ID_i})$.
4. Return (h_i, W_i) to Eve.

After the first stage, Eve chooses a pair of identities on which he wishes to be challenged on (ID_i, ID_j) . Note that Eve can not query the identity of ID_A . Then Eve outputs two plaintexts m_0 and m_1 . C chooses a bit $b \in \{0, 1\}$ and signcrypts m_b . To do so, he sets $R_1^* = cP$, obtains $k^* = H_2(\tau)$ from the hash function H_2 , and computes $c_b = E_{k_1^*}(m_b)$. Then C chooses $W^* \in G_1$ and sends the ciphertext $\sigma^* = (c_b, R_1^*, W^*)$ to Eve. Eve can perform a second series of queries like at the first one. At the end of the simulation, she produces a bit b' for which he believes the relation $\sigma^* = \text{Signcrypt}(m_{b'}, \{S_i\}_{i=1, \dots, t}, ID_j)$ holds. If $b = b'$, C outputs $\tau = e(R_1^*, S_{ID_i}) = e(cP, abP) = e(P, P)^{abc}$. Else, C outputs $\tau \neq e(P, P)^{abc}$. So C can solve the BDDH problem.

4.4 Conclusion

In this chapter, we show that the threshold signcryption scheme of Fagen Li et al. is vulnerable if the attacker can replace the group public key. Then we point out that the receiver uses the sender's public key without any verification in the unsigncrypt stage cause this attack. Further, we propose a probably-secure improved scheme to correct the vulnerable and give the unforgeability and confidentiality of our improved scheme under the existing security assumption.

Acknowledgment The authors would like to thank the editors and anonymous reviewers for their valuable comments. This work is supported by the National Natural Science Foundation of China under Grant No. 60873235 and 60473099, the National Grand Fundamental Research 973 Program of China (Grant No. 2009CB320706), Scientific and Technological Developing Scheme of Jilin Province (20080318), and Program of New Century Excellent Talents in University (NCET-06-0300).

References

1. Zheng Y (1997) Digital signcryption or how to achieve cost (signature & Encryption) \ll cost (signature) + cost (encryption). In: Proceedings of advances in CRYPTO'97, LNCS 1294. Springer, Berlin, pp 165–179
2. Bao F, Deng RH (1997) A signcryption scheme with signature directly verifiable by public key. PKC'98 LNCS, vol 1431. Springer, Berlin, pp 55–59
3. Chow SSM, Yiu SM, Hui LCK, Chow KP (2004) Efficient forward and provably secure ID-based signcryption scheme with public verifiability and public ciphertext authenticity. ICISC'03 LNCS, vol 2971. Springer, Berlin, pp 352–269
4. Boyen X, Multipurpose identity based signcryption: a swiss army knife for identity based cryptography. CRYPT'03 LNCS, vol 2729. Springer, Berlin, pp 383–399
5. Mu Y, Varadharajan V (2000) Distributed signcryption. INDOCRYPT'00. LNCS, vol 1977. Springer, Berlin, pp 155–164

6. Yang G, Wong DS, Deng X (2005) Analysis and improvement of a signcryption scheme with key privacy. ISC'05. LNCS, vol 3650. Springer, Berlin, pp 218–232
7. Steinfeld R, Zheng Y (2000) A signcryption scheme based on integer factorization. ISW'00. LNCS, vol 1975. Springer, Berlin, pp 308–322
8. Libert B, Quisquater J (2004) Efficient signcryption with key precovery from gap Diffie-Hellman groups. PKC'04. LNCS vol 2947. Springer, Berlin, pp 187–200
9. Baek J, Steinfeld R, Zheng Y (2002) Formal proofs for the security of signcryption. PKC'02. LNCS vol 2274. Springer, Berlin, pp 80–98
10. Shamir A (1984) Identity-based cryptosystems and signature schemes. CRYPTO'84. LNCS vol 196. Springer, Berlin, pp 47–53
11. Boneh D, Franklin M (2001) Identity-based encryption from well pairing. CRYPTO'01. LNCS vol 2139. Springer, Berlin, pp 213–229
12. Barreto PSLM, Libert B, McCullagh N, Quisquater JJ (2005) Efficient and provably-secure identity-based signatures and signcryption from bilinear maps. ASIACRYPT'05. LNCS, vol 3788. Springer, Berlin, pp 515–532
13. Li F, Hu X, Nie X (2009) A new multi-receiver ID-based signcryption scheme for group communication. ICCAS'2009. IEEE Press, San Jose, pp 296–300
14. Han Y, Gui X (2009) Multi-recipient signcryption for secure group communication. ICIEA 2009, pp 161–165
15. Jin Z, Wen Q, Du H (2010) An improved semantically-secure identity-based signcryption scheme in the standard model. *Comput Electr Eng* 36(3):545–552
16. Huang X, Susilo W, Mu Y, Zhang E (2005) Identity-based ring signcryption schemes: cryptographic primitives for preserving privacy and authenticity in the ubiquitous world. 19th international conference on advanced information networking and applications, Taiwan, pp 649–654
17. Liu Z, Hu Y, Zhang X, Ma H (2010) Certificateless signcryption scheme in the standard model. *Inf Sci* 180(3):452–464
18. Yu Y, Bo Y, Sun Y, Zhu S-l (2009) Identity based signcryption scheme without random oracles. *Comput Stand Interfac* 31(1):56–62
19. Desmedt Y (1987) Society and group oriented cryptography: a new concept. CRYPTO'87. LNCS, vol 293. Springer, Berlin, pp 120–127
20. Desmedt Y, Frankel Y (1991) Shared generation of authenticators and signatures. CRYPTO'91. LNCS, vol 576. Springer, Berlin, pp 457–469
21. Rivest RL, Shamir A, Adleman L (1978) A method for obtaining digital signatures and public-key cryptosystems. *Commun ACM* 2(2):120–126
22. Baek J, Zheng Y (2004) Identity-based threshold signature scheme from the bilinear pairings. International conference on information technology, Las Vegas, pp 124–128
23. Duan S, Cao Z, Lu R (2004) Robust ID-based threshold signcryption scheme from pairings. International conference on information security, Shanghai, pp 33–37
24. Peng C, Li X (2005) An identity-based threshold signcryption scheme with semantic security. *Computational intelligence and security 2005*. LNAI, vol 3902. Springer, Berlin, pp 173–179
25. Libert B, Quisquater JJ (2003) A new identity based signcryption schemes from pairings. IEEE information theory workshop, Paris, pp 155–158
26. Li F, Yu Y (2008) An efficient and provably secure ID-based threshold signcryption scheme, ICCAS. Springer, Xiamen, pp 488–492
27. Malone LJ (2002) Identity based signcryption. In: *cryptology ePrint archive*. Report, (14):098–106
28. Chow SSM, Yiu SM, Hui LCK, Chow KP (2004) Efficient forward and provably secure ID-based signcryption scheme with public verifiability and public ciphertext authenticity. In: Lin J-I, Lee D-H (eds) ICISC 2003, LNCS, vol 2971. Springer, Berlin, pp 352–369
29. Boyen X (2003) Multipurpose identity based signcryption: a Swiss army knife for identity based cryptography. In: Boneh D (ed) CRYPTO 2003. LNCS, vol 2729. Springer, Berlin, pp 383–399

Chapter 5

Cryptanalysis of an Enhanced Event Signature Protocols for Peer-to-Peer Massively Multiplayer Online Games

Wei Yuan, Liang Hu, Hongtu Li and Jianfeng Chu

Abstract In 2008, Chan et al. presented an event signature (EASES) protocol for peer-to-peer massively multiplayer online games (P2P MMOGs). The authors declare that the EASES protocol is efficient and secure, and could achieve non-repudiation, event commitment, save memory, bandwidth and reduce the complexity of the computations. In 2010, Li et al. found a replay attack on the EASES protocol and proposed an enhanced edition to improve it. However, our works show their enhancement is still not secure as well. Finally, we made a discussion about this problem and point the weakness existence in this protocol.

Keywords Hash-chain · Cryptanalysis · Massive multiplayer online games · One-time signature · Peer-to-peer networks · MMOG · Network security

5.1 Introduction

Multiplayer on line games are a rapidly growing segment of Internet applications in the recent years. By providing more entertainment and sociability than single-player games, is fast becoming a major form of digital entertainment. In this kind

W. Yuan · L. Hu · H. Li · J. Chu (✉)

Department of Computer Science and Technology, Jilin University, Changchun, China
e-mail: chujf@jlu.edu.cn

W. Yuan
e-mail: yuanwei1@126.com

L. Hu
e-mail: hul@mails.jlu.edu.cn

H. Li
e-mail: li_hongtu@hotmail.com

of games, all players should connect with the server to send and receive event updates. An event update is cryptographic protocol by which a player generates an event message and sends it to the server for updating the game states. Traditional massively multiplayer online games (MMOGs) are conventional client–server models that do not scale with the number of simultaneous clients that need to be supported. To resolve conflicts in the simulation and act as a central repository for data, peer-to-peer (P2P) architecture is increasingly being considered as replacement for traditional client–server architecture in MMOGs. P2P MMOGs have many advantages over traditional client–server systems due to their network connectivity and basic network services in a self-organizing manner. Whenever a player wants to play the finger-guessing game, an event message is sent to the server and the server processes all the events and updates the game states to ensure a global ordering for game executions and fair plays. However, P2P MMOGs communicate on the Internet raise the security issues such as cheating that a dishonest player can get valuable virtual items and even be sold for moneymaking. Recently, there are more and more efforts mounted to focus on event update protocols for online games in respect to the protection of sensitive communication and the provision of fair play.

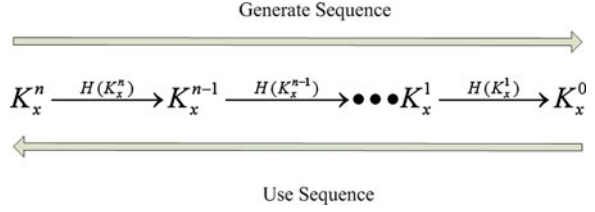
In 2004, Dickey et al. [1] proposed a low latency and cheat-proof event ordering based on digital signatures and voting mechanism for P2P games. However, Corman et al. [2] later show that Dickey et al.'s protocol is unable to prevent all cheats as claimed, and propose an improvement called secure event agreement protocol. As digital signature requires a large amount of computations. To reduce heavyweight computations in every round of a game session, in 2008, Chan et al. [3] proposed an efficient and secure event signature (EASES) protocol using one-time signature with hash-chain key and claimed that their protocol has low computation and bandwidth costs, and is thus applicable to P2P-based MMOGs. Then they proposed a dynamic EASES protocol to avoid the pre-generation of hash-chain keys. Unfortunately, the EASES protocol is not secure and attackers can easily forge a series of update event to replace the original one. In 2010, Li et al. [4] found a replay attack on the EASES protocol and suggested a simple enhanced edition. However, their enhanced protocol still suffered from our attack.

In this chapter, we briefly review the enhanced protocol proposed by Li et al. Further, we introduce attacking methods to crack this protocol. Finally, we make a discussion on why our attack does.

5.2 Review of Chan et al.'s Event Signature Protocol for P2P MMOGs

Chan et al.'s event signature protocol has four phases: the Initialization Phase, Signing Phase, Verification Phase, Re-initialization Phase.

Fig. 5.1 Construction of hash-chain keys



5.2.1 Initialization Phase

In this phase, player P_i generates a series of one-time signature keys for a session and performs the following operations:

1. P_i chooses a master key MK_i to compute the n th one-time signature key $K_i^n = h(MK_i)$, where n represents the maximum number of rounds in a session.
2. P_i computes the other r th round one-time signature keys $K_i^{r-1} = H(K_i^r)$, where $r = (n - 1), \dots, 0$.
3. P_i signs the first one-time signature key by its private key to get the signature $\Delta_i = S_{sk_i}(K_i^0)$. The hash-chain keys K_i^r will be used in the reverse order of their production during the subsequent r th rounds, where $r = 0, 1, 2, \dots, n - 1$. Figure 5.1 shows the production of hash-chain keys.

5.2.2 Signing Phase

In this phase, if P_i wants to submit event update messages to other online players in a game session with n rounds, he/she performs the following operations:

1. P_i computes the 1st round one-time signature key δ_i^1 by computing $\delta_i^1 = H(K_i^1 \parallel U_i^1), \Delta_i, K_i^0$. Then, P_x submits the first round message to other online players.
2. P_i computes the second round one-time signature key $\delta_i^2 = H(K_i^2 \parallel U_i^2), U_i^1, K_i^1$ and submits it to other online players.
3. P_i computes the r th round one-time signature key $\delta_i^r = H(K_i^r \parallel U_i^r), U_i^{r-1}, K_i^{r-1}$ and submits it to other online players in the subsequent r th round, where $r = 3, 4, \dots, n$.

5.2.3 Verification Phase

In this phase, each online player P_j receives the event update message $\delta_i^1 = H(K_i^1 \parallel U_i^1), \Delta_i, K_i^0$ from the player P_i and performs the following operations:

1. In the first round, P_j first verifies $\Delta_i \stackrel{?}{=} D_{pk_i}(K_i^0)$. If it holds, P_j confirms that the key K_x^0 is legitimate.
2. In the subsequent r th round, P_j verifies $K_i^{r-2} \stackrel{?}{=} H(K_i^{r-1})$ to check if the signature key K_i^{r-1} is legitimate, where $r = 2, 3, 4, \dots, n$.
3. If above holds, P_j verifies $\delta_i^{r-1} \stackrel{?}{=} H(K_i^{r-1} \parallel U_i^{r-1})$ to check whether the update has been altered or not. If it passes verification, P_j convinces that no player has tampered with the update from P_i .

5.2.4 Re-Initialization Phase

If P_i wants to extend his/her game session for a few rounds, P_i regenerates a new master key and performs the following operations:

1. In the n th round, P_i chooses a new master key MK_i' and generates the new one-time signature keys $NewK_i^0, \dots, NewK_i^n$. Then P_i has the new signature key $NewK_x^0$ with the key $K_i^n = H(MK_i')$ to generate $\delta_i^n = H(K_i^n \parallel U_i^n \parallel NewK_i^0), U_i^{n-1}, K_i^{n-1}$. P_i sends $\delta_i^n, U_i^{n-1}, K_i^{n-1}$ to other players as usual.
2. In the $(n + 1)$ th round, P_i sends $\delta_i^{n+1} = H(NewK_i^1 \parallel U_i^{n+1}), U_i^n K_i^n NewK_i^0$ to other players.
3. In the $(n + 2)$ th round, P_i sends $\delta_i^{n+2} = H(NewK_i^2 \parallel U_i^{n+2}), U_i^{n+1}, NewK_i^1, MK_i$ to other players.

Upon receiving new one-time signature keys from P_i , the other player, P_j , should perform the following verifiable operations:

1. In the $(n + 1)$ th round, P_j verifies $\delta_i^n \stackrel{?}{=} H(K_i^n \parallel U_i^n \parallel NewK_i^0)$ to check if the new signature key $NewK_x^0$ is legitimate.
2. In the $(n + 2)$ th round, in addition to the regular verifications, P_j must also verify $K_i^n \stackrel{?}{=} H(MK_i)$. If the above passes verification, P_j confirms the validity of $NewK_x^0$. The series of new one-time signature keys $NewK_i^0, \dots, NewK_i^n$ can be used after the $(n + 2)$ th rounds.

5.3 A Replay Attack on Chan et al.'s Protocol

In 2010, Li et al. has presented a replay attack to Chan et al.'s protocol then they made an enhanced one. In this section, we review the replay attack on Chan et al.'s protocol.

After one session ends, the attacker can get the one-time signature keys $K_i^0, K_i^1, \dots, K_i^n$, the signature $\Delta_i = S_{sk_i}(K_i^0)$ and the master key, MK_i , which is transmitted from legal player P_i to other players. Then he can forge event updates

$U_k^0, U_k^1, \dots, U_k^n$ with the valid signature keys $K_i^0, K_i^1, \dots, K_i^n$ to cheat other players in P2P based MMOG like formula (5.1):

$$\begin{cases} \delta_i^r = H(K_i^1 \parallel U_k^1), \Delta_i, K_i^0 & \text{in the first round} \\ \delta_i^{r'} = H(K_i^r k^r), \Delta_i^{r-1}, K_i^0 & \text{in the } r\text{th round} \end{cases} \quad (5.1)$$

Upon receiving event messages from the attacker, the other player computes the hash value of a given signature key K_i^r and fake event updates U_i^r to verify its equality to the previously received signature key δ_i^r , by computing $\delta_i^r = H(K_i^r \parallel U_k^r)$. Thus, the replay attack can not be prevented in Chan et al.'s protocol. Then Li et al. proposed their enhanced protocol.

5.4 Review of Chun-Ta Li et al.'s Enhanced Event Signature Protocol for P2P MMOGs

Li et al.'s enhanced EASES protocol also has four phases: the initialization phase, signing phase, verification phase, re-initialization phase.

5.4.1 Initialization Phase

In the initialization phase, player P_x generates a series of one-time signature keys for a session and performs the following operations:

1. P_x chooses a master key MK_x to compute the n th one-time signature key $K_x^n = h(MK_x)$, where n represents the maximum number of rounds in a session.
2. P_x computes the other r th round one-time signature keys $K_x^{r-1} = H(K_x^r)$, where $r = (n - 1), \dots, 0$.
3. P_x signs the first one-time signature key by its private key to get the signature $\Delta_x = S_{sk_x}(K_x^0 \parallel gno\#)$. Note that hash-chain keys K_x^r will be used in the reverse order of their production during the subsequent r th rounds, where $r = 0, 1, 2, \dots, n - 1$.

5.4.2 Signing Phase

If P_x wants to submit event update messages to other online players in a game session with n rounds, he/she performs the following operations:

1. P_x computes the first round one-time signature key δ_x^1 by computing $\delta_x^1 = H(K_x^1 \parallel U_x^1 \parallel gno\#)$, $\Delta_x, K_x^0, gno\#$. Then, P_x submits the first round message it to other online players.

2. P_x computes the second round one-time signature key $\delta_x^2 = H(K_x^2 \parallel U_x^2 \parallel gno\#)$, $U_x^1, K_x^1, gno\#$ and submits it to other online players.
3. P_x computes the r th round one-time signature key $\delta_x^r = H(K_x^r \parallel U_x^r \parallel gno\#)$, $U_x^{r-1}, K_x^{r-1}, gno\#$ and submits it to other online players in the subsequent r th round, where $r = 3, 4, \dots, n$

5.4.3 Verification Phase

In the verification phase, each online player P_y receives the event update message from the player P_x and performs the following operations:

1. In the first round, P_y first verifies $\Delta_x \stackrel{?}{=} D_{pk_x}(K_x^0 \parallel gno\#)$. If it holds, P_y confirms that the key K_x^0 is legitimate and $gno\#$ is not a duplicate value; if not, it stops.
2. In the subsequent r th round, P_y verifies $K_x^{r-2} \stackrel{?}{=} H(K_x^{r-1})$ to check if the signature key K_x^{r-1} is legitimate, where $r = 2, 3, 4, \dots, n$.
3. If above holds, P_y verifies $\delta_x^{r-1} \stackrel{?}{=} H(K_x^{r-1} \parallel U_x^{r-1} \parallel gno\#)$ to check whether the update has been altered or not. If it passes verification, P_y convinces that no player has tampered with the update from P_x .

5.4.4 Re-Initialization Phase

Whenever P_x wants to extend his/her game session for a few rounds, P_x regenerates a new master key and performs the following operations:

1. In the n th round, P_x chooses a new master key MK'_x and generates the new one-time signature keys $NewK_x^0, \dots, NewK_x^n$. Then P_x has the new signature key $NewK_x^0$ with the key $K_x^n = H(MK'_x)$ to generate $\delta_x^n = H(K_x^n \parallel U_x^n \parallel NewK_x^0 \parallel gno\#)$. P_x sends $\delta_x^n, U_x^{n-1}, K_x^{n-1}$ and $gno\#$ to other players in this round.
2. In the $(n + 1)$ th round, P_x computes $\delta_x^{n+1} = H(NewK_x^1 \parallel U_x^{n+1} \parallel gno\#)$, $U_x^n K_x^n NewK_x^0$ and sends δ_x^{n+1} to other players.
3. In the $(n + 2)$ th round, P_x sends $\delta_x^{n+2} = H(NewK_x^2 \parallel U_x^{n+2} \parallel gno\#)$, $U_x^{n+1}, NewK_x^1, gno\#, MK_x$ to other players.

Upon receiving new one-time signature keys from P_x , the other player, P_y , should perform the following verifiable operations:

1. In the $(n + 1)$ th round, P_y verifies $\delta_x^n \stackrel{?}{=} H(K_x^n \parallel U_x^n \parallel NewK_x^0 \parallel gno\#)$ to check if the new signature key $NewK_x^0$ is legitimate.

2. In the $(n + 2)$ th round, in addition to the regular verifications, P_y must also verify $K_x^n \stackrel{?}{=} H(MK_x)$. If the above passes verification, P_y confirms the validity of $NewK_x^0$. The series of new one-time signature keys $NewK_x^0, \dots, NewK_x^n$ can be used after the $(n + 2)$ th rounds.

5.5 Cryptanalysis of Li et al.'s Enhanced Protocol

Li et al.'s Enhanced protocol is based on Chan et al.'s EASES, they found the EASES protocol easily suffered from the replay attack, and try to add a unique game number, "gno#", to solve this problem. Unfortunately, the problem they found is not the main issue, thus attacker can crack the enhanced protocol in the similar way. The detailed steps are described as follows:

1. When P_x starts to send the first message, $\delta_x^1 = H(K_x^1 \parallel U_x^1 \parallel gno\#), \Delta x, K_x^0, gno\#$, to P_y , attacker P_z intercepts it and records K_x^0 .
2. When P_x sends the second message, $\delta_x^2 = H(K_x^2 \parallel U_x^2 \parallel gno\#), U_x^1, K_x^1, gno\#$, to P_y , P_z intercepts it and record K_x^1 . Then P_z forges a new message $\delta_x^1 = H(K_x^1 \parallel U_x^{1*} \parallel gno\#), \Delta x, K_x^0, gno\#$ and sends it to P_y , P_y verifies $\Delta x \stackrel{?}{=} D_{pk_x}(K_x^0 \parallel gno\#)$ and records K_x^0 to his memory if the equation holds.
3. When P_x sends the third message, $\delta_x^r = H(K_x^r \parallel U_x^r \parallel gno\#), U_x^{r-1}, K_x^{r-1}, gno\#$, $r = 3, \dots, n$, to P_y , P_z intercepts it and record K_x^{r-1} . Then P_z forges a new message $\delta_x^{r-1} = H(K_x^{r-1} \parallel U_x^{(r-1)*} \parallel gno\#), U_x^{(r-2)*}, K_x^{r-2}, gno\#$, and sends it to P_y , P_y verifies $K_x^{r-3} \stackrel{?}{=} H(K_x^{r-2})$ and $\delta_x^{r-2} \stackrel{?}{=} H(K_x^{r-2} \parallel U_x^{(r-2)*} \parallel gno\#)$ and record $U_x^{(r-2)*}$ and K_x^{r-2} to his memory if the two equations hold.
4. Finally, all update event U_x^1, \dots, U_x^{n-2} are replaced to $U_x^{1*}, \dots, U_x^{(n-2)*}$. Thus, Li et al.'s enhanced protocol still suffers from our attack.

5.6 Discussions

The aim that Chan et al.'s protocol is to reduce the computational cost. They believe that public-key cryptosystem require a large amount of computations. Then they propose the EASES and the dynamic EASES protocol. In these two protocols, only the first signature needs to be based on public-key cryptography, while others are based on the relationship between the hash-chain keys. As the above attack shows, they do not achieve their aim. Because attackers can easily tamper the hash value based on the public message, like the associated key. Further, the update event can be forged and the receiver can not find any questions.

Thus, in Chan et al.'s protocol, the public key is necessary. Li et al. notice the replay attack on Chan et al.'s protocol and make corresponding enhancement. However, the core problems are still existence in their enhanced protocol, so our attack does as well.

Acknowledgments The authors would like to thank the editors and anonymous reviewers for their valuable comments. This work is supported by the National Natural Science Foundation of China under Grant No. 60873235 and 60473099, the National Grand Fundamental Research 973 Program of China (Grant No. 2009CB320706), Scientific and Technological Developing Scheme of Jilin Province (20080318), and Program of New Century Excellent Talents in University (NCET-06-0300).

References

1. Dickey C, Zappala D, Lo V, Marr J (2004) Low latency and cheat-proof event ordering for peer-to-peer games. In: Proceedings of ACM international workshop on network and operating system support for digital audio and video, pp 134–139
2. Corman A, Douglas S, Schachte P, Teague V (2006) A secure event agreement (SEA) protocol for peer-to-peer games. In: The first international conference on availability, reliability and security
3. Chan M-C, Hu S-Y, Jiang J-R (2008) An efficient and secure event signature (EASES) protocol for peer-to-peer massively multiplayer online games. *Comput Netw* 52(9):1838–1854
4. Li C-T, Lee C-C, Wang L-J (2010) On the security enhancement of an efficient and secure event signature protocol for P2P MMOGs. *ICCSA 2010 LNCS* 6016:599–609

Chapter 6

Design of CAN Bus and Wireless Sensor Based Vehicle Tire Pressure Monitoring System

Peng He-huan, Zheng Hong-ping and Ma Ze-yun

Abstract The quality of tires, as key parts, makes significant impact on the safety of vehicles. Abnormal events to tire pressures, usually resulting in overheat or failure, may bring with inconvenience to driving as well as the useful lives and efficiency of tires. With on-going researches, a pressure monitoring device that gives warns in case of abnormal events, plays a very positive role in avoiding failures. With this regard, deep researches are made on can bus and wireless sensor based vehicle tire pressure monitoring systems.

Keywords Can bus · Wireless sensor · Tire pressure monitoring system (TPMS)

6.1 Introduction

The vehicular electronic controlling system (VECS) [1], of which the TPMS services as a key component [2], collects information including ignition switch status, speed, temperature in the process of driving, which improves the performance of TPMS significantly. A lot of methods for the improvement are available and for example, connecting the TPMS with high-speed local area communication networks of vehicles. As such, the overall performance of TPMS will be improved that not only the cost is reduced, but also the comfortableness is enhanced. In addition, the inherent identifying function of TPMS is helpful to the initial installation, ties positioning as well as location adjustment.

P. He-huan (✉) · Z. Hong-ping · M. Ze-yun
Tianmu College of Zhejiang A&F University, No. 252, Yijin Street,
Jincheng Community, Lin'an, Zhejiang Province, China
e-mail: peng6907hh@163.com

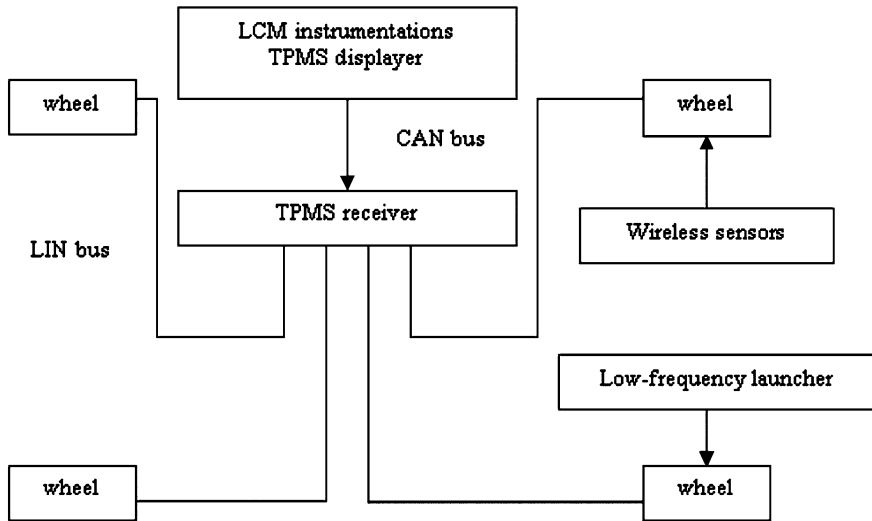


Fig. 6.1 Structure of TPMS

6.2 The Design of TPMS

The structure of direct TPMS is as shown in Fig. 6.1. Components involved consist of wireless sensor, TPMS displays indicating tires information, TPMS receiver, low-frequency trigger, CAN bus and LIN bus.

With wireless sensor installed in wheels, the pressure and stability of air in tires need to be effectively and periodically tested, usually once every 5 s, with the results sent to TPMS receivers through wireless channel. The TPMSs are usually placed in the cab and are mainly for the receiving, processing and sending information through CAN bus to ICM instrumentations for further processing and exchanging information with bus nodes as BCM vehicular controlling modules and others. Low-frequency triggers are usually placed in the inner side of vehicle baffles, connected with LIN bus through TPMS receivers and are able to search the wireless sensor quickly in case of abnormal events, to ensure the precision of information being received by TPMS receivers, and further to identify the pressure and temperature of air in tires and for drivers to take measures by referring indications from TPMS devices.

The CAN bus communication controllers possess the functions of both physical and data link layers and are able to perform adjustments on communication date framings, such as zero insertion/deletion, data block coding, cyclic redundancy checking and priority determining.

LIN bus is mainly for refining vehicular network performance. In case advanced communication CAN bus is not necessary, such as controlling doors, windows or rearview mirror, smart sensors and brakes are able to support

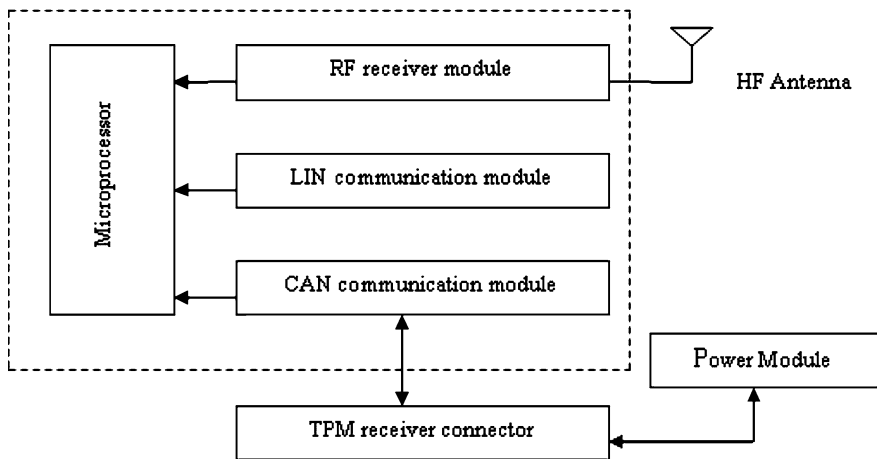


Fig. 6.2 Frame of receiver circuit

communication and thus are helpful for controlling low-frequency triggers and also for controlling cost. The transport speed of LIN bus reaches 20 bit/s, in all 16 nodes can be supported in one channel bus circular and the length of bus cable reaches 40 m.

6.3 Design of System

6.3.1 Design of Hardware

6.3.1.1 TPMS Receivers

TPMS receives are usually placed under the instrumentations board [3], as shown in Fig. 6.2 and consist of: RF receiver module, CAN communication module, microprocessor, CAN controller, LIN communication module, power module. Information received by HF Antenna is mainly sent by wireless sensors and need to be processed by RF receiver module before sending to microprocessor; such processing includes: filtering, amplifying, mixing, demodulating and others. Microprocessors are primarily for performing works on processed instructions, including send data package to instrumentations through CAN communication module, received information useful in bus and send low-frequency signals to wireless sensors with support from low-frequency triggers controlled by LIN communication module. The functions of battery module, including filtering, anti-reverse-pressure, level shifting, are for vehicular batteries. As connectors specifically designed for vehicles, receiver connectors support the connections

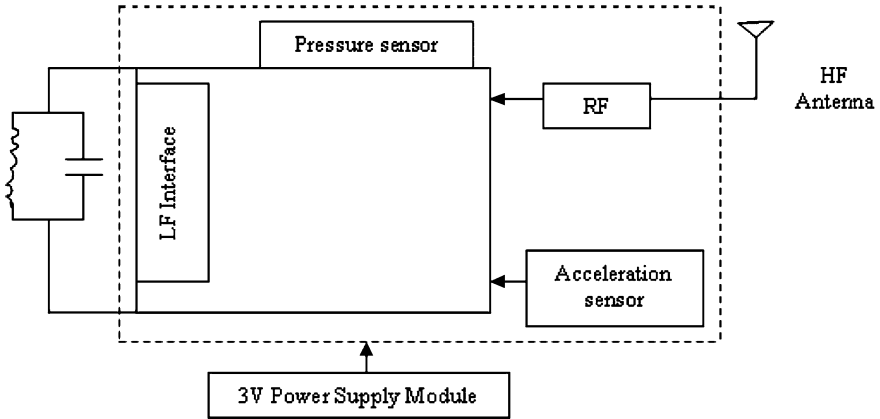


Fig. 6.3 Frame of wireless sensor circuit

between the power, ground lines, CAN harnesses and LIN harnesses of TPMS receivers, to ensure effective communications between nodes of TPMS receivers and vehicular CAN bus network.

6.3.1.2 Wireless Sensors

The way valve wireless sensor are most popular and are usually placed in the bottom of rim valve mouth stem. Wireless sensors are mainly used in controlling pressure and temperature of air in tires under monitory and send information to TPMS receivers through wireless control. A wireless sensor consist of: sensors, signal conditioning, low-frequency interfaces, RF circuits, battery.

Bridge electronic air pressure sensor is able to sense the pressure in tires and then transfer the pressure signals into electronic data and send useful information to date processing center for further works, including amplifying, compensation, etc. The low-frequency module consists of mainly low-frequency antennas and low-frequency interfaces, is able to low-frequency signals at 125 kHz as received from low-frequency triggers and then perform operations in accordance with signals from triggers and data center. In case of abnormal situations, wireless sensors are able to send signals to TPMS receivers through high-frequency antennas, in which case the central frequency of signals is 315 MHz (Fig. 6.3).

6.3.1.3 Low-Frequency Triggers

Low-frequency triggers are usually placed in the inner side of vehicle baffles, consist of LIN communication module, low-frequency driver and launch, micro-processors and power module, as shown in Fig. 6.4.

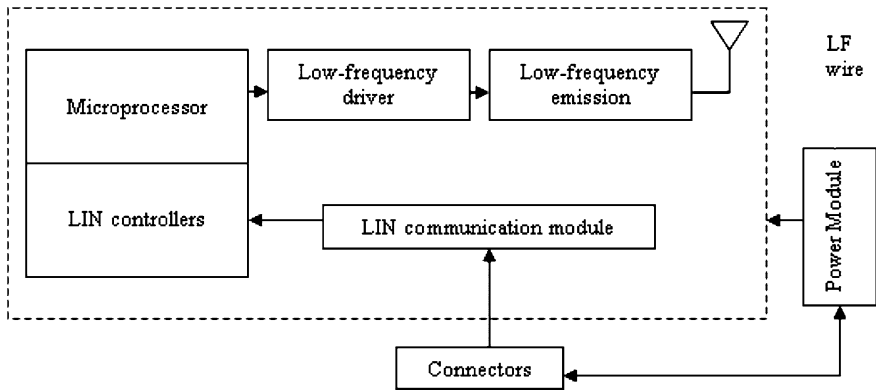


Fig. 6.4 Low-frequency trigger circuit

After processing information in TPMS receivers with LIN bus, LIN communication module is able to send signals to wireless sensors through low-frequency drivers and low-frequency launches, to improve multiple communication between TPMS and wireless sensors. Therefore, TPMS system is able to not only trigger each individual wireless sensors, but also test identification of each wireless sensors. Such function is helpful to control tires positions and also to set relevant factors of wireless sensors. The power module is for level shifting and filtering for battery, and provides power to certain devices, such as microprocessor and low-frequency drivers.

6.3.1.4 TPMS Displayers

After collecting and processing, TPMS receivers is able to send high-frequency signals from wireless sensors to TPMS displayers in ICM instrumentations module through CAN bus for drivers to refer. Information to be displayed includes air pressure warning, system problems cautious and others.

6.3.2 Design of System Software

TPMS inside the high-speed CAN bus communication network consists of: ware-communication and wireless communication controlling module, each component is significantly for the ongoing TPMS system and will result in problems in system operation in case of absence.

Table 6.1 Data frame send by TPMS receivers to ICM instrumentations

Name of signals	Lowest position	Name of signals	Lowest position
TPMS problem	0	Rapid leakage	10
Auto-identification of tires position	1	High pressure	12
Tires position	2	High temperature	14
Reserve	4	Air pressure in tires	16
Low air pressure	8	Temperature in tires	24

Note The identifier for data frame ID is 32Ch; the speed for bus transportation is 500 kb/s; the length is 4 bit; the signal type is periodical; the data refreshing frequency is second

6.3.2.1 Software Controlling Policy and Communication Protocol for TPMS Receivers

After making internal adjustments, TPMS receivers are able to identify tires and communication process to normal operation of the system. The automatic identification of tires positions is achieved through functioning of TPMS receivers with low-frequency triggers, which amplifying wireless signals to ensure the ID information and pressure date be collected. As a result, the TPMS receiver is able to perform practices on factors and position of tires. The network communication includes network management frame, data processing and problems probation, etc.

The nodes in CAN bus electronic control system are operated under standard communication protocol, their data frame are operated in a standard way and the format, communication speed as well as refreshing frequency are processed. Data frame send by TPMS receivers to ICM instrumentations are shown in Table 6.1.

6.3.2.2 Software Controlling Policy and Communication Protocol for Wireless Sensors

The software for wireless sensors usually consists of: initialization, interrupt, parameter testing, data demodulation and others. After data initialization, the wireless sensors entered power-saving module and the data testing represents basically periodical waking every several seconds. The wireless sensors in operating status performs well that they are able to test the air pressure, temperature and voltage, process relevant information, gives warnings through wireless channels if abnormal situation relevant to pressure or other factors is figured out from data, and then return to power-saving module. The abnormal situations relevant to pressure are usually low pressure or rapid leakage. Low pressure represents the status that the real air pressure is $\frac{1}{4}$ more lower than designed cold air pressure level. Rapid leakage represents the status that the pressure is decreasing at a speed extending designed range.

Table 6.2 TPMS receiver node Bus statistics

Signal names	Receiving/sending status of IPM nodes	ID	Data frame number	Standard deviation	Time diff of data refreshing/ms	Remote frame number	Error frame number
TPMS	Tx	32Ch	450	5.325	103	0	0
CLM	Rx	320 h	446	23.2	103	0	0
BCM	Rx	45Bh	224	25.5	2×103	0	0
ICM	Rx	42Ah	9×103	0	50	0	0

6.4 Testing

The research in this paper involves designs in many aspects and two phases of tests are performed on TPMS: 1) test with Vector CANalyzer bus analyzer, being majorly 400 h bench test on auto-identifiable TPMS; 2) 104 km car road test of TPMS in M-Car domestically made. In the analysis, the CAN bus communication and wireless data transportation are researched.

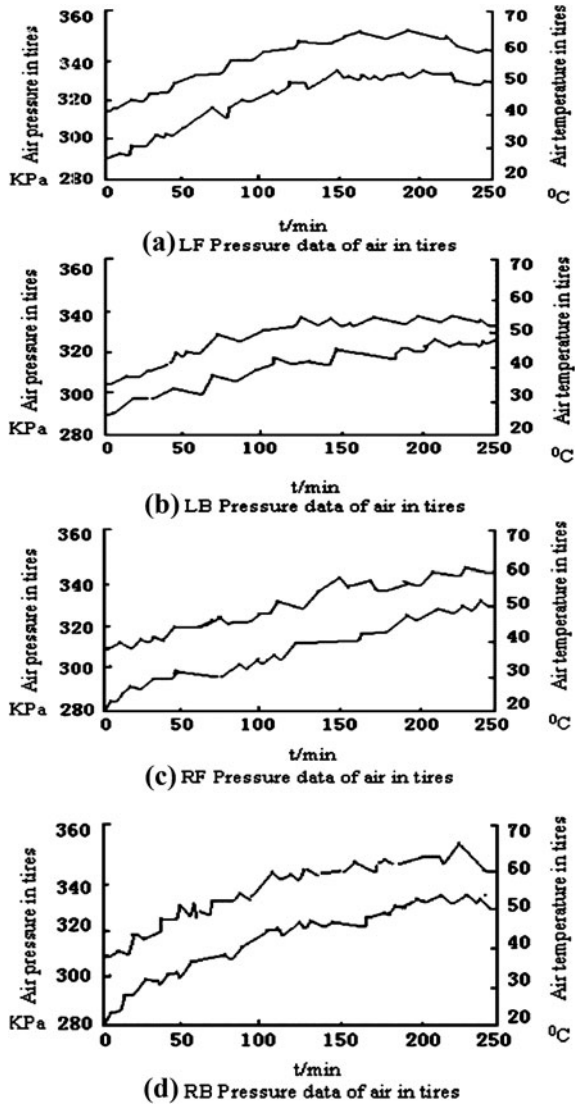
6.4.1 CAN Bus Communication Test

Bus analyzers containing TPMS nodes are connected with CAN communication network bench, and then data process adjustment is performed through network platform. As the final result illustrated in Table 6.2, that after periodical refreshing, frequency of data frame used by TPMS is even more accurate and no remote frame or error frame is identified, evidencing the fact that TPMS nodes improve the accuracy of information transportation in network.

6.4.2 Test of Wireless Transportation

Wireless sensors are usually placed in wheels and after connecting to CAN network, TPMS special data testing system will test the performances of vehicles. Figure 6.5a–d illustrate the movement of 4 h air pressure and temperature changes status in left-front, right-front, left-back and right-back wheel respectively. TPMS car load test is on high-speed road and the time necessary for wireless sensor to send a frame of date is 5 min, which means that if the performances of tires are satisfying, the speed of vehicle will reach 60–180 min/h after sending a frame of data. During the test, the out-door temperature is 25–30°C, the air pressures are 0.2, 0.2, 0.21 and 0.21 PMa in left-front, right-front, left-back and right back wheel respectively. Figure 6.5 illustrates that both the temperature and pressure for the car tested reached 46 frames and over 95% of information is received and air pressure and temperature changed significantly; as such, the inner situation of the car is illustrated.

Fig. 6.5 Data Curve of PTMS Car Road Test



6.5 Conclusion

After installation of TPMS in vehicular bus communication network, if no abnormal events occurs with TPMS nodes and frames are refreshed in standard periods, the overall operation of the system will be stable and actual performance will be satisfactory. Will application of the system in driving, responses can be made to air pressure, temperature and other situations to ensure correctness of factors.

References

1. Zang H, Tian C, Zhao B (2008) The design of tire pressure monitor system [J]. J Beijing Inst Technol 10:871–873
2. Guibao T, Li P (2008) Design of pressure-sensor based tire pressure monitoring system [J]. J Chongqing Univ 01:023–029
3. Zheng Q, Jianguo H, Lina Y, Zhixian L, Li T (2009) Research and application of TPMS [J]. Microprocessor 02:153–158

Chapter 7

Analysis and Test of the Exposure Synchronization of the Multi-Sensor Aerial Photogrammetric Camera System

Zhuo Shi, Li Yingcheng and Qu Lei

Abstract Detailed analysis of exposure synchronization is made aiming at high precision matching and combining the images captured by multi sensor camera system. Exposure mechanism and motion of shutter blades or curtains become the main source of asynchronic error among individual cameras. Two methods are used to test the exposure asynchronization according to both between lens shutter and focal plane shutter configurations. Analysis of the results shows the consistency of theoretical analysis and test measurements.

Keywords Exposure synchronization · Multi-sensor aero photogrammetric camera · Shutter configuration · Ground sample distance

7.1 Introduction

Multi-sensor aerial photogrammetric camera introduces larger format, longer base line, less flight course, more efficiency and less after-work under the same task situation, making it possible to reduce task cost [1]. Such kind of camera is included into resource, grid line supervision, land mapping, city planning and other applications widely.

Z. Shi (✉) · L. Yingcheng
Chinese Academy of Surveying and Mapping, 28 Lianhuachi West Road, 100830,
Beijing, China
e-mail: Zhuoshi9988@163.com

L. Yingcheng · Q. Lei
China TopRS Technology Co.Ltd, 16 Bei Tai Ping Road, 100039, Beijing, China

Digital sensor could not achieve the traditional film aero photogrammetric cameras' format due to limitation on semiconductor technics and sensor construction [2]. Multiple sensors are integrated to form a large format [3]. Each sensor and corresponding lens form an imaging system, which records images and exposure information separately. Images are matched and combined after to form large format digital images. Two factors influence the combining process: the relative outer orientation elements and the exposure asynchronization of each camera system. Stability of relative outer orientation elements is guaranteed by precise design and producing, thus the exposure asynchronization becomes the key factor affecting the image combination precision. Synchronic exposing control is one of the most important technologies in multi-sensor aero photogrammetric system integration, for the asynchronous error effects the whole image acquisition and processing progress. Asynchronous error must be reduced in an acceptable range to guarantee the combining precision.

Considering digital cameras' construction, exposure mechanism and signal transmitting, analysis and tests are made for both sensitization aspect and shutter motion aspect. Testing data is analysed to validate stability.

7.2 Analysis of Exposure Synchronization of Multi-Sensor System

Camera exposes at a specific position on signal sent by Flight Management System (FMS). Instead of sending different exposure signal for each camera system, the FMS sends one single exposure signal. Images captured by each camera will not accord the designed relative position if cameras do not expose synchronically. The combination process highly relies on steady relative position. The asynchronous exposure leads to harder combination processing, or even wrong results. Figure 7.1 shows the relative position of four images under synchronic exposure (a) and asynchronous exposure (b).

7.2.1 Exposure Process and Analysis of Asynchronous Error Source

Exposure process can be idealized as a step response process [4]. The shutter opens after receiving the exposure signal and allows the light to go through. After the required shutter speed has been reached, the shutter closes to block the light. The response time for the step process equals to the required shutter speed (Fig. 7. 2).

Exposure process can be divided into the electronically controlled phase and the mechanical phase. In the electronically controlled phase, the time difference for synchronization is small since the exposure signal is an impulse, which is a low voltage, low frequency signal, and transmits a short distance. The mechanical

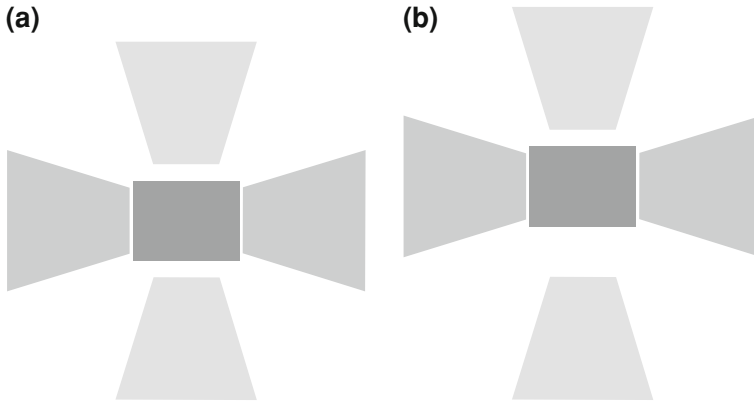
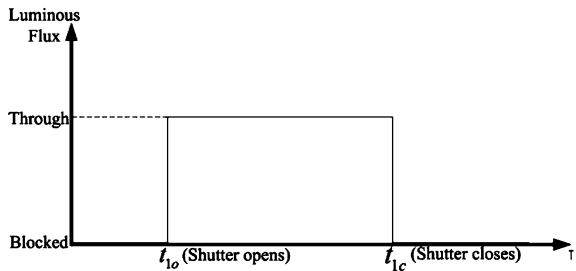


Fig. 7.1 Relative position of four images. **a** Synchronic exposure, **b** asynchronic exposure

Fig. 7.2 Step response of the shutter



phrase begins when the shutter receives the electronic signal. The driver unit drives the blades to expand to a required period and then closes them [5]. The various types of driver unit among different products result in uncontrollability and randomness in mechanical phrase.

The source of error for the exposure asynchronization mainly comes from the variety of the blades and driver units. The error exhibits a certain degree of randomness and can be analyzed by using a large amount of observed data (Fig. 7.2).

7.2.2 Calculation of the Maximum Exposure Synchronization Difference

The actual shutter speed $\Delta T_i (i = 1, 2, \dots, 4)$ is not exactly the same in practice for a multi-sensor camera system [6, 7]. However, the difference can be neglected since it has little impacts on the synchronization. Analysis aims at the moment

when the shutter closes. Δt is defined as the time difference between the moment when the first camera finishes the exposure process and the moment when the last camera finishes the exposure process. The allowed maximum exposure asynchronization is calculated under the following conditions: (1) camera and plane orientations are ignored; (2) plane moving at a constant speed without the impacts of wind; (3) the maximum displacement caused by the exposure asynchronization should be within one Ground Sampling Distance (GSD).

$$\Delta L_{\max} = \Delta t_{\max} \times v_G \quad (7.1)$$

$$\Delta t_{\max} \times v_G \leq GSD \quad (7.2)$$

where: v_G stands for the ground speed of the plane at the moment of exposure
 Δt_{\max} stands for the allowed maximum error in exposure
 ΔL_{\max} stands for the allowed maximum displacement of images
GSD stands for the ground sampling distance

Taking the ground speed for middle and large size aerial photography plane as 250 km/h and GSD as 10 cm, the allowed maximum exposure difference among the cameras is calculated.

$$\Delta t_{\max} \leq \frac{GSD}{v_G} = \frac{10 \text{ cm}}{250 \text{ km/h}} \approx 1.44 \text{ ms} \quad (7.3)$$

Factors such as the wind speed at the moment of exposure, the instantaneous change of the orientation of plane and adjustment of the camera orientation during flight, will affect the value of ΔL . Therefore, the value of Δt_{\max} should be adjusted to satisfy the accuracy and quality required by the data process considering the above factors.

7.3 Test of Exposure Asynchronization

The following two test methods are used [8]:

1. Taking photographs of a timer with high accuracy

Use two cameras to take photographs of the high accurate timer simultaneously for several times. The time difference showed between the two images is the exposure time difference in synchronization. Results are then recorded and analyzed. This method is easy to operate and does not require extra apparatus. Also it can be applied to various types of shutters. The accuracy solely depends on the accuracy of the timer.

2. Using photodiode circuit to capture the moment when shutters are fully opened

Several lenses are placed aiming at the same light source as a group. Each photodiode is placed at the position where shutter is fully opened. Protections are

Fig. 7.3 Testing images from Cam1 and Cam2



made to avoid stray light. Exposure signals are sent simultaneously to each shutters. And the signals from the photodiode are captured by the oscilloscope. The time difference of synchronization is the difference between the rising edges of each signals. The camera which finishes first is treated as the reference, and time difference of synchronization of other cameras is calculated. Statistical methods are employed to analyze all the results. This method requires constructing photodiode circuit to produce a more accurate result and is only measurable for between-the-lens shutters (Fig. 7.3).

7.3.1 Taking Photographs of a Timer with High Accuracy

This method is used for the full frame camera with a focal plane curtain shutter since the movement of shutters cannot be tested through physical tools due to camera structure. The tested camera model is Canon 5D Mark II, the target is a timer with a minimum display accuracy of 0.001 s. The procedure is as follows:

1. Change of the timer is modeled as a moving target, and photographs of the timer are taken simultaneously from each camera .
2. Compare the photographs from each camera and the difference between them is considered as the time difference in synchronization.
3. Record the time difference interval.

Parts of the test results are presented in the following.

Among all the 100 tested photographs, 92% indicate time difference less than 1 ms; 8% show very different timing between the two photographs in one test. This phenomenon results from the fact that the timer was actually refreshing at the

Fig. 7.4 Parts of the test results

	A	B	C	D
1	No.	CAM1	CAM2	Time Difference (ms)
16	15	1:42.969	1:42.969	0
17	16	1:44.749	1:44.749	0
18	17	1:47.146	1:47.146	0
19	18	1:49.187	1:49.187	0
20	19	1:51.148	1:51.148	0
21	20	1:53.235	1:53.235	0
22	21	1:55.481	1:55.481	0
23	22	1:57.570	1:57.570	0
24	23	1:59.731	1:59.731	0
25	24	2:02.072	2:02.073	1
26	25	2:04.319	2:04.319	0
27	26	2:06.729	2:06.729	0
28	27	2:08.857	2:08.857	0

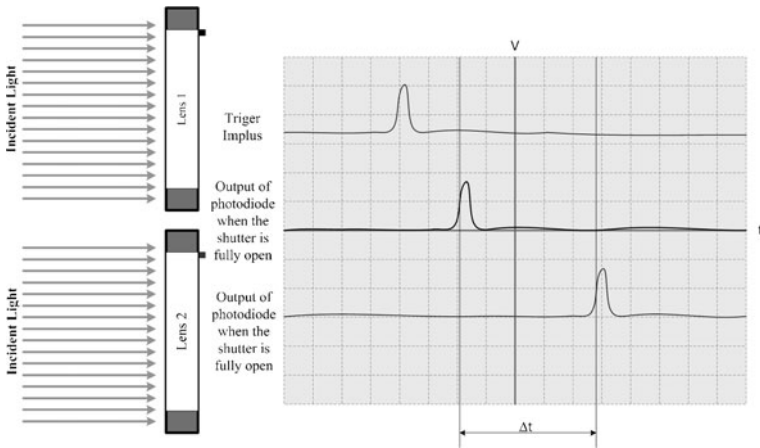


Fig. 7.5 Sketch map of the test

moment the photo was taken. Such results should be eliminated in the result analyzing. The usable accuracy should be no larger than 10 ms due to the refresh period of the timer.

This method relies highly on the accuracy of the timer and the refresh frequency of the display facility, so the accuracy of the testing result is limited. Further study will be done based on 7-segment LED monitor timer which has a refreshing period less than 0.1 ms. However, this method does not require the modification of the camera structure and is easy to implement. The result is directly readable and the method can be applied on various types of cameras, especially for full frame cameras with focal plane curtain shutter in low cost unmanned aerial vehicle applications (Fig. 7.4).

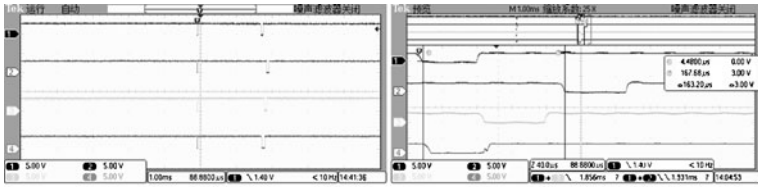


Fig. 7.6 Negative impulses captured by the oscilloscope

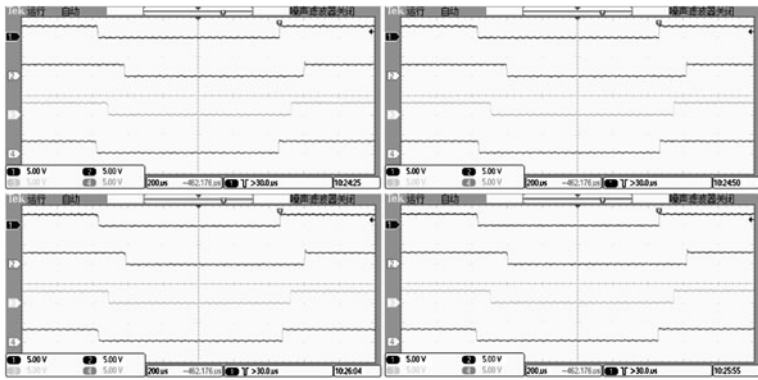


Fig. 7.7 Impulses captured by the oscilloscope under continuous exposure mode

7.3.2 Using Photodiode Circuit to Capture the Moment when Shutters are Fully Opened

Place lenses under the same light source and adjust the aperture to its maximum. Place photodiodes at the position where the shutter is fully opened, and keep the photodiode away from unexpected light. Apply the exposure signals to each shutter at the same time. The exposure signal is shown as the upper wave in Fig. 7.5. Capture and record the outputs from each photodiode with the oscilloscope. The difference between the rising edges of the two signals is the time difference of asynchronization. The targets are four Rollei Electronic Shutters in the Schneider Apo Digital lenses.

Oscilloscope monitor display is shown below for all four camera channels. Two negative impulses are captured, and the falling edge of the second negative impulse indicates when the shutter is 90% open. The timing period between the first and the last cameras' second falling edge is the exposure asynchronization (Fig. 7.6).

Tests are implemented in manual exposure mode for five times, producing the following exposure asynchronization: 122.40, 149.00, 154.60, 157.60, 163.20 μs. Tests are implemented in continuous exposure mode for 50 times. And results

	A	B	C	D
1	NO.	First exposure(us)	Last exposure(us)	Time difference(us)
2	11	1011.1	847.1	164.0
3	12	1034.2	866.2	168.0
4	13	1030.0	861.0	169.0
5	14	64.5	105.9	170.4
6	15	4.4	167.6	163.2
7	16	1766.8	1922.2	155.6
8	17	8.6	140.2	148.8

Fig. 7.8 Parts of observed time difference

show identity with above. The maximum timing difference recorded is less than 200 μ s. Some of the results are shown as follows (Figs. 7.7, 7.8).

7.4 Conclusion

Causation of the exposure asynchronization and its effects to after-process are analyzed based on the manufacturing of a certain large format combining-sensor system. Two methods are introduced to measure the timing difference of multi-exposure. Results show identity between the actual exposure asynchronization of the tested system and the theoretical analysis.

Method based on photodiode detecting can be introduced as a highly required test for such systems before shipping. Method such as using high accuracy timer can be applied as periodical test without disassembly. Both before-shipping and regular inspection and test are recommended to be involved into the industry standard and/or national standard to ensure the consistency and applicability of the images produced by such multi-sensor aero photogrammetric camera systems.

References

1. Deren LI, Wang S, Zhou Y (2008) Introduction to photogrammetry and remote sensing. Surveying and Mapping Publication, Beijing
2. Qian Y (2006) Principle and use of modern camera. Zhejiang Photography Publication, Zhejiang
3. Du Y, Liu B, Cao J, Zhang BH, Hu BL, Tang Y, Wang H, Lu M (2008) Control system of aviation digital camera based on P89C668 SCM. Comput Eng Des 29(1):129–141
4. Zhang Z (2004) Aspects on aerial digital cameras. Eng Surv Mapp 13(4):1–5
5. Talley BB (1938) Multiple lens aerial cameras, chapter IV in aerial and terrestrial photogrammetry. Pitman Publishing Corporation, Chicago, pp 91–116
6. D Light (2004) A basis for estimating digital camera parameters. Photogramm Eng Remote Sens 70:297–300
7. Jacobsen K (1988) Handling of panoramic and extreme high oblique photographs in analytical plotters, ISPRS congress Kyoto. Int Arch PhRS XXVII B2, pp 232–237
8. Schroter RW (2008) The challenge of process and management optimization for photogrammetric mapping. In: Proceedings of the colloquium for the 60th birthday of Prof. M. Ehlers, Osnabruck

Chapter 8

Design and Application of Custom RS485 Communication Protocol

Lei Zhou, Hu Liu and Quanyin Zhu

Abstract As RS485 network has the character of simple construction, lower cost and easy extension, it has been widely applied to many application systems based on Radio Frequency Identification (RFID). In RS485 network the mainframe can access many readers. Aiming at the demand of a RFID application system in which the mainframe needs transmitting information to display terminal, this paper proposes a solution. Through connecting display terminal and readers into RS485 network, defining custom communication protocols between the mainframe and display terminal, the mainframe can access display terminal like the other readers. The implementation results show that the solution has achieved the anticipated target and run stably and reliably.

Keywords RFID · RS485 network · Custom communication protocol · Display terminal

8.1 Introduction

Radio Frequency Identification (RFID) is a non-contact automatic identification technology, and the identify process does not need manual intervention. Its basic principle is utilizing the space (electromagnetic or inductance coupling)

L. Zhou (✉) · H. Liu · Q. Zhu
Faculty of Computer Engineering, Huaiyin Institute of Technology, Huaian, China
e-mail: zl_gxy@163.com

H. Liu
e-mail: jsliuhu@163.com

Q. Zhu
e-mail: hyitzqy@gmail.com

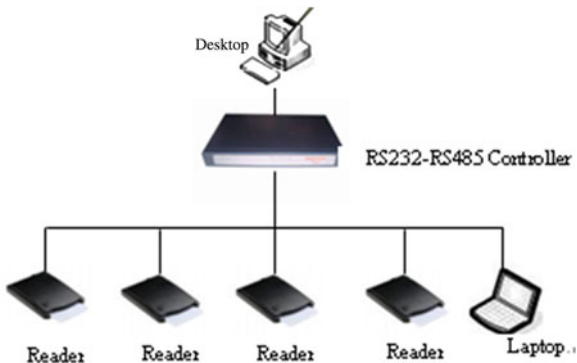
transmission characteristics of radio frequency signals to realize the automatically identification of objects [1–3]. Now RFID technology has been applied to logistic, access control and manufacture area etc. In a typical RFID application system, there are usually many readers to read RF tags' information. The readers need communicating with the mainframe, such as transmitting tag messages or response information to the mainframe, receiving an instruction from the mainframe. In the case that the mainframe can communicate with many readers at the same time, RS485 network can be adopted. Because RS485 network has advantages such as simple construction, lower cost and easy extension, in addition, an increasing number of equipments provide 485 ports, so RS485 network is widely used in many fields now.

8.2 System Background

We designed a meeting attendance system for a company. In order to identify conventioners automatically, we assigned a RF tag for each conventioner, and placed a reader in each entrance of meeting hall. When conventioners entered, readers could read their tags and obtained attendance information transmitting to the mainframe. There were many entrances in meeting hall, and the mainframe needed to access many readers, so we adopted RS485 network. In this communication style, the mainframe connects to RS485 network via RS232/RS485 converter, and each RS485 network can connect with many readers. According to the user demands, the meeting president hoped to obtain real-time data during the meeting attendance. As the mainframe running the attendance system was in the monitor room which was far away from the rostrum, it would be very inconvenient for the system runner to phone the meeting president about attendance information always. To solve this question, meeting organizers determined to display the latest attendance data in a computer which placed in the rostrum. We can call the computer as display terminal for convenient. How to achieve communication between the mainframe and display terminal? As we must implement the function that the mainframe can access many readers with programming, if we write communication procedure especially for the mainframe and display terminal in addition, it may bring many questions. For example, the procedure may have influences on access sequence between the mainframe and the readers and bring about program errors, on the other hand, programming especially for display terminal increases the complexity of program control. After careful consideration, we decided to take display terminal as a reader and connect it to RS485 network like a reader, the mainframe can access display terminal as same as the readers. The connection structure is shown in Fig. 8.1.

In this communication style, the laptop is used as display terminal and system takes display terminal as a reader. Because RS485 network works in bus communication model, only one reader can communicate with the mainframe at the same time. In RS485 network there are many readers communicating with the

Fig. 8.1 The system connection



mainframe and conventioners can trigger tags at the same time. In order to avoid access conflicts caused by the operations that several readers transmit information to the mainframe simultaneously, we must adopt active polling programming mode. In this model the mainframe sends instruction to a reader at regular intervals, obtains corresponding return data to deal with, and then sends an instruction to another reader analogizes in turn. As display terminal is taken as a reader, the mainframe can also send attendance information to it in polling programming. When display terminal receives attendance information from the mainframe, it can process data and display result in the monitor. As the polling interval is very short, display terminal can always get the latest attendance information.

From the above, we can see that the scheme which takes display terminal as a reader can simplify process procedures, but in order to accomplish the function of unified access control, we must define communication protocols between the mainframe and display terminal, and they must be consistent with the reader protocols.

8.3 Custom Communication Protocols Definition

As we know, custom communication protocols between display terminal and the mainframe must conform to the rules of reader protocols, so we must study reader protocols firstly. According to the user demands, conventioners need not trigger tags over long distance, low frequency reader can meet the system demand as well as have lower cost, so we adopted 125 kHz WM-02H readers which were produced by Beijing Wan-Mei corp [4].

WM-02H reader communication protocols provide many operation functions such as read-record, set-address and communication test instruction, etc., its communication protocol format is shown as follows [5].

|start flag | slave machine address | information length | instruction code and parameter | slave machine return data | verify bytel.

Table 8.1 Custom communication instructions

Instruction code	Parameter	Function
A4	Reader address	Set display terminal address
A6	Test value	Communication test
AB	Attendance data	Send attendance data to display terminal

Next the function of each instruction will be described in detail

Next every part meaning of the protocol will be explained. Start flag is two bytes length, it determines the command is from the mainframe or a slave machine, and the control instruction from the mainframe has start flag AAH FFH, while the return data from a slave machine has start flag BBH FFH; slave machine address has one byte, ranges from 00H to FEH; information length indicates total bytes of instruction and parameters, but start flag, slave machine address and verify byte are not included in; instruction code and parameter are the content of the order; verify byte is one byte length which mainly is used to check the instruction correctness. If the instruction is transmitted correctly, verify byte should be equal to the result executing XOR operator on all other bytes.

According to the format of WM-02H reader communication protocol, we can define the protocols between display terminal and the mainframe, and the custom protocol instructions are shown in Table 8.1.

8.3.1 Set Address Instruction (A4)

When the mainframe sends an instruction to display terminal, it must specify the address, so we must define an instruction to set display terminal address. There's one point which needs attention that slave machine address must be FFH in the instruction.

According to the protocol format, if we want to set display terminal address 08H, the instruction should be AA FF FF 02 A4 08 04, then the meaning of each byte in order will be explained. Start flag AA FF indicates it's an instruction from the mainframe, next FF is the broadcast address, 02 indicates the instruction length is two bytes, A4 is the instruction code, 08 is the display terminal address, 04 is verify byte. If the instruction is executed successfully, display terminal will return response information BB FF 08 01 08 45 to the mainframe. In this response message, BB FF indicates this data was sent by display terminal, 08 is display terminal address, 01 indicates the instruction length, 08 is the address been set, 45 is verify byte. If the address sets error, display terminal return information BB FF FF 01 33 89, 33 indicates the command has error, and the address can not be set successfully.

8.3.2 Communication Test Instruction (A6)

Before the mainframe begins to communicate with each reader it must send communication test command to check if the communication line is normal. In the same way the mainframe needs to do communication test to display terminal. When display terminal receives communication test instruction, it checks instruction content firstly. If the instruction is correct, it returns response message to the mainframe and the mainframe checks the communication line normal or not according to return information; if the instruction is error, display terminal return command code 33H which shows communication test has problems.

According to the protocol format, if the mainframe sends communication test instruction to display terminal which has address 08H, assumption test value is 77 88, then the instruction should be AA FF 08 03 A6 77 88 07. If display terminal accepts correctly it returns information BB FF 08 02 77 88 B1 to the mainframe, otherwise it returns information BB FF 08 01 33 8E. Then the mainframe analyzes the return information, if the return data is 33H, that shows the communication line is not normal, but if the test data returned is as same as the data sent before, it shows the communication line is ok, and communication test instruction has executed successfully.

8.3.3 Send Information Instruction (AB)

In order to display real-time attendance data in display terminal, we need to define send information instruction. During active polling programming, the mainframe can send meeting attendance data such as total number and attendance number, display terminal receives the data and calculates attendance number, absence number and attendance rate so as to display them. As the system is mainly for small and medium-sized conference, total conventioners number is not more than several thousands, so the instruction parameter parts which store total number and attendance number are assigned both two bytes.

According to the send information protocol format, if the mainframe sends the instruction to display terminal, assumption its address is 08H, meeting total number is 425(01A9H) and attendance number is 385(0181), then the instruction should be AA FF 08 05 AB 01 A9 01 81 DB. Display terminal receives the instruction and obtains attendance data to display on the screen. As processing speed of different computers has difference, in addition, there is much difference between display terminal and readers, it is very hard to set suitable delay interval while the mainframe waiting for response information from display terminal. In order to avoid conflict caused by the unsuitable interval, the instruction was designed not to return information, that means display terminal only receives the instruction and do not return any information to the mainframe. Because active polling interval is very short, even if errors occur in transmitting process, display

terminal has not accepted correct message, there is no influence to the next message to be displayed correctly.

8.4 The Mainframe System Programming

As display terminal is taken as a reader in RS485 network, the mainframe can access it with the other readers. The mainframe takes active polling programming mode to access each reader including display terminal, but we must pay attention to the difference of access process between readers and display terminal. The mainframe sends read-record instruction to a reader in order to obtain tag information which has stored in the reader, and then it will wait for response information from the reader. While the mainframe executes send-information instruction and sends attendance data to display terminal, it need not wait for return data. The following is the processing procedure.

- Step 1: User selects display terminal and readers which have connected into the RS485 network, and system saves their addresses.
- Step 2: System sends communication-test instruction to each reader in order at a specified interval and waits for return information. If return data from a reader is correct, it means the communication line between the reader and the mainframe is normal, then system sends instruction to the next reader until all readers have been visited. If response from a reader is error, there are at most three times for the reader to receive instruction and return information to the mainframe. If a communication line is not normal really, the instruction has failed to execute three times, system will pop up a message box and end the program. If communication test for all readers are passed, the program will go to the next step.
- Step 3: After communication test, system begins to access every reader in turn at a certain interval. According to slave machine address system can know operation object is a reader or display terminal. If the address belongs to a reader, system will send read-record instruction and delay a certain interval in order to receive response information of the reader; while if it is display terminal address, system will execute send-information instruction which has no return information, and then send instruction to the next reader indirectly. If the reader visited is the last one, system will repeat above procedure from the first reader until user ends the program.

There is one point to pay special attention that how to set a suitable polling interval. Polling interval is affected by many factors such as serial communication baud rate, instruction length, return information length, etc. For guaranteeing system reliability, only take all factors into consideration and after lots of test can we set the value of polling interval. If the interval is set too big, it will reduce

efficiency of system, while too short value will bring communication conflict and cause system run-time error.

8.5 Display Terminal Programming

Display terminal module is to accept instruction from the mainframe and obtain attendance information to display. As it only receives instruction passively, need not send information to the mainframe, so we can use passive programming model utilizing serial port communication control.

In Visual Basic integrated development environment we can use Mscomm control to implement serial port communication. Before using MSComm control there are many properties to be set. Supposing there is a MSComm control named mscomm1, we may complete the property setting of mscomm1 control as follows.

```
Private Sub Form_Load ()
    'Set serial port number
    MSComm1.CommPort = 1
    'Set serial port communication baud rate
    MSComm1.Settings = ''9600,N,8,1''
    'Obtain all bytes of buffer each time
    MSComm1.InputLen = 0
    MSComm1.InputMode = comInputModeBinary
    'Set response byte length when System triggers serial
port event
    MSComm1.RThreshold = 1
    'Set receive buffer size
    MSComm1.InBufferSize = 1024
    'Open the communication port
    MSComm1.PortOpen = True
End Sub
```

When serial port has received one byte system will trigger MSComm control event automatically. Due to RS485 bus uses broadcast communication model, every instruction from the mainframe will be sent to all readers, including display terminal. When a reader receives an instruction, it judges firstly the instruction receive address, only the reader owning the same address will accept the instruction to deal with, so when display terminal receives a instruction through serial port, it also must check the instruction receive address in event procedure. The serial port event procedure may be written as follows.

```
'Serial port event procedure
Private Sub MSComm1_OnComm()
'Deal with serial port event
Select Case MSComm1.CommEvent
```

```

Case comEvReceive ' Received data
  `Delay a certain interval to obtain enough data
    Timer1.Enabled = True
    Do While Timer1.Enabled
      DoEvents
    Loop
  `Read data from serial port buffer
    buffer = MSComm1.Input
  `Check the instruction length which has received
  .....
  `Judge the instruction correctness
  .....
  `Judge the instruction receive address
  .....
  `Judge the instruction type
  .....
  `Deal with the instruction according to its type, if it
  needs return information to the mainframe, then send
  response messages which conform to instruction format
  .....
End Sub

```

Next we will take send-information instruction for example and describe the processing procedure of display terminal.

- Step 1: Once serial port buffer of display terminal receives bytes, control event will be triggered automatically. In event processing procedure, system delays a certain interval and obtains all bytes from the buffer, then stores them in an array.
- Step 2: System looks for the position of instruction start byte which begins with AAFF, and executes XOR operator on all bytes to check the instruction is right or not.
- Step 3: If verify operation is passed, system will compare instruction receive address with display terminal address according to instruction format. If the instruction is not sent to display terminal, system will clear up the buffer space and wait for the next instruction data, otherwise get attendance data and show them in the procedure interface of display terminal.

8.6 Conclusions

In RS485 network the mainframe can communicate with many readers. According to the system demands which need to accomplish information transmission between the mainframe and display terminal in the chapter , we defined custom

communication protocols between the mainframe and display terminal, connected display terminal into rs485 network with other readers, took display terminal as a reader and achieved unified access control at last. The system has run stably for a long time till now, and the running results show that the solution can fully meet the system requirements.

References

1. Jiang X, Wang X (2008) Study on logistic information acquisition technology in steelmaking practice based on RFID. Proceedings of IEEE symposium world congress on intelligent control and automation (WCICA 08), IEEE Press, June 2008, pp 7946–7950
2. Ding Z (2009) Research and realization on key technologies of RFID, University of Science and Technology of China, May 2009
3. Fu R, Tong Y (2009) Design and realization of attendance check management system based on Net and RFID. *Comput Digit Eng* 37:198–201
4. Zhou L, Hu L, Ying J (2008) Application of RF card in cinema ticket issue system. *Microcomput Inf* 24:243–244, 240
5. Zhou L, Hu L (2008) Application research of notebook computer simulating card reader in attendance system. *Comput Sys Appl* 17:62–65

Chapter 9

Sensitivity of Neurons Exposed to AC Induction Electric Field

Xiu Wang, Jiang Wang, Yanqiu Che, Chunxiao Han, Bin Deng and Xile Wei

Abstract Neuronal coding is one of the characteristics that exhibit the response of the neuron to the external stimulus. Based on a simplified Hodgkin–Huxley model, this paper firstly establishes a new neuronal model under the effect of alternating current (AC) induction electric field, and investigates the mechanism of neuronal encoding and the sensitivity of firing rate to noises. According to the model established, this paper obtains the bifurcation point of the model with the membrane voltage–time curves. Then by the data analysis of the mean firing rate–electric field frequency or the mean firing rate–amplitude curves, it obtains the relationship between the neuronal firing rate and noises as well as the electric field intensity, proving that the fluctuations of electric field frequency or amplitude can affect the sensitivity of neurons.

Keywords Sensitivity · Mean firing rate · AC induction electric field · OU noise

9.1 Introduction

Neurons are the minimal unit of structure and function in the nervous system, determining the specific characteristics of brains, such as memory and cognition. When subjected to certain stimulus, neurons will generate spiking. Different action potential sequence represents the firing pattern induced by different stimulation, namely neural coding, which is one of the most fundamental problems in the

X. Wang · J. Wang (✉) · Y. Che · C. Han · B. Deng · X. Wei
School of Electrical and Automation Engineering, Tianjin University, Weijin Road No. 92, Tianjin 300072, China
e-mail: jiangwang@tju.edu.cn

nervous system [1]. The information coding of neurons changes when stimulation parameters change, such as input and noises, which is called sensitivity.

A large number of theoretical studies and experimental results show that the external electrical field can change the firing characteristics of neurons, thereby change the function and properties of the nervous system [2]. On the one hand, the electromagnetic radiation of the external electric field will change the normal properties of neurons, leading the production of some neuropathies [3]. On the other hand, the electrical stimulus has become one of the main approaches of physical therapy for neuropathies, such as transcranial magnetic stimulation, TMS for short [4], and electro-axupuncture therapy. Since neurons are in an environment with a variety of noises, the noises may change the initial firing time of neurons, thereby affect neuronal coding and transmission of information [5], here it is called sensitivity. Lundstrom et al. interpreted neuronal sensitivity to mean or variance as conferring on the neuron integrator-like or differentiator-like properties [6]. Jing Yang et al. investigated the “critical sensitivity” phenomena of neural firing pacemaker via physiological experiments [7]. Although some literature refers to sensitivity of neurons, but all do not explore the neuronal sensitivity under AC induction electric field.

This chapter establishes a new model of neurons under external electric field on the basis of HH model, introduces the mean firing rate to describe the output of a neuron, and analyze neuronal sensitivity under noisy AC electric field from the view of comparing $F-A$, $F-f_c$ curves, then reveal the regulation of AC induction electric field to neuronal sensitivity.

9.2 Model Simplification

Hodgkin–Huxley Model is a four-dimensional model with the features of multi-variable, nonlinearity and strong-coupling. In order to investigate the sensitivity, this paper makes some simplification of the HH model: eliminate the time-dependence of m considering the relatively small time constant of m , adopt the m 's stable value m_∞ rather than m , and make h linearly related to n [8], thereby reduce the order of the model. Then superimpose a induced voltage V_e on the membrane of the neuron which represents the effect of electromagnetic radiation, while noises are superimposed on stimulus current I in the term of I_{noise} , then obtains the following two-dimensional model:

$$\begin{aligned}
 C \frac{dV}{dt} &= -G_{Na} m_\infty^3 h (V + V_e - E_{Na}) - G_K n^4 (V + V_e - E_K) \\
 &\quad - G_{leak} (V + V_e - E_{leak}) + (I + I_{noise}) + I_e \\
 \tau \frac{dn}{dt} &= n_\infty - n
 \end{aligned} \tag{9.1}$$

Where $C = 1\text{nF/cm}^2$, $G_{Na} = 50\text{mS/cm}^2$, $G_k = 36\text{mS/cm}^2$, $G_{leak} = 5\text{mS/cm}^2$, $E_{Na} = 50\text{ mV}$, $E_k = -77\text{ mV}$, $E_{leak} = -54\text{ mV}$, $k_m = 7$, $V_m = -40\text{ mV}$, $V_n = -45\text{ mV}$, $\tau = 5\text{ ms}$. There are five other relational expressions as below:

$$m_\infty = 1/(1 + \exp((V_m - V)/k_m))$$

$$n_\infty = 1/(1 + \exp((V_n - V)/k_n))$$

$$h = 0.89 - 1.1n$$

$$V_e = A \sin(\omega t)$$

$$I_e = C \frac{dV_e}{dt} = CA\omega \cos(\omega t)$$

The model has two variables: a fast variable V is the membrane voltage of the model, while a slow variable n is a combined variable denoting the probability of the Na^+ channel inactivation and K^+ channel activation. τ is the time constant of recovery variable n . V_e is external AC induction electric field. I_e is the current flowing across the membrane capacitance caused by induction electric field. I_{noise} represents the external OU noise whose intensity is determined by variable D .

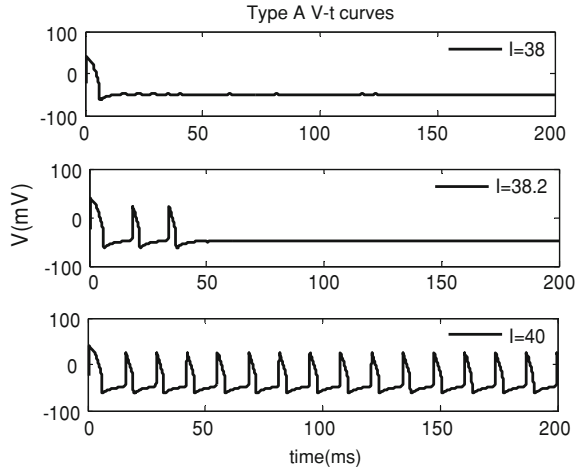
Lundstrom et al. [9] classify the neurons into three types: Type A, B+ and B− on the basis of how their F–I curves response to fluctuation input. This chapter focuses on the firing sensitivity of Type A neuron, choosing G_{Na} and τ as neural firing pacemaker, letting $G_{Na} = 50\text{ mS/cm}^2$, $\tau = 5\text{ ms}$ in Eq. 9.1. The noise used here is equivalent to an Ornstein–Uhlenbeck process which is called OU noise. The mean of OU noise is zero, the noise intensity D is variable. The Eq. 9.1 is solved via a fourth-order Runge–Kutta solver with a 0.02 ms step.

9.3 Analysis of Neuronal Sensitivity

9.3.1 Firing Patterns

Firstly, we obtain the Voltage–time curves (V – t curves) of the model under the direct current (DC) I without noises and induction electric field (Fig. 9.1). It shows that the firing pattern of the model neuron changes with the increasing stimulus: the neuron exhibits single-spiking when $I < 38.2\ \mu\text{A/cm}^2$ and begins periodic firing when $I = 38.2\ \mu\text{A/cm}^2$, then the firing rate increases with increasing current exhibiting continuously periodic firing. It is indicated that $I = 38.2\ \mu\text{A/cm}^2$ is the bifurcation point of the neuron, where the firing pattern transforms from single-spiking to periodic firing.

Fig. 9.1 Comparing the V-t curves of Type A neuron when $I = 38, 38.2, 40 \mu\text{A}/\text{cm}^2$



9.3.2 Sensitivity Under DC Electric Field

Analyze the sensitivity of Type A neuron under the input of noisy DC via the firing rate of different noises [9]. The noise intensity is taken as $D = 0, 10, 20, 30 \mu\text{A}/\text{cm}^2$ respectively, then obtain a series of firing rate by gradually change the DC stimulus I under each value of D in Fig. 9.2. Figure 9.2 shows: Type A neuron has different firing rate under different intensity of noises when $I \leq 40 \mu\text{A}/\text{cm}^2$, which indicates that noises have some effect on the firing pattern of Type A neuron and increase the firing rate, Type A neuron has a little sensitivity to noises; when $I \geq 40 \mu\text{A}/\text{cm}^2$ Type A neuron has almost the same firing rate under different intensity of noises, indicating that noises have no effect on the firing rate and Type A neuron has no sensitivity to noises which implies that Type A neuron has self-adaptation.

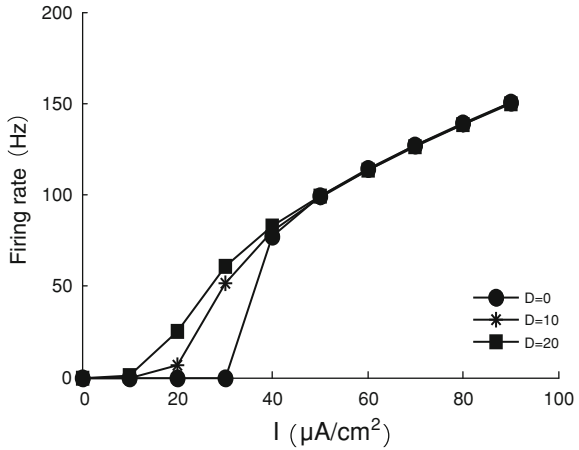
9.3.3 Sensitivity Under AC Induction Electric Field

The current is taken as the bifurcation point of Type A, namely $I = 38.2 \mu\text{A}/\text{cm}^2$, while AC Voltage $V_e = A \sin(2\pi ft)$.

1. Fixed the amplitude A and change the frequency f_e

Fixed A , then obtain the firing rate–voltage frequency curves $F-f_e$ under noise intensity $D = 0, 10, 20, 30 \mu\text{A}/\text{cm}^2$ respectively (Fig. 9.3): (a) Firing rate and voltage frequency have a nonlinear relationship. By comparing the $F-f_e$ curve of no noise in Fig. 9.3 (blue solid line) can it show that noises can increase the firing rate of Type A neuron obviously, and exhibit the sensitivity with certain laws. The sensitivity of Type A neuron to noises is in a fluctuating state transform between

Fig. 9.2 F-I curves of Type A neuron under the input of different noisy DC. The three curves (*star, square, circle curves*) correspond to noises of $D = 0, 10, 20, 30 \mu\text{A}/\text{cm}^2$ respectively



strong (at $f_e = 0, 10, 30, 50$ Hz) and weak (at $f_e = 20, 40, 60$ Hz). (b) Neurons show little sensitivity to noises, and have slight sensitivity only in the case of $f_e \in [0, 30]$ Hz. When $f_e \in [0, 30]$ Hz, firing rate has an inverse proportion to voltage frequency, while has a direct proportion when $f_e \in [30, 60]$ Hz. Comparing with the F- f_e curve of no noise indicates that the range of inverse proportion is decreased and the range of direct proportion is increased. Noises make the firing rate of Type A neuron obviously decrease (c-f).

Type A neuron has almost the same sensitivity curves of different noises which exhibit no strong sensitivity.

In conclusion, when $A \leq 20$ mV, adding noises can obviously change the firing rate of neurons, indicating that noisy AC electric field has effect on the sensitivity of the neuron in this case, increases the sensitivity of Type A neuron that has no sensitivity under the input of high noisy DC and make Type A neuron exhibits various sensitivity. While $A \geq 20$ mV, Type A neuron has slight sensitivity in the case of $f_e \in [0, 20]$ Hz or $f_e \in [0, 10]$ Hz and exhibits no sensitivity that has no obvious changes of the firing rate in the case of $f_e \geq 20$ Hz because of the Type A's self-adaptaion.

Fixed the the frequency f_e and change the amplitude A

Fixed the frequency f_e , then obtain the firing rate-voltage amplitude curves F-A with noise intensity $D = 0, 10, 20, 30 \mu\text{A}/\text{cm}^2$ respectively (Fig. 9.4). (a) The neurons exhibit weaker sensitivity in the case of $A \in [0, 10]$ mV while exhibit no in the other range. Firing rate has an inverse proportion to voltage amplitude and firing rate in the range of $A \in [17, 25]$ mV is decreased compared with the one that has no effect of noises, increased in the other range. (b) Similar to (a), the neuron exhibits weaker sensitivity in the case of $A \in [0, 10]$ mV while exhibits no in the other range. Firing rate in the range of $A \in [16.5, 30]$ mV is decreased compared with the one that has no effect of noises, increased in the other range. In the case of $A \geq 30$ mV, the effect of the AC electric field is so strong and the neuron has the

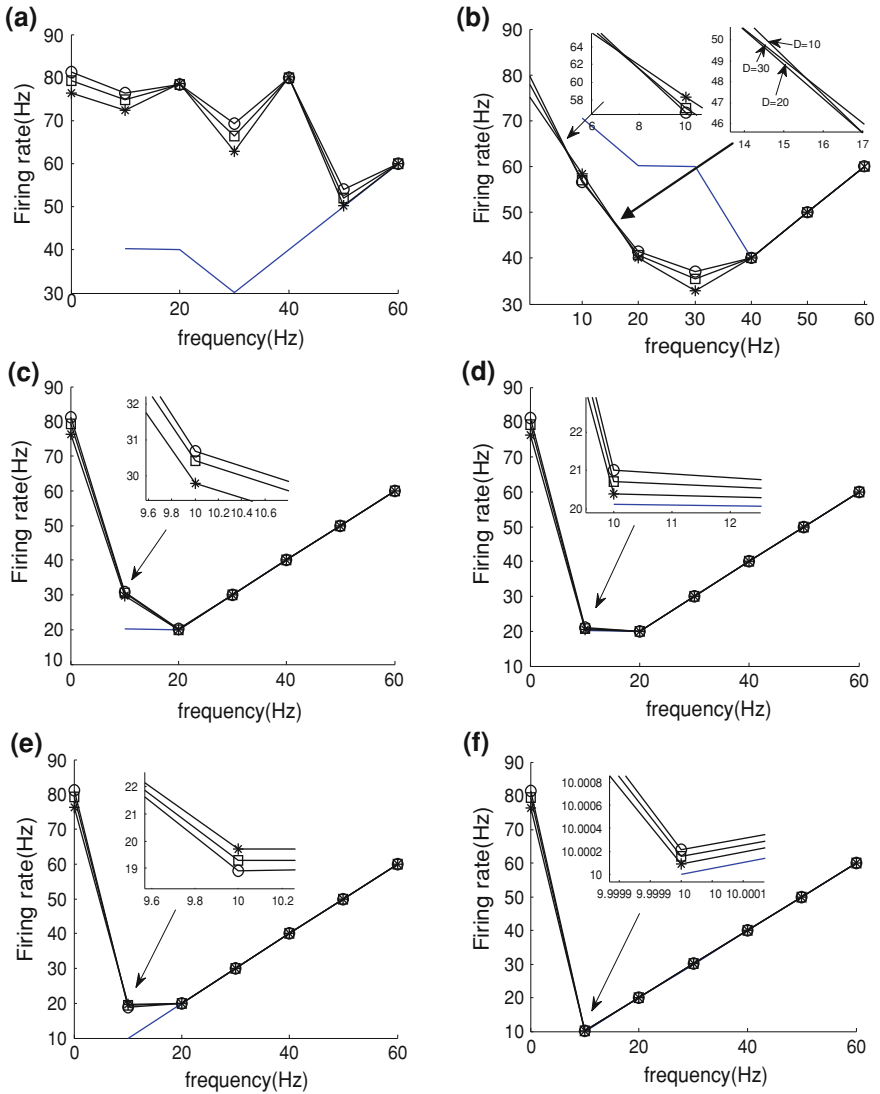


Fig. 9.3 F–f_e curves of Type A neuron under different noises of each fixed amplitude A . **a** A = 10 mV. **b** A = 20 mV. **c** A = 30 mV. **d** A = 40 mV. **e** A = 50 mV. **f** A = 60 mV. The three curves (*star, square, circle* curves) correspond to noises of D = 0, 10, 20, 30 $\mu\text{A}/\text{cm}^2$ respectively

adaptation that the firing rate of the neuron equals to the frequency of external electric field and does not change any more. (c) when $A \in [0, 30]$ mV, the neuron exhibits relatively strong sensitivity to noises and firing rate increases with the decrease of the voltage frequency. When $A \geq [0, 30]$ mV, the neuron exhibits no

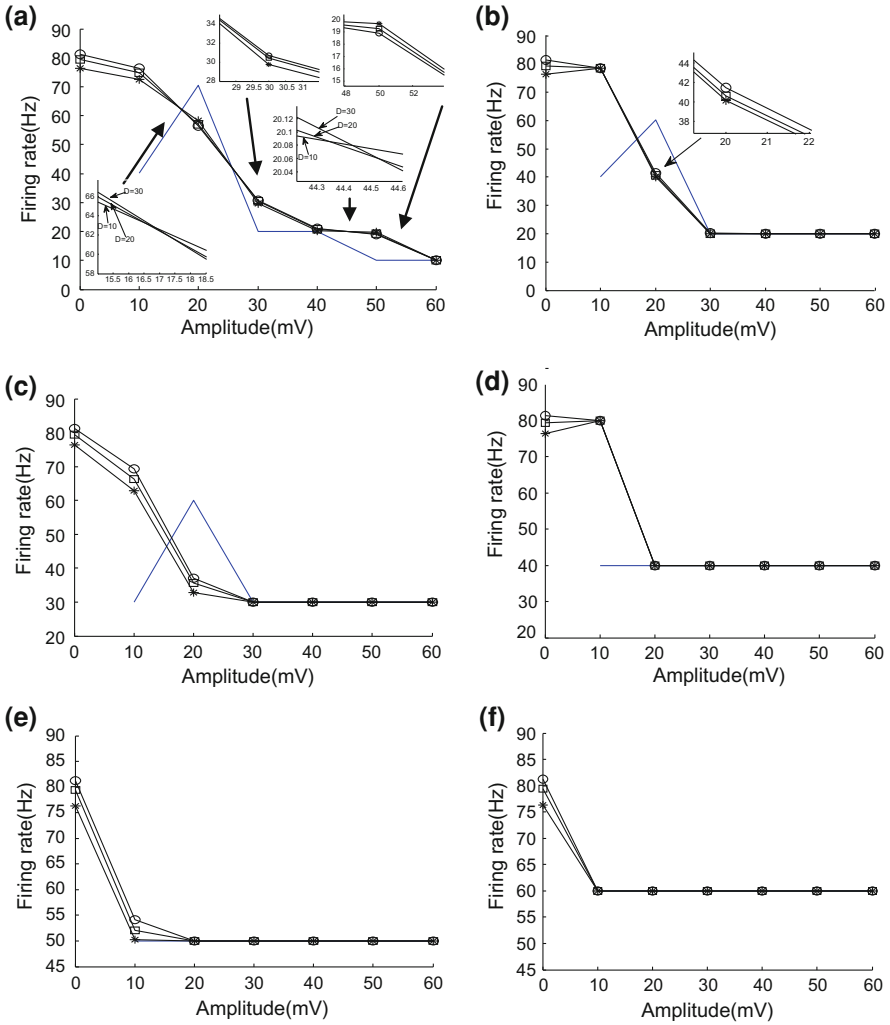


Fig. 9.4 F–A curves of Type A neuron under different noises of each fixed frequency. **a** $f_e = 10$ Hz. **b** $f_e = 20$ Hz. **c** $f_e = 30$ Hz. **d** $f_e = 40$ Hz. **e** $f_e = 50$ Hz. **f** $f_e = 60$ Hz. The three curves (*star*, *square*, *circle* curves) correspond to noises of $D = 0, 10, 20, 30 \mu\text{A}/\text{cm}^2$ respectively

sensitivity and the firing rate equals to the frequency of external electric field. Firing rate in the case of $A \in [15.5, 30]$ mV is decreased compared with the one of no effects of noises, while increased when $A \in [0, 15.5]$ mV. The result in the case of $A \geq 30$ mV is similar to (b) and (d). In the case of $A \in [0, 10]$ mV, the firing rate increases as the noise intensity increases, and the more stronger the noise is, the more the firing rate increases. And the neuron has slight sensitivity in this region.

When $A \geq 10$ mV, the neuron has adaptation that the noise has no effect on the firing rate. As the amplitude increases, the firing rate decreases until $A = 20$ mV, it equals to the external voltage frequency. (e) The neurons exhibit a little strong sensitivity in the case of $A \in [0, 20]$ mV and the firing rate has an inverse proportion to the voltage amplitude. When $A \in [0, 20]$ mV, noises make the firing rate increase compared with the one of no noisy effects. While $A \geq 20$ mV, the neuron exhibits no sensitivity and has adaptation that the firing rate equals to the voltage frequency. (f) The neurons exhibit a little weaker sensitivity in the case of $A \in [0, 10]$ mV and the firing rate in this range increases because of noises. As the amplitude A increases to 10 mV, the neuron exhibit adaptation that the firing rate remain the frequency of the external electric field.

In conclusion, noises make Type A neuron exhibits relatively strong sensitivity in the range of lower voltage amplitude, that is $A \in [0, 10]$ mV with $f_e = 10, 20, 40$ or 60 Hz, $A \in [0, 20]$ mV with $f_e = 50$ Hz, $A \in [0, 30]$ mV with $f_e = 30$ Hz. And in these range the firing rate increases compared with the one that has no influence of noises—the stronger the noise intensity is, the more the firing rate increases. In the case that AC electric field has large amplitude or frequency (refer to the analysis above), Type A neuron has self-adaptation that exhibit no sensitivity.

9.4 Conclusions

This paper firstly establishes a new neuronal model under the effect of AC induction electric field on the basis of a simplified HH model, then investigates the coding mechanism of Type A neuron to the stimulus and the sensitivity of firing rate to OU noises when the neuron is in an AC induction electric field. Obtain the conclusions as follows: Under the noisy AC electric field, Type A neuron has certain sensitivity in the case of $A \in [0, 10]$ mV or $A \in [10, 20]$ mV with $f_e \in [0, 50]$ Hz or $A \in [20, 30]$ mV with $f_e \in [0, 30]$ Hz. While Type A neuron exhibits no sensitivity in the case of $A \in [10, 20]$ mV with $f_e \geq 50$ Hz or $A \in [20, 30]$ mV with $f_e \geq 30$ Hz or $A \geq 30$ mV with any frequency. Because when the intensity of the electric field is low, noises superimposed on stimulus current have larger effect on the neuron that make the firing rate change, thereby make the neuron exhibit certain sensitivity, and when the intensity of the electric field is too high, the effect of the noise is much weaker than the one of the electric field, so the effect of noise is ignored and the neuron is equivalent to only have the effect of electric field stimulus, namely the neuron has self-adaptation. Therefore, in this case, no matter how the intensity of the noise changes, the firing rate has no change maintaining the frequency of external electric field. Namely, the neuron has no sensitivity.

References

1. Dayan P, Abbott LF (2001) *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Massachusetts Institute of Technology Press, Cambridge
2. Song B, Gu Y, Pu J, Reid B, Zhao ZQ, Zhao M (2007) Application of direct current electric fields to cells and tissues in vitro and modulation of wound electric field in vivo. *Nat Protoc* 2:1479–1489
3. Wertheimer N, Leeper ED (1979) Electrical wiring configurations and childhood cancer. *Am J Epidemiol* 109:273–284
4. Kleim JA, Kleim ED, Cramer SC (2007) Systematic assessment of training-induced changes in corticospinal output to hand using frameless stereotaxic transcranial magnetic stimulation. *Nat Protoc* 2:1754–2189
5. Wen ZH, Hu SJ, Dong XZ (2004) Noise-induced changes of dynamic characteristics of neurons. *J Fourth Mil Med Univ* 25:948–949
6. Lundstrom BN, Hong S, Higgs MH, Fairhall AL (2008) Two computational regimes of a single-compartment neuron separated by a planar boundary in conductance space. *Neural Comput* 20:1239–1260
7. Yang J, Duan Y, Xing J (2004) Direction character of “critical sensitivity” phenomena in neural firing pacemaker. *Acta Biophysica Sinica* 20:31–36
8. Gerstner W, Kistler WK (2002) *Spiking neuron models: single neurons, populations, plasticity*. Cambridge University Press, Cambridge
9. Lundstrom BN, Famulare M, Sorensen LB, Spain WJ (2009) Sensitivity of firing rate to input fluctuations depends on time scale separation between fast and slow variables in single neurons. *J Comput Neurosci* 27:277–290

Chapter 10

Spectra of Discrete Multi-Splitting Waveform Relaxation Methods to Determining Periodic Solutions of Linear Differential-Algebraic Equations

Xiaolin Lin, Liming Liu, Hong Wei, Yuan Sang, Yumei Wang
and Ronghui Lu

Abstract This chapter proposed spectra of discrete multi-splitting waveform relaxation (DMSWR) method to determine the periodic solutions of linear differential-algebraic equations. Based on the spectral radius of the derived operator by decoupled process, we obtained some convergent conditions for DMSWR method. The DMSWR method is an acceleration technique of the periodic waveform relaxation. A numerical example in circuit simulation is provided to further confirm the theoretical analysis and also to show that the multi-splitting technique can effectively accelerate the convergent performance of the iterative process.

X. Lin (✉) · L. Liu · Y. Sang

Faculty of Science, Shaanxi University of Science and Technology,
Xi'an 710021, People's Republic of China
e-mail: linxl@sust.edu.cn

L. Liu

e-mail: liuliming@sust.edu.cn

Y. Sang

e-mail: sangyuan@sust.edu.cn

L. Liu · Y. Sang · Y. Wang

College of Electric and Information Engineering, Shaanxi University of Science
and Technology, Xi'an 710021, People's Republic of China
e-mail: 0906032@sust.edu.cn

H. Wei

Manage Center of Internet, Shaanxi University of Science and Technology,
Xi'an 710021, People's Republic of China
e-mail: weihong@sust.edu.cn

R. Lu

Center of Computer Teach and Experiment, Wuyi University,
Wuyishan 354300, People's Republic of China
e-mail: lrh-mail@163.com

Keywords Discrete multi-splitting waveform relaxation • Linear differential-algebraic equations • Periodic solutions • Spectra of linear operators • Circuit simulation

10.1 Introduction

We have known that waveform relaxation (WR) method is a basic and efficient iteration technique for solving ordinary differential equations (ODEs) and differential-algebraic equations (DAEs) either in initial value problems or two point boundary problems in engineering applications, such as circuit simulation and mechanical modeling. In fact WR was first proposed to simulate MOS VLSI circuits [1–3]. Numerical algorithms incorporated with WR are relaxation-based methods and they are suitable for scientific computations of transient responses for very large dynamic systems. Many researchers have given convergence conditions of WR [4–9] and multi-splitting waveform relaxation (MSWR) [10, 11].

The resulted iterative systems with periodic constraint can be numerically solved by the sophisticated codes of DAEs or ODEs on boundary problem in public domain. In WR method, there are many decouple techniques such as Jacobian Iteration, Gauss–Seidel Iteration and so on. In order to accelerate the speed of convergence of WR, we present the multi-splitting waveform relaxation (MSWR) method [10], it is a novel splitting technique in engineering applications.

Consider the DAEs as the following:

$$\begin{cases} M\dot{x}(t) + Ax(t) + By(t) = f_1(t), & x(\mathbf{0}) = x(T), \\ Cx(t) + Ny(t) = f_2(t), & t \in [\mathbf{0}, T]. \end{cases} \quad (10.1)$$

where M and N are, respectively, $n_1 \times n_1$ and $n_2 \times n_2$ nonsingular matrices, A is an $n_1 \times n_1$ matrix, B is an $n_1 \times n_2$ matrix, C is an $n_2 \times n_1$ matrix, $f_1(t) \in R^{n_1}$ and $f_2(t) \in R^{n_2}$ ($t \in [\mathbf{0}, T]$) are two known input functions with period T , $x(t) \in R^{n_1}$ and $y(t) \in R^{n_2}$ ($t \in [\mathbf{0}, T]$) are to be computed. It is also obvious that $y(\mathbf{0}) = N^{-1}(f_2(\mathbf{0}) - Cx(\mathbf{0}))$ is for (10.1). Further, $y(\mathbf{0}) = y(T)$ is resulted from $x(\mathbf{0}) = x(T)$ and $f_2(\mathbf{0}) = f_2(T)$. We assume that the boundary condition on periodic solutions of (10.1) means that the condition $x(\mathbf{0}) = x(T)$ implies $\dot{x}(\mathbf{0}) = \dot{x}(T)$ and $y(\mathbf{0}) = y(T)$.

Let $M = M_{1l} - M_{2l}$, $A = A_{1l} - A_{2l}$, $B = B_{1l} - B_{2l}$, $C = C_{1l} - C_{2l}$, $N = N_{1l} - N_{2l}$ ($l = \mathbf{1}, \mathbf{2}, \dots, L$) and $(x^{(0)}(\cdot), y^{(0)}(\cdot))^T$ is a given initial guess. Now, we present the MSWR algorithm to compute the steady-state periodic response over one period for (10.1). The MSWR algorithm of (10.1) is:

$$\begin{cases} M_{1l}\dot{x}^{k,l}(t) + A_{1l}x^{k,l}(t) + B_{1l}y^{k,l}(t) = M_{2l}\dot{x}^{(k-1)}(t) + A_{2l}x^{(k-1)}(t) + B_{2l}y^{(k-1)}(t) + f_1(t), \\ C_{1l}x^{k,l}(t) + N_{1l}y^{k,l}(t) = C_{2l}x^{(k-1)}(t) + N_{2l}y^{(k-1)}(t) + f_2(t), \\ x^{k,l}(\mathbf{0}) = x^{k,l}(T), \quad y^{k,l}(\mathbf{0}) = y^{k,l}(T), \quad l = \mathbf{1}, \mathbf{2}, \dots, L, \\ x^{(k)}(t) = \sum_{l=1}^L E_l x^{k,l}(t), \quad y^{(k)}(t) = \sum_{l=1}^L \tilde{E}_l y^{k,l}(t), \quad t \in [\mathbf{0}, T], \quad k = \mathbf{1}, \mathbf{2}, \dots \end{cases} \quad (10.2)$$

where we suppose that M_{1l} and N_{1l} ($l = \mathbf{1}, \mathbf{2}, \dots, L$) are nonsingular, E_l and \tilde{E}_l ($l = \mathbf{1}, \mathbf{2}, \dots, L$) are non-negative diagonal matrix and $\sum_{l=1}^L E_l = I$, also $\sum_{l=1}^L \tilde{E}_l = I$ in this chapter. In order to preserve the consistency of the boundary conditions for every periodic iteration an initial guess $(x^{(0)}(t), y^{(0)}(t))^T$ in (10.2) should satisfy $(x^{(0)}(\mathbf{0}), y^{(0)}(\mathbf{0}))^T = (x^{(0)}(T), y^{(0)}(T))^T$ and $\dot{x}(\mathbf{0}) = \dot{x}(T)$. For any constant guess, the required boundary conditions are naturally held. Often, for a linear system we only consider its MSWR solutions in $C([\mathbf{0}, T], C^n)$ or $L^2([\mathbf{0}, T], C^n)$, here $n = n_1 + n_2$. This treatment can greatly simplify the theoretical analyzes on the MSWR. The convergence behaviors of the MSWR are mainly decided by the corresponding MSWR operators in these functions spaces and decouple process.

The WR solutions of initial value problems of equations as in (10.1) were reported in [12]. The expressions of spectra and pseudo-spectra for their WR operators were also clearly understood [13]. However, so far as we known, most of these theoretical convergence results are about the WR, and there are few chapters to theoretically analyze the spectra of the DMSWR operator for linear dynamic systems in the WR literatures. In this chapter we discuss the DMSWR operator derived from (10.2) where an analytic expression of its spectra is obtained. A finite-difference method is then used to solve the decoupled systems (10.2) in our test examples. The results of paper [6] are the special case of our results in this chapter. At same time, a numerical example in circuit simulation is provided to further confirm the theoretical analysis and also to show that the multi-splitting technique can effectively accelerate the convergent performance of the iterative process.

10.2 Spectra of DMSWR Operators and Finite-Difference for Solving MSWR Solutions

In this section we consider the discrete case of Sect. 10.2 and give a finite-difference formula for solving the MSWR solution of (10.1).

10.2.1 Spectra of DMSWR Operators

Now we discuss the application of linear multi-step method in the MSWR algorithm (10.2). For this purpose, let us fix the time increment $\tau = T/N$ and discretize (10.2) by a linear multi-step method, where its characteristic polynomials are $a(\xi)$ and $b(\xi)$, i.e., $a(\xi) = \sum_{j=0}^m \alpha_j \xi^j$ and $b(\xi) = \sum_{j=0}^m \beta_j \xi^j$, to obtain

$$\left\{ \begin{array}{l} \frac{1}{\tau} M_{1l} \sum_{j=0}^m \alpha_j x_{p-m+j}^{k,l} + A_{1l} \sum_{j=0}^m \beta_j x_{p-m+j}^{k,l} + B_{1l} \sum_{j=0}^m \beta_j y_{p-m+j}^{k,l} \\ = \frac{1}{\tau} M_{2l} \sum_{j=0}^m \alpha_j x_{p-m+j}^{(k-1)} + A_{2l} \sum_{j=0}^m \beta_j x_{p-m+j}^{(k-1)} + B_{2l} \sum_{j=0}^m \beta_j y_{p-m+j}^{(k-1)} + \sum_{j=0}^m \beta_j (f_1)_{p-m+j} \\ C_{1l} x_p^{k,l} + N_{1l} y_p^{k,l} = C_{2l} x_p^{(k-1)} + N_{2l} y_p^{(k-1)} + (f_2)_p, \quad l = 1, 2, \dots, L \\ x_p^{(k)} = \sum_{l=1}^L E_l x_p^{k,l}, \quad y_p^{(k)} = \sum_{l=1}^L \tilde{E}_l y_p^{k,l}, \quad p = 0, \pm 1, \pm 2, \dots, \quad k = 1, 2, \dots \end{array} \right. \quad (10.3)$$

In the above algorithm we assume that $a(\xi)$ and $b(\xi)$ have no common roots where $a(1) = 0$ and $\dot{a}(1) = b(1)$. In practical codes one adopts a convergent linear multi-step method to solve DAEs of (10.2). A special case of the linear multi-step method is the backward differentiation formula (BDF) where $b(\xi) = \xi^m$. The m-step constant BDF method converges to $O(\tau^m)$ for $m < 7$ (see [1]).

Let $x_\tau^{k,l}$ and $y_\tau^{k,l}$ stand for the infinite sequences $\{x_p^{k,l}\}_{p=-\infty}^\infty$ and $\{y_p^{k,l}\}_{p=-\infty}^\infty$ for all $l = 1, 2, \dots, L$ and similarly let $x_\tau^{(k)}$, $y_\tau^{(k)}$, $x_\tau^{(k-1)}$, $y_\tau^{(k-1)}$, $(f_1)_\tau$ and $(f_2)_\tau$ stand for the infinite sequences $\{x_p^{(k)}\}_{p=-\infty}^\infty$, $\{y_p^{(k)}\}_{p=-\infty}^\infty$, $\{x_p^{(k-1)}\}_{p=-\infty}^\infty$, $\{y_p^{(k-1)}\}_{p=-\infty}^\infty$, $\{(f_1)_p\}_{p=-\infty}^\infty$ and $\{(f_2)_p\}_{p=-\infty}^\infty$. These infinite sequences are N-periodic, for example it means that $x_{p+N}^{(k)} = x_p^{(k)}$ ($p = 0, \pm 1, \pm 2, \dots$) for the sequence $\{x_p^{(k)}\}_{p=-\infty}^\infty$. Now we simply rewrite (10.3) as

$$\left\{ \begin{array}{l} \frac{1}{\tau} a M_{1l} x_\tau^{k,l} + b A_{1l} x_\tau^{k,l} + b B_{1l} y_\tau^{k,l} = \frac{1}{\tau} a M_{2l} x_\tau^{(k-1)} + b A_{2l} x_\tau^{(k-1)} + b B_{2l} y_\tau^{(k-1)} + b (f_1)_\tau \\ C_{1l} x_\tau^{k,l} + N_{1l} y_\tau^{k,l} = C_{2l} x_\tau^{(k-1)} + N_{2l} y_\tau^{(k-1)} + (f_2)_\tau, \quad l = 1, 2, \dots, L \\ x_\tau^{(k)} = \sum_{l=1}^L E_l x_\tau^{k,l}, \quad y_\tau^{(k)} = \sum_{l=1}^L \tilde{E}_l y_\tau^{k,l}, \quad k = 1, 2, \dots \end{array} \right. \quad (10.4)$$

where we denote the infinite sequences $\left\{ \sum_{j=0}^m \alpha_j M_s x_{p-m+j}^{(r)} \right\}_{p=-\infty}^\infty$, $\left\{ \sum_{j=0}^m \beta_j A_s x_{p-m+j}^{(r)} \right\}_{p=-\infty}^\infty$ and $\left\{ \sum_{j=0}^m \beta_j B_s y_{p-m+j}^{(r)} \right\}_{p=-\infty}^\infty$ by $a M_s x_\tau^{(r)}$, $b A_s x_\tau^{(r)}$ and $b B_s y_\tau^{(r)}$.

Definition For an N-periodic sequence w_τ , its discrete Fourier coefficients are

$$\tilde{w}_p = \frac{1}{N} \sum_{q=1}^N w_q e^{-ipq(2\pi/N)}, \quad p = 0, \pm 1, \pm 2, \dots$$

By use of Definition, we know that $w_\tau = \sum_{q=1}^{N-1} \tilde{w}_q \varepsilon_{\tau,q}$, where $\varepsilon_{\tau,q} = \{e^{ipq(2\pi/N)}\}_{p=-\infty}^{\infty}$.

Condition (S) For the characteristic polynomials $a(\zeta)$ and $b(\zeta)$, we assume that the matrix

$$\begin{pmatrix} \frac{1}{\tau} \frac{a}{b}(\zeta^q) M_{1l} + A_{1l} & B_{1l} \\ C_{1l} & N_{1l} \end{pmatrix}^{-1} \quad q = 0, 1, \dots, N-1 \quad (10.5)$$

exist for the splitting matrices $M_{1l}, A_{1l}, B_{1l}, C_{1l}$ and N_{1l} ($l = 1, 2, \dots, L$) in which $\zeta = e^{i(2\pi/N)}$.

Let $z_\tau^{(r)} = [(x_\tau^{(r)})^T, (y_\tau^{(r)})^T]^T$, if Condition (S) holds, for any fixed k the solution of (10.4) can be written as

$$z_\tau^{(k)} = \mathbf{K}_\tau z_\tau^{(k-1)} + \varphi_\tau, \quad (10.6)$$

here

$$\mathbf{K}_\tau z_\tau = \sum_{q=0}^{N-1} \begin{pmatrix} \sum_{l=1}^L E_l \left(\frac{1}{\tau} \frac{a}{b}(\zeta^q) M_{1l} + A_{1l} \right) & \sum_{l=1}^L E_l B_{1l} \\ \sum_{l=1}^L \tilde{E}_l C_{1l} & \sum_{l=1}^L \tilde{E}_l N_{1l} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{l=1}^L E_l \left(\frac{1}{\tau} \frac{a}{b}(\zeta^q) M_{2l} + A_{2l} \right) & \sum_{l=1}^L E_l B_{2l} \\ \sum_{l=1}^L \tilde{E}_l C_{2l} & \sum_{l=1}^L \tilde{E}_l N_{2l} \end{pmatrix} \tilde{z}_q \varepsilon_{\tau,q}$$

and

$$\varphi_\tau = \sum_{q=0}^{N-1} \begin{pmatrix} \sum_{l=1}^L E_l \left(\frac{1}{\tau} \frac{a}{b}(\zeta^q) M_{1l} + A_{1l} \right) & \sum_{l=1}^L E_l B_{1l} \\ \sum_{l=1}^L \tilde{E}_l C_{1l} & \sum_{l=1}^L \tilde{E}_l N_{1l} \end{pmatrix}^{-1} \tilde{f}_q \varepsilon_{\tau,q}$$

in which $\tilde{f}_q = [(\tilde{f}_1)_q^T, (\tilde{f}_2)_q^T]^T$. With the same approach given in [14], we can get the following theorem (we omit the proof here).

Theorem 1 Under Condition (S) the spectral set of the DMSWR operator K_τ in (10.6) is

$$\sigma(K_\tau) = \cup \left\{ \sigma \left(\begin{pmatrix} \sum_{l=1}^L E_l \left(\frac{1-a}{\tau b} (\xi^q) \right) M_{1l} + A_{1l} & \sum_{l=1}^L E_l B_{1l} \\ \sum_{l=1}^L \tilde{E}_l C_{1l} & \sum_{l=1}^L \tilde{E}_l N_{1l} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{l=1}^L E_l \left(\frac{1-a}{\tau b} (\xi^q) \right) M_{2l} + A_{2l} & \sum_{l=1}^L E_l B_{2l} \\ \sum_{l=1}^L \tilde{E}_l C_{2l} & \sum_{l=1}^L \tilde{E}_l N_{2l} \end{pmatrix} \right) : q = \mathbf{0}, \mathbf{1}, \dots, N-1 \right\} \quad (10.7)$$

where $\xi = e^{i(2\pi/N)}$.

10.2.2 Finite-Difference for Solving MSWR Solutions

In this section we compute the iterative waveforms $\left[(\dot{x}^{(k)})^T(t), (\dot{y}^{(k)})^T(t) \right]^T$ ($k = \mathbf{1}, \mathbf{2}, \dots$) in (10.2) at $m + 1$ time-points, $t_0 = \mathbf{0}, t_1, t_2, \dots, t_m = T$, with a constant step-size τ . For any fixed k we approximate the derivatives $\dot{x}^{(k)}$ and $\dot{y}^{(k)}$ in (10.2) with the implicit Euler method. As a simple case of the linear multi-step method presented in Sect. 10.3.1, we now may write out the iterative matrix for discrete waveforms without using the discrete Fourier series technique. We will follow this form to do our computations in the next section. For the purpose, we denote that

$$X^{(r)} = \left[\left(x^{(r)} \right)^T(t_1), \dots, \left(x^{(r)} \right)^T(t_m) \right]^T \in R^{mm_1},$$

$$Y^{(r)} = \left[\left(y^{(r)} \right)^T(t_1), \dots, \left(y^{(r)} \right)^T(t_m) \right]^T \in R^{mm_2},$$

$$F_1 = \left[f_1^T(t_1), f_1^T(t_2), \dots, f_1^T(t_m) \right]^T \in R^{mm_1},$$

$$F_2 = \left[f_2^T(t_1), f_2^T(t_2), \dots, f_2^T(t_m) \right]^T \in R^{mm_2}.$$

It is mentioned here that the order of discrete equations is different from that of Sect. 10.2.1 for the differential part and the algebraic part. By $x^{(r)}(t_0) = x^{(r)}(t_m)$ the discrete MSWR form of (10.2) is

$$\begin{cases} H_{1l}X^{k,l} + H_{2l}Y^{k,l} = J_{1l}X^{(k-1)} + J_{2l}Y^{(k-1)} + \tau F_1 \\ H_{3l}X^{k,l} + H_{4l}Y^{k,l} = J_{3l}X^{(k-1)} + J_{4l}Y^{(k-1)} + \tau F_2 \\ l = \mathbf{1}, \mathbf{2}, \dots, L \\ X^{(k)} = \sum_{l=1}^L E_l X^{k,l}, Y^{(k)} = \sum_{l=1}^L \tilde{E}_l Y^{k,l}, \quad k = \mathbf{1}, \mathbf{2}, \dots \end{cases} \quad (10.8)$$

where

$$H_{1l} = \begin{pmatrix} M_{1l} + \tau A_{1l} & 0 & \dots & 0 & -M_{1l} \\ -M_{1l} & \ddots & & & 0 \\ 0 & \ddots & M_{1l} + \tau A_{1l} & & \\ & & -M_{1l} & \ddots & \vdots \\ \vdots & & & \ddots & M_{1l} + \tau A_{1l} \\ & & & & \ddots & 0 \\ 0 & \dots & 0 & -M_{1l} & M_{1l} + \tau A_{1l} \end{pmatrix},$$

$$H_{2l} = \begin{pmatrix} \tau B_{1l} & & \\ & \ddots & \\ & & \tau B_{1l} \end{pmatrix}, \quad H_{3l} = \begin{pmatrix} C_{1l} & & \\ & \ddots & \\ & & C_{1l} \end{pmatrix}, \quad H_{4l} = \begin{pmatrix} N_{1l} & & \\ & \ddots & \\ & & N_{1l} \end{pmatrix}$$

and

$$J_{1l} = \begin{pmatrix} M_{2l} + \tau A_{2l} & 0 & \dots & 0 & -M_{2l} \\ -M_{2l} & \ddots & & & 0 \\ 0 & \ddots & M_{2l} + \tau A_{2l} & & \\ & & -M_{2l} & \ddots & \vdots \\ \vdots & & & \ddots & M_{2l} + \tau A_{2l} \\ & & & & \ddots & 0 \\ 0 & \dots & 0 & -M_{2l} & M_{2l} + \tau A_{2l} \end{pmatrix},$$

$$J_{2l} = \begin{pmatrix} \tau B_{2l} & & \\ & \ddots & \\ & & \tau B_{2l} \end{pmatrix}, \quad J_{3l} = \begin{pmatrix} C_{2l} & & \\ & \ddots & \\ & & C_{2l} \end{pmatrix}, \quad J_{4l} = \begin{pmatrix} N_{2l} & & \\ & \ddots & \\ & & N_{2l} \end{pmatrix}.$$

Let $E_l \in R^{m(n_1+n_2)}$ ($l = 1, 2, \dots, L$) be non-negative diagonal matrix and $\sum_{l=1}^L E_l = I$. For any fixed step-size τ , the convergence condition of the above iterative algorithm can be concluded in the following theorem.

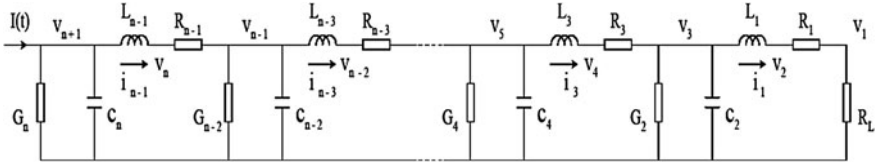


Fig. 10.1 A linear periodic DAEs circuit with n even

Theorem 2 The DMSWR algorithm (10.8) is convergent if

$$\rho \left(\sum_{l=1}^L E_l \begin{pmatrix} H_{1l} & H_{2l} \\ H_{3l} & H_{4l} \end{pmatrix}^{-1} \begin{pmatrix} J_{1l} & J_{2l} \\ J_{3l} & J_{4l} \end{pmatrix} \right) < \mathbf{1} \tag{10.9}$$

10.3 Numerical Experiments

We define that the iterative error is the sum of the squared difference of successive waveforms taken over all time-points.

10.3.1 Example One

Example one is a test circuit shown in Fig. 10.1 where n is 10. This circuit is taken from [15]. It is a general form of uniformly dissipative low-pass ladder filter circuit with a current-source input and a voltage output. The circuit equations have a form as (10.1) where

$$x(t) = [i_1(t), v_3(t), i_5(t), v_7(t), i_9(t), v_{11}(t)]^T$$

and $y(t) = [v_1(t), v_2(t), v_4(t), v_6(t), v_8(t), v_{10}(t)]^T \in R^6,$

$f_1(t) = [\mathbf{0}, \dots, \mathbf{0}, I(t)]^T \in R^{10}$ and $f_2(t) = [\mathbf{0}, \dots, \mathbf{0}]^T \in R^6,$ for any givent $t \in [0, T].$

Now, M, A, B, C and N in (10.1) are some concrete matrices. For $M, A \in R^{10 \times 10}$ we have $M = \text{diag}(L_1, C_2, L_3, C_4, \dots, L_9, C_{10})$ and $A = \text{diag}(\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_5)$ where

$$\tilde{A}_i = \begin{pmatrix} \mathbf{0} & -\mathbf{1} \\ \mathbf{1} & G_{2i} + R_{2i+1}^{-1} \end{pmatrix} \quad (i = \mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}), \quad \tilde{A}_5 = \begin{pmatrix} \mathbf{0} & -\mathbf{1} \\ \mathbf{1} & G_{10} \end{pmatrix}$$

The matrix $B \in R^{10 \times 6}$ and $C \in R^{6 \times 10}$ are

$$B = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -R_3^{-1} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -R_5^{-1} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -R_7^{-1} & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -R_9^{-1} \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$C = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -R_3^{-1} & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -R_5^{-1} & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -R_7^{-1} & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -R_9^{-1} & -1 & 0 \end{pmatrix}.$$

For $N \in R^{6 \times 6}$ we have $N = \text{diag}(\tilde{N}_1, \tilde{N}_2)$, where $\tilde{N}_2 = \text{diag}(R_3^{-1}, R_5^{-1}, R_7^{-1}, R_9^{-1}) \in R^{4 \times 4}$ and $\tilde{N}_1 = \begin{pmatrix} R_1^{-1} + R_2^{-1} & -R_1^{-1} \\ -R_1^{-1} & R_1^{-1} \end{pmatrix}$.

We seek its periodic responses by the MSWR algorithm. In our computations we use the discrete algorithm (10.8). For simplicity we let $n = 10$ and $T = 2\pi$, all circuit parameters are set to be one. The boundary values satisfy $x(\mathbf{0}) = x(2\pi)$ and $y(\mathbf{0}) = -N^{-1}Cx(\mathbf{0}) (= y(2\pi))$.

For example one, we use the Jacobi splitting to split the matrices M and N , i.e., M_1 and N_1 are diagonal matrices of M and N if we adopt the symbols in (10.2). The matrices B_1 and C_1 are

$$B_1 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in R^{10 \times 6},$$

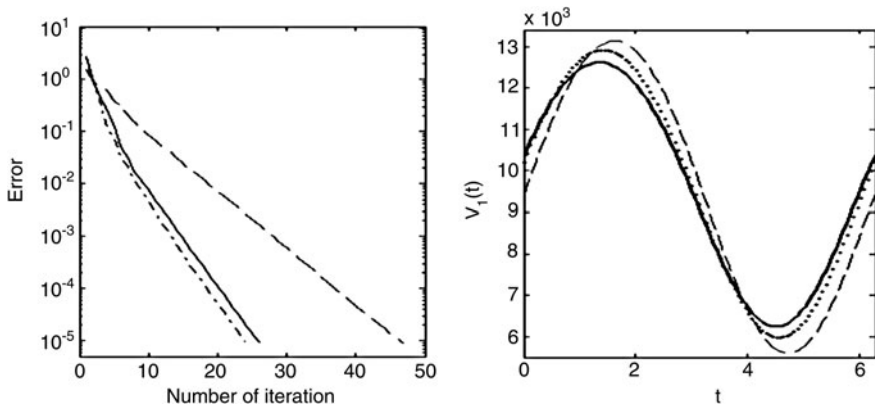


Fig. 10.2 Computed results for Example One. *Left* DMSWR iterations (*dashed line*, *solid line*, and *point line* for Jacobi splitting, Gauss–Siedel splitting and MSWR, respectively). *Right* approximate waveforms ($k = 20$ for Jacobi splitting, $k = 12$ for Gauss–Siedel splitting and $k = 6$ for DMSWR) of the voltage $v_1(t)$ (*solid line*)

$$C_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \end{pmatrix} \in R^{6 \times 10}.$$

For the matrix A we have two ways to treat its splitting, for $l = 1$, we simply do not split A , i.e., $A_1 = A$; for $l = 2$, we split A as

$$A_1 = \begin{pmatrix} 2 & & & & & & & & & \\ & 1 & & & & & & & & \\ & & \ddots & & & & & & & \\ & & & \ddots & & & & & & \\ & & & & 2 & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & \ddots & & & \\ & & & & & & & \ddots & & \\ & & & & & & & & 1 & 2 \end{pmatrix} \in R^{10 \times 10}$$

Let $\xi = i\zeta (\zeta \in R)$, the spectral of $\sigma(K)$ and $\sigma(\tilde{K}_\infty) = \overline{\cup\{\sigma(K(i\zeta)) : \zeta \in R\}}$ can be calculated for $E_1 = 0.6I$ and $E_2 = 0.4I$ in which $p = 0, \pm 1, \dots, \pm 50$ and $\zeta = 0, \pm 0.1, \dots, \pm 49.9 \pm 50$.

To compute the MSWR solution of the system, we let the input function $I(t) = I(t + 2\pi)$ satisfy $I(t) = \begin{cases} t, & 0 \leq t \leq 0.5\pi; \\ 0.5\pi, & 0.5\pi \leq t \leq 1.5\pi; \\ (2\pi - t), & 1.5\pi \leq t \leq 2\pi. \end{cases}$

The time-step is 0.02π sec and the initial guess is the zero function. The convergence results and three approximate waveforms for the voltage $v_1(t)$ are shown in Fig. 10.2.

10.4 Conclusions

We have successfully deduced an analytic expression of the spectral set on the DMSWR operator for a linear system of DAEs under a normal periodic constraint. The convergent conditions of the DMSWR algorithm on periodic solutions can be conveniently chosen from this useful expression, namely the DMSWR algorithm converges to the exact periodic response if the supremum value of spectral radii for a series of complex matrices is less than one. The convergent condition of the chapter is necessary and sufficient for the DMSWR algorithm.

Acknowledgments This work was supported by the Ph.D. Programs Foundation of Shaanxi University of Science and Technology (Grant No. BJ10-23).

References

1. Lelarsmee E, Ruehli AE, Sangiovanni-Vincentelli AL (1982) The waveform relaxation method for time-domain analysis of large scale integrated circuits [J]. *IEEE Trans CAD IC Syst* 1(3):131–145
2. White JK, Sangiovanni-Vincentelli A (1986) *Relaxation techniques for the simulation of VLSI circuits*. Kulwer, Boston
3. Gristede GD, Ruehli AE, Zukowski CA (1998) Convergence properties of waveform relaxation circuit simulation methods. *IEEE Trans Circ Syst Part I* 45(7):726–738
4. Ascher UM, Petzold LR (1998) *Computer methods for ordinary differential equations and differential-algebraic equations*. SIAM, Philadelphia
5. Jiang YL, Luk WS, Wing O (1997) Convergence-theoretics of classical and Krylov waveform relaxation methods for differential-algebraic equations. *IEICE Trans Fund Electr Commun Comput Sci* E80-A(10):1961–1972
6. Jiang YL, Chen RMM (2005) Computing periodic solutions of linear differential-algebraic equations by waveform relaxation. *Math Comp* 74:781–804
7. Fan ZC (2011) Waveform relaxation method for stochastic differential equations with constant delay. *Appl Num Math* 61:229–240
8. Bai ZZ, Yang X (2011) On convergence conditions of waveform relaxation methods for linear differential-algebraic equations. *J Comput Appl Math* 235:2790–2804
9. Wu SL, Huang CM (2009) Convergence analysis of waveform relaxation methods for neutral differential-functional systems. *J Comput Appl Math* 223:263–277
10. Lin XL, Jiang YL, Wang Z (2008) Multi-splitting waveform relaxation methods to determining periodic solutions of linear differential-algebraic equations. In: *Proceedings of the 4th international conference on natural computation and the 5th international conference on fuzzy systems and knowledge discovery*, vol 1, pp 325–330
11. Lin XL, Wang L (2010) Computing periodic solutions of linear differential equations by multi-splittings waveform relaxation [J]. *Proc 2010 Int Conf Comput Appl Syst Model* 6:V6366–V6370

12. Jiang YL, Luk WS, Wing O (1997) Convergence-theoretics of classical and Krylov waveform relaxation methods for differential-algebraic equations. *IEICE Trans Fund Electr Commun Comput Sci* E80-A(10):1961–1972
13. Jiang YL, Wing O (2000) A note on the spectra and pseudo-spectra of waveform relaxation operators for linear differential-algebraic equations. *SIAM J Numer Anal* 38(1):186–201
14. Vandewalle S, Piessens R (1993) On dynamic iteration methods for solving time-periodic differential equations. *SIAM J Numer Anal* 30(1):286–303
15. Weinberg L (1962) *Network analysis and synthesis*. McGraw-Hill Book Company, New York

Chapter 11

A Low Power Limiting Amplifier Designed for the RSSI of a 5.8 GHz ETC Receiver

Hang Yu, Yan Li, Wongchen Wei, Lai Jiang,
Shengyue Lin and Zhen Ji

Abstract A low power limiting amplifier designed for received signal strength indicator of a 5.8 GHz ETC receiver is presented in this chapter. Systematic analysis is carried out first to determine the optimal number of gain stages and the gain-bandwidth product of each stage. Then the detailed circuit design is discussed. The limiting amplifier consists of twelve AC-coupled gain stages, and is implemented using the standard CMOS 0.18 μm technology. The design is validated under Zeni design environment, and the simulation results show that the limiting amplifier achieves overall small signal gain of 98 dB, with 22 MHz bandwidth. Powered by a 1.8 V DC supply voltage, the total power consumption of the limiting amplifier is only 5.5 mW.

Keywords CMOS · Electronic tolling collection (ETC) · Limiting amplifier · Received signal strength indicator (RSSI)

H. Yu · Y. Li (✉) · W. Wei · L. Jiang · S. Lin · Z. Ji
Shenzhen City Key Laboratory of Embedded System Design,
College of Computer Science and Software Engineering,
Shenzhen University, Shenzhen, 518060
Guangdong, People's Republic of China
e-mail: liyan@szu.edu.cn

H. Yu
e-mail: yuhang@szu.edu.cn

L. Jiang
e-mail: jianglai@szu.edu.cn

Z. Ji
e-mail: jizhen@szu.edu.cn

11.1 Introduction

In many major cities in china, traffic congestion is a prominent problem with the rapid increase of automobiles. Electronic tolling collection (ETC) is known to be an efficient method for congestion alleviation, for in an ETC system, the vehicle is allowed to be charged without slowing down, and therefore the traffic speed can be largely improved. The Standardization Administration of the People's Republic of China (SAC) has already released the Chinese standard of Dedicated Short Range Communications (DSRC), in which the ETC operating frequencies is defined to be from 5.8 to 5.9 GHz [1].

Currently, Shenzhen University is developing an ETC system. In the system, direct conversion architecture is adopted to implement the 5.8 GHz receiver due to its simple, robust structure and low power consumption. In order to prevent the undesired DC level presented at the base-band, which is primarily caused by the leakage from the local oscillator to the receiver front-end, from degrading the receiver performance, received signal is first down-converted to an intermediate frequency of 10 MHz.

In the envisioned application, when multiple vehicles are presented at the tolling gate, each should be assigned a distinct communication channel. To help select the vacant channel automatically, a received signal strength indicator (RSSI) is included in the ETC receiver design.

The RSSI monitors the received signal strength while the receiver scans all the communication channels, and bi-directional link between the vehicle and the tolling station can only be established once a vacant channel is found.

The RSSI is generally realized in logarithmic form because the wide dynamic variation of the received signal that can be represented within a limited indication range. The logarithmic characteristic can be realized through piecewise linear approximation by using a limiting amplifier with multiple cascaded gain stages [2]. As shown in Fig. 11.1, The output of each stage is rectified, summed together, and translated into electrical quantities, such as an output current or voltage, that are proportional to the input power level in terms of dBm.

This chapter focuses on the limiting amplifier design for RSSI in the ETC system. In Sect. 11.2, a systematic level optimization process is presented to help determine key parameters of the limiting amplifier. The detailed circuit implementation of the cascaded gain stages is shown in Sect. 11.3. The design is validated under the Zeni design environment, and the simulation results are discussed in Sect. 11.4. A concluding remark is given in Sect. 11.5.

11.2 Systematic Optimization

For a limiting amplifier design, the number of the cascading gain stages associated with their gain, bandwidth, and power determines the overall circuit performance. In order to achieve a large dynamic range, and at the same time minimize the

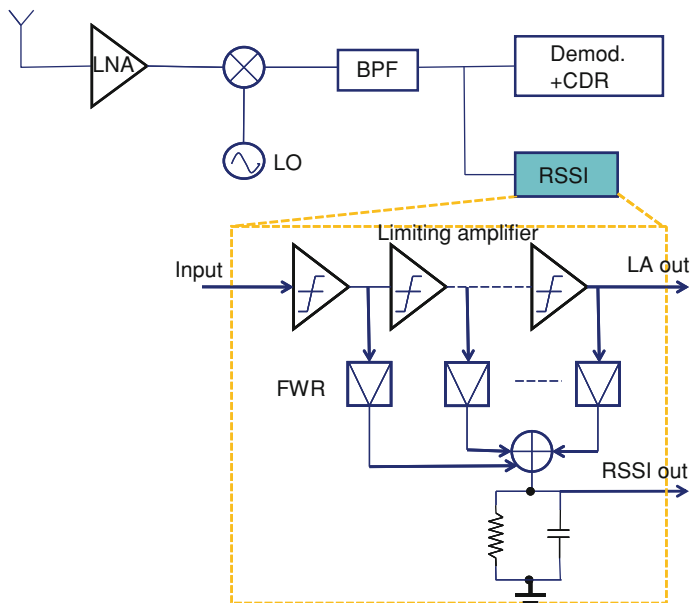


Fig. 11.1 Architecture of the ETC receiver and RSSI under developing in Shenzhen University

overall power consumption, optimal design parameters, such as the number of gain stages and the gain-bandwidth product (GBW) of each gain stage, have to be determined.

For design simplification, the cascading gain stages are kept identical. In its 1st order approximation, the identical gain stage is assumed to be linear, and the transfer function with voltage gain A can be written as:

$$\begin{aligned} V_{out} &= AV_{in} \quad \text{for } V_{in} < V_s \\ V_{out} &= V_L \quad \text{for } V_{in} \geq V_s, \end{aligned} \tag{11.1}$$

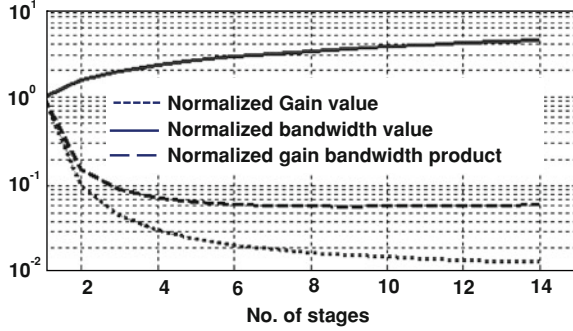
in which V_s is the threshold voltage, above which a gain stage is saturated, and V_L is the corresponding output voltage.

If the input signal is small enough, all gain stages operate in the linear region, and the output voltage of the limiting amplifier is $AN \times V_{in}$, in which N is the number of gain stages. As the input signal strength increases, the gain stages are driven into saturation one by one starting from the last stage, resulting to an approximated logarithmic response.

Given the overall small signal gain and bandwidth of the limiting amplifier AV and fV , the required gain and bandwidth of the identical gain stage can be derived as [2]:

$$A_C = A_V^{(1/N)-1} \tag{11.2}$$

Fig. 11.2 Normalized gain, bandwidth and GBW of a identical gain stage for a limiting amplifier with 100 dB small signal gain



$$f_c = \frac{f_v}{\sqrt{2^{1/N} - 1}} \quad (11.3)$$

For a targeted 100 dB overall small signal gain, the normalized gain, bandwidth of the identical gain stage for specified number of stages are plots in Fig. 11.2. The voltage gain for each gain stage is obviously reduced if the cascading number of stages increases, but it requires each gain stage to have larger bandwidth to compensate the increased number of poles. Figure 11.2 also includes the normalized GBW of the identical gain stage.

Total power consumption of the limiting amplifier is another key design parameter. A large number of gain stages not only results to alleviated GBW requirement, it can also generate a logarithmic response with less error [2]. However, the overall power consumption might increase. To facilitate the optimal design, the relation between the overall power consumption of the limiting amplifier PV and GBW is derived.

PV is the product of the number of stage N and the power consumption of the identical gain stage, which can be expressed as [3]

$$P_V = N \times P_C \propto N \times (\text{GBW})^2 \quad (11.4)$$

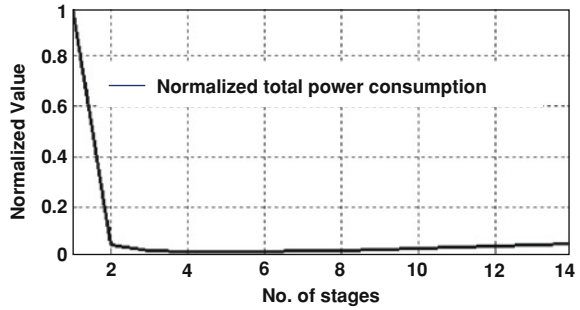
GBW in (11.4) is the ratio of the gain stage trans-conductance and the input capacitance of the following stage. It can be shown the GBW is also proportional to square root of the identical gain stage power consumption P_C , as in (11.5).

$$\text{GBW} = \frac{g_m}{C} = \frac{\sqrt{2\mu C_{ox}(W/L)I_d}}{kWL} \propto \sqrt{\frac{I_d}{WL^3}} = \sqrt{\frac{P_C}{V_{DD}WL^3}} \quad (11.5)$$

In (11.5), I_d is the bias current of the identical gain stage, W and L are the width and length of the MOS transistor that provides the trans-conductance.

The total power consumption of the limiting amplifier for 100 dB overall small signal gain is plotted in Fig. 11.3. Clearly, for the specified gain, the total power consumption decreases as the number of stages becomes larger. If the limiting amplifier consists of more than 4 stages, no obvious power increase can be

Fig. 11.3 Normalized total power consumption of a limiting amplifier with 100 dB small signal gain



observed. Therefore, more gain stages can be used to better approximate the required logarithmic response [4]. In this design, a total of twelve gain stages are used, with the gain of each set to 8.5 dB.

11.3 Circuit Implementation

The limiting amplifier consists of twelve cascading gain stages. Since the received signal is down-converted to 10 MHz intermediate frequency, the DC-offset cancellation scheme usually implemented in the RSSI designs is not required [5], and each stage can be AC-coupled to the following stage. Thus, the circuit design is greatly simplified.

Figure 11.4 shows the schematic of a single gain stage. A simple differential amplifier with NMOS input pair with load resistance R_D is utilized. To improve the amplifier linearity, a source degeneration resistor R_S is added. The amplifier input pair is self-biased by the passive low pass network consisting of R_1 , R_2 , and therefore only one bias voltage V_b is required. Neglecting the gate-bulk capacitance of the input NMOS pair, the overall voltage gain of the identical gain stage can be written as:

$$A_V = \frac{G_m \times R_D \parallel C_L}{1 + sR_1C_1} = \frac{sG_m R_D R_1 C_1}{(1 + sR_1C_1)(1 + sR_D C_L)} \quad (11.6)$$

The voltage transfer function of the gain stage has a bandpass characteristic, with its passing band determined by the two pole $1/R_1C_1$ and $1/R_D C_L$.

11.4 Design Validation

The limiting amplifier is implemented using CMOS 0.18 μm technology, and validated under the Zeni design environment.

In this design, C_1 , C_L and R_D are selected as 950 fF, 400 fF and 3.76 K Ω , and the simulated frequency response of a single gain stage is shown in Fig. 11.5.

Fig. 11.4 Schematic of the identical gain stage

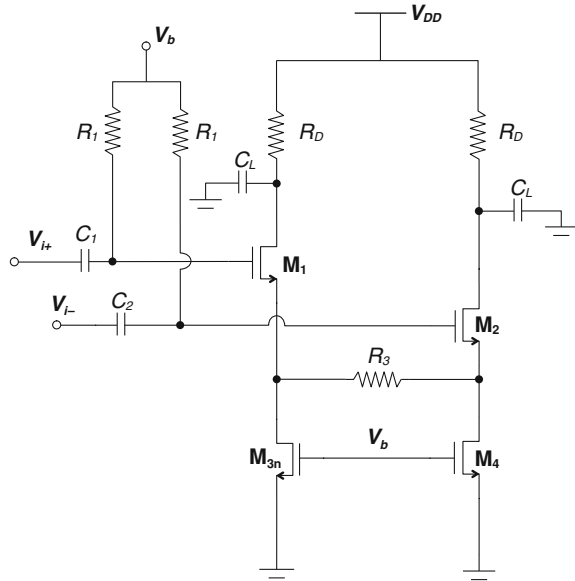
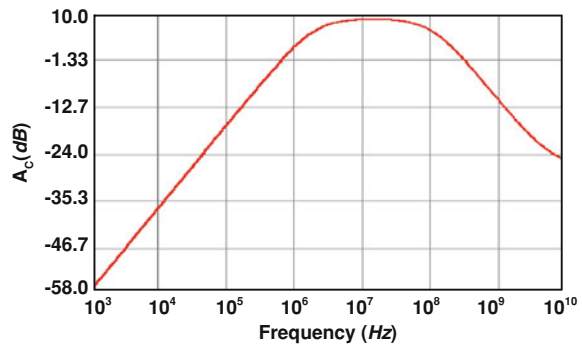


Fig. 11.5 Frequency response of the identical gain stage



A single gain stage has 8.5 dB small signal gain, and the 3 dB bandwidth is 2 ~ 110 MHz.

In total twelve identical gain stages are cascaded to form the limiting amplifier. The overall small signal gain of the completed limiting amplifier is simulated and shown in Fig. 11.6. The limiting amplifier achieves 98 dB overall small signal gain, and the 3 dB bandwidth is 5.6–22 MHz, approximately 1/5 of a single gain stage.

The -3 dB input sensitivity of the limiting amplifier is also simulated by varying the input signal strength from -100 to 0 dBm (Fig. 11.7). The limiting amplifier -3 dB input sensitivity can be defined as the input power that causes output power 3 dB lower than the saturated constant level, and is about -60 dBm for this design.

Fig. 11.6 Frequency response of the 12-stage limiting amplifier

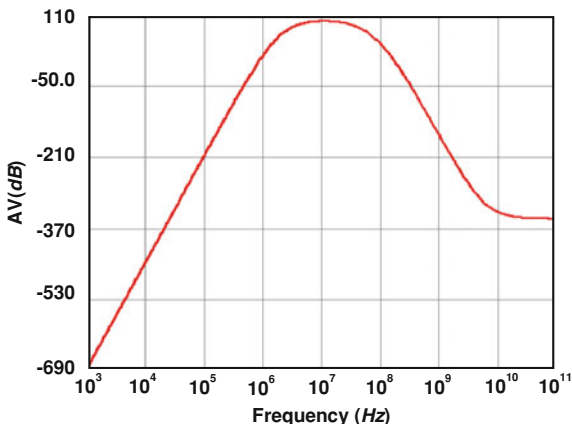
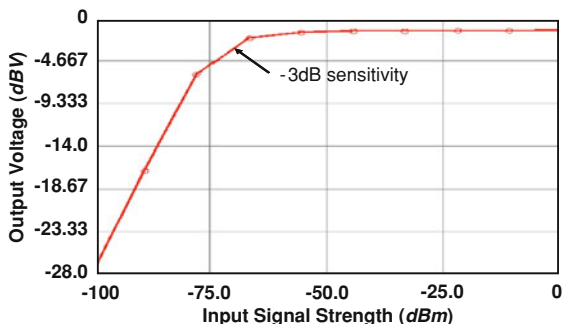


Fig. 11.7 The simulated -3 dB sensitivity of the limiting amplifier



With 1.8 V DC supply voltage, the power consumption of the completed limiting amplifier is 5.5 mW.

11.5 Conclusion

This chapter discusses a low power limiting amplifier designed for the RSSI of a 5.8 GHz ETC receiver. The number of stage and the gain bandwidth of a single gain stage are first determined through system level analysis, and then the detailed circuit implementation is presented. The limiting amplifier is designed and implemented using standard CMOS 0.18 μm technology. Formed by cascading twelve identical gain stages, the limiting amplifier achieves 100 dB overall small signal gain and 22 MHz bandwidth, while only consumes 5.5 mW of power from a 1.8 V voltage supply.

Acknowledgments The authors would like to thank the NSFC under grant No. 60901016, the project 10151806001000016 supported by Guangdong R/D Fund, the project JC201005280477A supported by Shenzhen city and the SRF for ROCS, SEM.

References

1. GB/T 20851-2007 (2007) Electronic toll collection-dedicated short range communication interface with roadside unit and lane controller
2. Holdenried CD, Haslett JW, McRory JG (2002) A DC-4-GHz true logarithmic amplifier: theory and implementation. *IEEE J Solid State Circ* 37(10):1290–1299
3. Huang PC, Chen YH, Wang CK (2000) A 2-V 10.7-MHz CMOS limiting amplifier/RSSI. *IEEE J Solid State Circ* 35(10):1474–1480
4. Barber WL, Brown ER (1980) A true logarithmic amplifier for radar IF applications. *IEEE J Solid State Circ* SC-15(3):291–295
5. Kim HS, Ismail M, Olsson HB (2001) CMOS limiters with RSSIs for bluetooth receivers. *MWSCAS*, vol 2. IEEE, Dayton, pp 812–815

Chapter 12

A 3rd order Opamp-Based Tunable Low-Pass Filter Design for Data Demodulator of a 5.8 GHz ETC RF Receiver

Yan Li, Hang Yu, Shengyue Lin, Lai Jiang, Rongchen Wei
and Zhen Ji

Abstract Electronic tolling collection (ETC) will be applied in the construction of the new national highway network in China and the Chinese standard has been released. A 5.8 GHz receiver satisfied the standard is one of the key components in the ETC system. In this work, a 3rd order low-pass filter was designed in a 0.18 μm CMOS process for the data demodulation block of the receiver. An operational amplifier was firstly realized and then a tunable low pass filter was designed based on the amplifier. Detailed simulation results are presented and the main design issues were discussed.

Keywords ETC · Tunable low-pass filter · Operational amplifier

Y. Li (✉) · H. Yu · S. Lin · L. Jiang · R. Wei · Z. Ji
Shenzhen City Key Laboratory of Embedded System Design,
College of Computer Science and Software Engineering,
Shenzhen University, Shenzhen 518060, Guangdong, People's Republic of China
e-mail: liyan@szu.edu.cn

H. Yu
e-mail: yuhang@szu.edu.cn

L. Jiang
e-mail: jianglai@szu.edu.cn

Z. Ji
e-mail: jizhen@szu.edu.cn

12.1 Introduction

The Standardization Administration of the People's Republic of China (SAC) has released the Chinese standard of Dedicated Short Range Communications (DSRC), which set its working frequencies ranging from 5.8 to 5.9 GHz [1]. A 5.8 GHz Radio Frequency (RF) receiver that complies with China's ETC standard is the key component to such a system. In order to simplify system architecture and avoid DC leakage from VCO, direct conversion architecture is applied in the ETC system. In such architecture, the receiver consists of a low noise amplifier (LNA), a mixer, a band-pass filter (BPF), a log-amplifier/received signal strength indicator (Log-Amp/RSSI) and a data demodulator. After the received 5.8 GHz radio frequency signal is converted to a 10 MHz base-band signal by the LNA and the mixer, the BPF is used to extract the useful signal. The signal is then amplified by the Log-Amp/RSSI block and the data demodulator is necessary to correctly recover the input data.

According to [1], amplitude-shift keying (ASK) will be used as the data modulation form in the Chinese ETC system and the data rate will be up to 2 Mbit/s. Fig. 12.1 gives a general structure for ASK data demodulator, which is composed of a low-pass filter (LPF), a data slicer and a comparator. Because the output signal of the Log-Amp/RSSI block is 10 MHz, the LPF plays an important role to the overall performance of the data demodulator.

An operational amplifier based low-pass filter was design in this work. Comparing with chebyshev and bessel filters, butterworth filters have a most flat amplitude response and a good compromise between gain, phase and delay. Thus the butterworth type filter was chosen in this work. As the data rate is far below the intermediate frequency (10 MHz) in the base band for signal processing, a 3rd order filter is enough to correctly extract the data. The chapter is organized as follows: in Sect. 12.2, a CMOS operational amplifier was firstly designed. The design issues of its swing, gain and bandwidth are discussed in detail. The filter design is given in Sect. 12.3. A cascade 3rd order butterworth low-pass filter was realized using the operational amplifier. Detailed simulation results are given in Sect. 12.4 and the conclusion is given in Sect. 12.5.

12.2 Operational Amplifier Design

A cascode two-stage structure is used in our design. Because the signal from the Log-Amp/RSSI block is in differential mode while the signal required by the data slicer is single-ended, a differential pair that achieves the differential to single-ended conversion is chosen. Active current load is used in this stage in order to achieve a high gain. The second stage uses a common-source block with an R-C compensation network. The CMOS operational amplifier circuit is given in Fig. 12.2.

Fig. 12.1 The structure of data demodulator

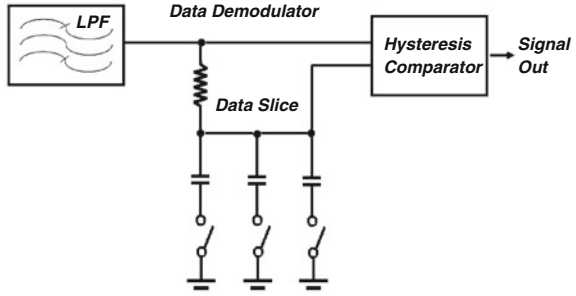
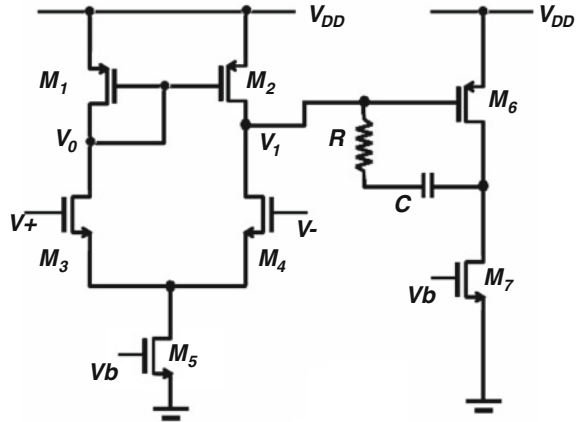


Fig. 12.2 The two-stage cascade CMOS operational amplifier



The first stage of the operational amplifier is a differential pair with active current loads that converts the input differential signal to single-ended. Assume that the input common mode voltage is V_{in} , the transistors M_3 and M_4 will be turn on when V_{in} is larger than the threshold voltage of M_1 and M_2 . If the bias current is I_{ss} , the minimum value of V_{in} that ensures all the transistors to work in the saturation region is:

$$V_{in} = \sqrt{\frac{I_{ss}}{\mu_n C_{ox} \left(\frac{W}{L}\right)_1}} + V_{TH} + \sqrt{\frac{2I_{ss}}{\mu_n C_{ox} \left(\frac{W}{L}\right)_5}}, \quad (12.1)$$

where V_{TH} is the threshold voltage of M_3 and M_4 , $(W/L)_1$ and $(W/L)_5$ are the width-to-length ratio of M_1 and M_5 , respectively.

As V_0 and V_1 decreases when V_{in} increases, the transistors M_3 and M_4 will enter the triode region if V_{in} exceeds a certain value. When M_1 is in the saturation region, we have the drain voltage V_0 and V_1 of the transistors M_3 and M_4 :

$$V_0 = V_1 = V_{DD} - |V_{GS1}| = V_{DD} - \sqrt{\frac{I_{ss}}{\mu_p C_{ox} \left(\frac{W}{L}\right)_1}} - |V_{TH1}|, \quad (12.2)$$

where V_{TH1} is the threshold voltage of M_1 and M_2 .

Thus, the value of V_{in} should satisfy the following equation:

$$V_{in} = V_{DD} - \sqrt{\frac{I_{ss}}{\mu_p C_{ox} \left(\frac{W}{L}\right)_1}} - |V_{TH1}| + V_{TH3}, \quad (12.3)$$

where V_{TH3} is the threshold voltage of M_3 and M_4 .

The maximum value of V_{in} is then defined by (12.3). If the value of V_{in} is chosen between the minimum and the maximum value, all the MOS transistors of the first stage are in the saturation region. The differential mode small signal gain of the first stage can be estimated using (12.4):

$$A_{DM} = g_{m4}(r_{02} \parallel r_{04}), \quad (12.4)$$

where g_{m4} is the transconductance of M_4 , r_{02} and r_{04} are the channel resistance of M_2 and M_4 , respectively.

The common mode gain and the corresponding CMRR value can be estimated in (12.5) and (12.6), which are given below

$$A_{CM} \approx \frac{-\frac{1}{2g_{m1,2}} \parallel \frac{r_{01,2}}{2}}{\frac{1}{2g_{m3,4}} + r_{05}} = -\frac{1}{1 + 2r_{05}g_{m3,4}} \frac{g_{m3,4}}{g_{m1,2}} \quad (12.5)$$

$$CMRR = \left| \frac{A_{DM}}{A_{CM}} \right| = (1 + 2r_{05}g_{m3,4}) \frac{g_{m1,2}g_{m4}(r_{02} \parallel r_{04})}{g_{m3,4}}, \quad (12.6)$$

where $r_{01,2}$ is the channel resistance of M_1/M_2 , r_{05} is the channel resistance of M_5 , $g_{m1,2}$ is the transconductance of M_1/M_2 and $g_{m3,4}$ is the transconductance of M_3 or M_4 .

The second stage has an R-C compensation configuration. The capacitor C provides Miller compensation, which moves the inter stage pole towards the origin and the output pole away. This configuration allows a much greater bandwidth than that obtained by merely connecting the compensation capacitor from one node to ground [2]. However, the right half plane zero in two-stage CMOS operational amplifier, given by $g_{m6}/(C + C_{GD6})$, is a serious issue because the C is chosen large enough to position the dominant pole properly. Thus, placing a resistor R in series with capacitor C is to compensate the impact of the zero frequency. For pole-zero cancellation to occur, it is required:

$$R = \frac{C + C_L}{g_{m6}C}, \quad (12.7)$$

where g_{m6} is the transconductance of M_6 and C_L is the load capacitors of the circuit.

The small signal gain of the whole operational amplifier can be estimated as follows:

$$\frac{V_{out}}{V_{in}} = g_{m4}(r_{02} \parallel r_{04}) \cdot \frac{\frac{1}{R + \frac{1}{sC}} - g_{m6}}{\frac{1}{r_0} + \frac{1}{R + sC}} \quad (12.8)$$

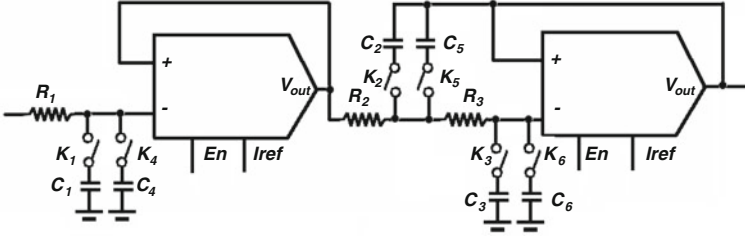


Fig. 12.3 Cascode tunable low-pass filter architecture

Table 12.1 Capacitor values of the 3rd low-pass filter

Cut-off frequency (MHz)	C_{A1} (pF)	C_{A2} (pF)	C_{A3} (pF)
1	15.9	31.9	7.96
2	7.96	15.9	3.98

$C_{A1} = (K_1C_1 + K_4C_4)$, $C_{A2} = (K_2C_2 + K_5C_5)$, $C_{A3} = (K_3C_3 + K_6C_6)$

12.3 Opam-Based 3rd Order Tunable Low-Pass Filter Design

By biquad synthesis method, the 3rd order low-pass filter is built by cascading one 1st order block and one 2nd order block. A programmable capacitor array is added to tune the cut-off frequency. The proposed architecture is given in Fig. 12.3.

As given in Fig. 12.3, three pairs of capacitor, three resistors and two operational amplifiers are used. The cut-off frequency is chosen by the states of the switches K_1 to K_6 . When K_1 , K_2 and K_3 are turned on, the cut-off frequency is set to 1 MHz. Otherwise, when K_4 , K_5 and K_6 are turned on, a 2 MHz bandwidth is achieved.

The transfer function of the first and the second stages are given in (12.9) and (12.10), respectively:

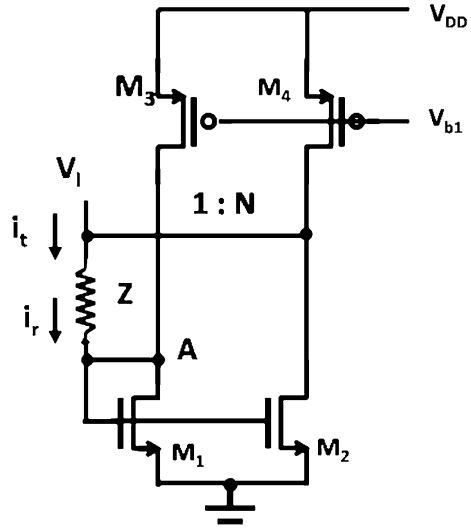
$$H_1(\omega) = \frac{1}{SR_1(K_1C_1 + K_4C_4) + 1} \tag{12.9}$$

$$H_2(\omega) = \frac{1}{S^2(K_2C_2 + K_5C_5)(K_3C_3 + K_6C_6)R_2R_3 + S(K_3C_3 + K_6C_6)(R_2 + R_3) + 1} \tag{12.10}$$

The values of all the resistances are fixed to 10 KΩ and the normalized values of different capacitors were firstly calculated thanks to the pole positions of the 3rd order butterworth function [3]. Then, by using the denormalization factor, which is defined as $C/(2f_cR)$, the real values can be obtained. Table 12.1 summarizes all the values.

In order to reduce the capacitor surface, impedance scaled down technique is used [4]. The principle is illustrated in Fig. 12.4. As the equivalent ac impedance

Fig. 12.4 Impedance scaling



of M_1 is $1/gm$ (seen from point A), the total impedance seen from V_1 is $Z + 1/gm$. When the (W/L) ratio of M_1 is large enough, the ac current flowing through Z is $i_r = Vi/Z$. Since the (W/L) ratio between M_1 and M_2 is $1/N$, the total current at the input node should be: $(N + 1) i_r$. Therefore, an effective capacitance equaling $(N + 1) C$ is obtained.

12.4 System Implementation and Validation

The whole system was implemented using TSMC 0.18 μm CMOS technology and the system performance is fully studied under the Zeni design environment.

Under different values of bias current I_{ss} , the DC voltage transfer characteristic of the first stage was first studied. The results are given in Fig. 12.5.

In the first stage, when all the MOS transistors are in the saturation region, the value of current drawn through M_1 and M_2 is equal, and therefore, the voltage of the first stage output terminal is a constant value. From (12.2), the output voltage value is controlled by the bias current. In our studies, the bias currents of 20 (A), 40 (B), 60 (C) and 80 μA (D) were studied. Curve B is the best result because the second stage needs a 1.2 V input voltage in order to work in the saturation region. The differential small signal gain of the operational amplifier was also studied under the above bias currents and the results are given in Fig. 12.6.

As shown in Fig. 12.6, the highest gain is obtained with a bias current of 40 μA . The reason is that the output voltage of the first stage is 1.2 V under such a condition, which sets the second stage in saturation.

The gain and phase responses of the 3rd order filter are given in Figs. 12.7 and 12.8, respectively. The design requirements are achieved.

Fig. 12.5 The first stage output voltage transfer characteristic

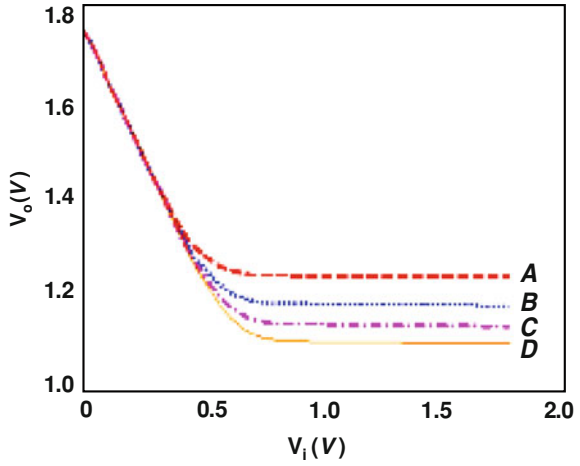


Fig. 12.6 AC simulation results of the operational amplifier

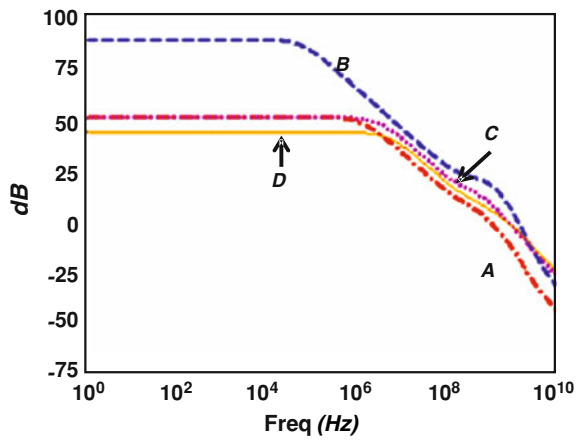


Fig. 12.7 AC simulation results of the low pass filter: gain response

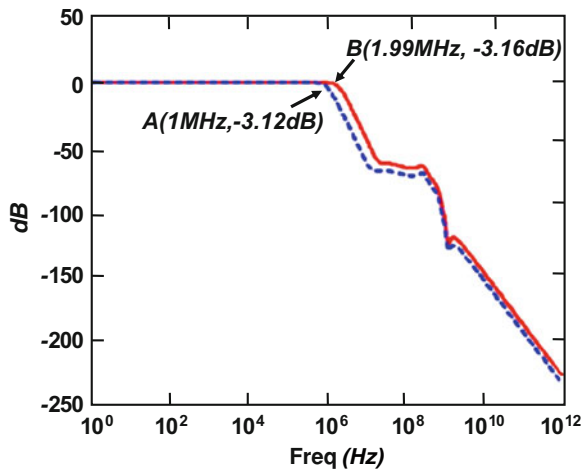
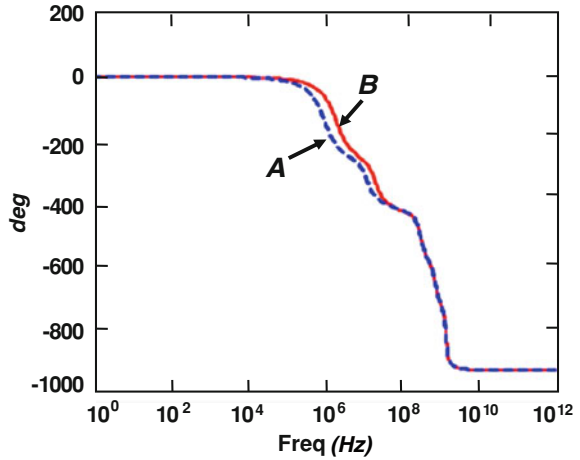


Fig. 12.8 AC simulation results of the low pass filter: phase response



12.5 Conclusion

A 3rd order low-pass butterworth filter for the data demodulator block of the 5.8 GHz ETC receiver was designed in a 0.18 μm CMOS technology. An operational amplifier of a 75 dB gain is firstly designed and the filter is then realized using a two-stage biquad structure based on it. Impedance scaled down technique is used for surface saving. Simulation results show that the system satisfies the design requirements.

Acknowledgments This work is partially supported by the project 60901016 supported by NSFC, the project 10151806001000016 supported by Guangdong R/D Fund, the project JC201005280477A supported by Shenzhen city and the SRF for ROCS, SEM.

References

1. GB/T 20851-2007 (2007) Electronic toll collection-dedicated short range communication interface with roadside unit and lane controller 06
2. Fan XH, Mishra C, Edgar SS (2005) Single Miller capacitor frequency compensation technique for low-power multistage amplifiers. *IEEE JSSC* 40(3):584–592
3. Winder S (2002) Analog and digital filter design. Newnes, Amsterdam
4. Allen PE, Holberg DR (2002) CMOS analog circuit design. PHEI, Beijing

Chapter 13

Sensor Fusion Using Improvement of Resampling Algorithm Particle Filtering for Accurate Location of Mobile Robot

Xiang Gao and YaPing Fu

Abstract Presented chapter deals with the information fusion through used a Divisional Resampling of Particle Filter algorithm. The Divisional Resampling is based on the Multinomial Resampling and the Stratified Resampling, which divides the random number any arrangement into several sub-interval arranged by ascending. The chapter combined Divisional Resampling algorithm and the feedback control algorithm and then integrated the measurement information from odometer and sonar sensor, so that estimating the state vector of a mobile robot to achieve the aim of accurate location.

Keywords Mobile robot · Particle filter · Sensor fusion

13.1 Introduction

More recently, the Extended Kalman Filter (EKF) is an estimation algorithm that performs optimization in the least mean squares sense [1]. So EKF has been successfully applied to neural networks training and sensor fusion problems. However, EKF algorithm can result to unsatisfactory representations of the non-linear functions.

X. Gao (✉) · Y. Fu
Department of Automation Engineering, Nanjing University of Posts and
Telecommunications, Nanjing, China
e-mail: gaonj@gmail.com

Y. Fu
e-mail: fuyaping1986@126.com

To overcome these shortcomings [2], Particle Filter (PF) is mentioned. The particle filter is an alternative nonparametric implementation of the Bayes filter. In particle filters, the samples of a posterior distribution are called particles and are denoted. $X_t := \{x_t^1, x_t^2, \dots, x_t^M\}$, each of particle x_t^m (with $1 < m < M$) is concrete instantiation of the state at time t. The input of this algorithm is the particle set x_{t-1} , then the most recent control u_t and the most recent measurement z_t .

The algorithm of Particle Filtering can provide optimal estimation in nonlinear non-Gaussian state models and nonlinear models which has improved its performance over the established nonlinear filtering approaches [3]. The basic idea of Particle Filter is based on state space experience conditional distribution of random sample to produce a set of these samples, which is called Particle and then on the basis of measurements constantly to adjust to the weight and position of Particle. In this chapter the Particle Filter has been employed for the localization of mobile robot by fusing data coming from odometer and sonar sensors which are collecting environmental information to realize the position of mobile robot's velocity and angle information, so as to achieve target tracking.

13.2 The Nonlinear Measurement Model Particle Filter

13.2.1 The Divisional Resampling of Particle Filter

The algorithm of Particle Filter has a drawback which is that after some numbers of iteration k, the particles' weight w_k^i will become 0 [4]. In the ideal case all the weights should converge to the value $1/N$. So particles are degenerated, they become unnecessary to modify the algorithm.

The basic idea of Resampling is removing small weight particle, so focused on right big particle particles. At present a wide range of present application of heavy sampling algorithms are Multinomial Resampling Residual Resampling, Stratified Resampling, and Systematic Resampling 7.

This chapter uses a kind of compromise resampling algorithm—the Divisional Resampling.

The steps are as follows:

Step1: $d = \lfloor \sqrt{N} \rfloor, R = N \% d, N$ is the size of the particle collection

Step2: (1) if $R = 0$, then $d = \lfloor \sqrt{N} \rfloor, N^d = N/d, x = 1/N^d$

For $i = 1, \dots, N^d A_i = \left\{ S_j^i \right\}_{1 \leq j \leq d}, S_j^i \sim U(((i-1)x, ix]),$

END

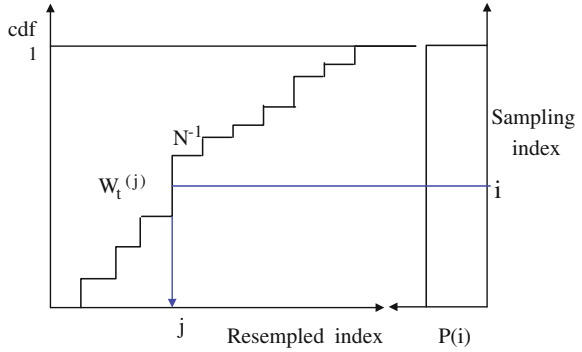
(2) else

$N^d = \lfloor N/d \rfloor + 1, x = (1 - R/N)/(N^d - 1)$

For $i = 1, \dots, (N^d - 1)$

$A_i = \left\{ S_j^i \right\}_{1 \leq j \leq d}, S_j^i \sim U(((i-1)x, ix])$

Fig. 13.1 Resampled index



END

$$A_{N^d} = \left\{ S_j^{N^d} \right\}_{1 \leq j \leq R}, S_j^{N^d} \sim U(((i-1)x, 1])$$

Step3:

$$U = \{U_i\}_{1 \leq i \leq N} = \bigcup_{i=1}^{N^d} A_i, I^i = D_w^{inv}(U^i), \tilde{\zeta}^i = \zeta^{I^i}, \tilde{w}^i = 1/n, i = 1, \dots, n$$

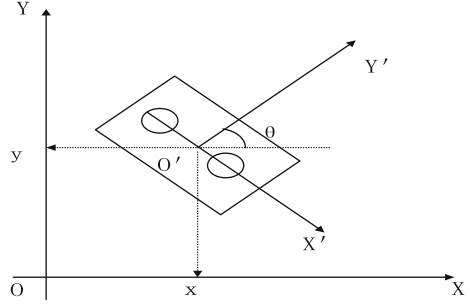
The process of Resampling as follows: after calculating the accumulative of particle probability, i is random sampling index, then we use i onto the probability of accumulative and regional, projection position and indicate the position of accumulative and corresponding domain, j is a new particles index. As shown in figure one shows.

The purpose of resampling is to reduce the number of particles less weight to focus on the larger particle weight. Basic idea is posteriori probability density function of sampling N times again [5]. The posteriori probability density function is $p(x_k|z_{1:k}) \approx \sum_{i=1}^N w_k^i \delta(x_k - x_k^i)$, Where x_k^i is the state of a sample of robot position in the K time. w_k^i is the robot state x_k^i probability in the K time. Resampling N times, resulting in a new set of support points $\{x_k^{i*}\}_{i=1}^N$ of particles, with $p(x_k^{i*}) = p(x_k^i) = w_k^i$. Resampling is an independent distribution, so the weight is resetting to $w_k^i = 1/N$ (Fig. 13.1).

13.2.2 The Prediction Stage

In the prediction stage, we mainly calculate $p(x(k)|Z^-)$ where $x(k)$ is time sequence of states in the k time, $Z^- = \{z(1), \dots, z(n-1)\}$ is observation sequence from 1 to t time. Through calculation of the recursive estimate edge distribution $p(x_k|z_{1:k})$, then there is no need to save the system state historical value.

Fig. 13.2 Mobile robot motion model



13.2.3 Update Stage

Measurement update, given [6, 7] $z(k)$, and compute the new value of the state vector

$$p(x(k)|z) = \sum_{i=1}^N w_k^i \delta(x_k - x_k^i) \tag{13.1}$$

$$w_k^i = \frac{(w_{k-1}^i p(z(k)|x(k) = \zeta_{k-1}^i))}{\sum_{j=1}^N w_{k-1}^j p(z(k)|x(k) = \zeta_{k-1}^j)}$$

$$\zeta_k^i = \zeta_{k-1}^i$$

In the process of practical operation of weights, we need to normalize weights. After for normalization $w_k^{i*} = w_k^i / \sum_{i=1}^N w_k^i$. A posteriori probability density in the system in $k-1$ time is $p(x_k|z_{k-1})$. Time measurement information obtained in the k time, in accordance with this article resampling algorithm selected a random sample of N points, after time and status update, N particles of the posterior density can be approximately considered $p(x_k|z_k)$.

13.3 Simulation Results

13.3.1 Improved Particle

Filter will be applied to mobile robot localization

- (1) Mobile robot motion model

The object of the chapter studied is commonly used in two rounds of indoor robot, shown as follows (Fig. 13.2).

There is an encoder on the driving wheels and it provides a measurement of the incremental angle. The odometer sensor is used to obtain an estimation of the robot's angular velocity $w(t_k)$ and linear velocity $v(t_k)$. Given a sampling period is $\Delta t_k = t_{k+1} - t_k$.

Mobile robot's in continuous-time kinematic equation is:

$$\begin{aligned} \dot{x}(t) &= v(t) \cos(\theta(t)) & \dot{y}(t) &= v(t) \sin(\theta(t)) & \dot{\theta}(t) &= w(t) \end{aligned} \quad (13.2)$$

Where $(x, y)^T$ point is mobile robot's position and θ is the orientation of robot body rotation movement.

Establish its probability for motion model: $p(l_t | l_{t-1}, u_{t-1})$. The robot in $t-1$ time relative to the global coordinates the pose is $l_{t-1} = (x_{t-1}, y_{t-1}, \theta_{t-1})^T$, After time t , reached the position l_t , then the movement model is:

$$l_t = l_{t-1} + \begin{bmatrix} \cos(\theta_{t-1}) & 0 \\ \sin(\theta_{t-1}) & 0 \\ 0 & 1 \end{bmatrix} u_{t-1} + v_{t-1} \quad (13.3)$$

With: $u_{t-1} = (\delta_{t-1}, \Delta\theta_{t-1})^T$ is the odometer model input, during the time of $(t-1, t)$, δ_{t-1} is robot's center displacement, $\Delta\theta_{t-1}$ is the angle of the robot rotated, v_{t-1} obeyed Gaussian measure noise. Defined the control law 11:

$$\begin{aligned} v &= v_d \cos(\theta_d - \theta) + k_1((x_d - x) \\ &\quad \cos \theta + (y_d - y) \sin \theta) \\ w &= w_d + k_2 v_d \frac{\sin(\theta_d - \theta)}{\theta_d - \theta} \\ &\quad ((y_d - y) \cos \theta - (x_d - x) \sin \theta + k_3(\theta_d - \theta)) \end{aligned} \quad (13.4)$$

k_1, k_2, k_3 is Gain.

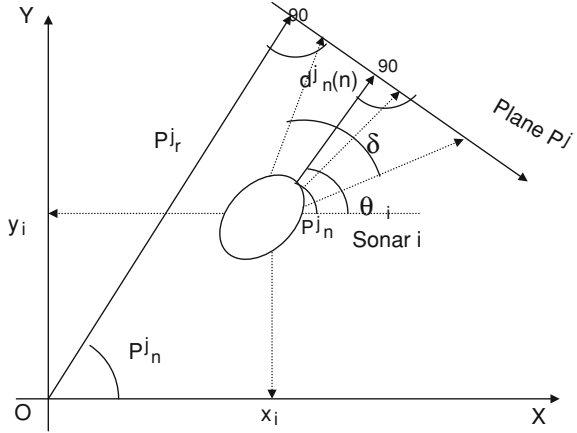
Dynamics of the tracking error:

$$\begin{aligned} e_x &= x_d - x, e_y = y_d - y \\ \ddot{e}_x + K_{d1} \dot{e}_x + K_{p1} e_x &= 0 \\ \ddot{e}_y + K_{d1} \dot{e}_y + K_{p1} e_y &= 0 \end{aligned} \quad (13.5)$$

(2) Coordinate system and the sensor modeling

XOY is the plane the global coordinate system [8], $X'O'Y'$ is XOY rotated θ . This chapter uses odometer and ranging sonar sensors to collect mobile robot information. Mobile robot collects surrounding environment information through the sonar sensors. Sonar through the transmitter, meet obstacles issued pulse waves reflected back, then be receiver receive, the reflection of the Angle can be sonar sensors detected. Robot sonar sensors through the collection of information in the

Fig. 13.3 Sonar position in the coordinate system



environment, Sonar pulsing sound through a transmitter, an obstacle to be reflected back, and then received by the receiver, the angle of reflection can be detected by sonar sensors. Sonar sensors in the coordinate system in the position $XOY (x_i, y_i)$, Fig. 13.3 shows. Its direction angle θ_i , δ is the opening angle with sonar. Sonar feature information extraction environment, it is to get the raw data corresponding to the distance and angle information.

Sonar sensor model shown in Fig. 13.3:

Suppose K sampling time, the mobile robot pose $X(k) = (x(k), y(k), \theta(k))$ [9], after a rotation of the coordinate transformation, to the only sonar sensors i -coordinate system in the coordinate $OXY (x_i, y_i)$ conversion to $O'X'Y'$ coordinate system.

System (x'_i, y'_i) , the sonar sensor model is:

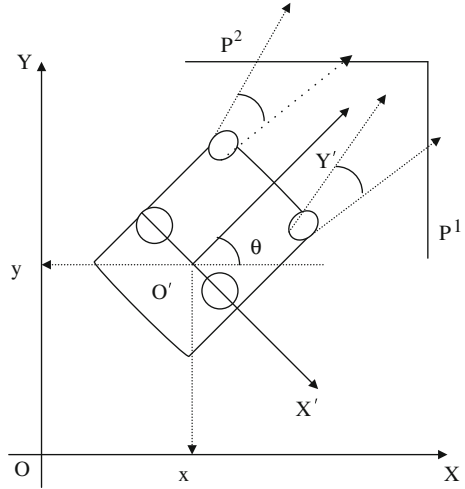
$$\begin{aligned}
 x_i(k) &= x(k) + x'_i \sin(\theta(k)) + y'_i(\cos \theta(k)) \\
 y_i(k) &= y(k) - x'_i(\cos \theta(k)) + y'_i \sin(\theta(k)) \\
 \theta_i(k) &= \theta(k) + \theta_i
 \end{aligned}
 \tag{13.6}$$

Mobile robot wall as reflected in your environment P^j , sonar sensors are available with distance from the planes P_r^j, P_n^j, P_r^j is the distance between the origin O to the plane.

Sonar i to plane P^j distance of obstacles in k time for $d_i^j(k)$: i is the number of sonar sensors, j is the number of planes, Simplified model used in this article, as shown in Fig. 13.3. $i = 1, j = 1$, then $d_i^j(k)$ is: $d_i^j(k) = P_r^j - x_i(k) \cos(P_n^j) - y_i(k) \sin(P_n^j)$ (14)

Where $z(k\Delta t_k)$ is the odometer and sonar sensor measurements [10], $v(k\Delta t_k)$ is a Gaussian white noise sequence $\sim N(0, R(k\Delta t_k))$, $z((k + 1)\Delta t_k) = G(X(k + 1) \Delta t_k) + v(k\Delta t_k)$, $z(k\Delta t_k)$ can be decomposed into two sub-vectors of the form:

Fig. 13.4 Mobile robot odometer and sonar sensor model



$$\begin{aligned}
 z_1(k+1) &= [x(k) + v_1(k), \\
 y(k) + v_2(k), \theta(k) + v_3(k)] \\
 z_2(k+1) &= [d_1^j(k) + v_4(k)]
 \end{aligned}
 \tag{13.7}$$

(Fig. 13.4).

13.3.2 Based on Particle Filter Multi-Sensor

Information fusion for mobile robot localization [11]:

Suppose the robot's initial position in \$OXY\$: \$x(0) = 1m, y(0) = 0m, \theta(0) = 45^\circ\$

In the \$O'X'Y'\$ coordinate system, location of sonar sensors:
 $x'_1 = 0.5m, y'_1 = 0.5m, \theta'_1 = 0^\circ$

Plane position is \$P^1 : p_r^1 = 15.5, p_n^1 = 45.00\$.

State noise: \$w(k) = 0, p(0) = diag[0.1, 0.1, 0.1], R(0) = diag[0.1, 0.1, 0.1]\$.

Kalman filter gain \$K(k) \in R^{3 \times 4}\$

13.4 Simulation Results are as Follows

Figure 13.5, Fig. 13.6

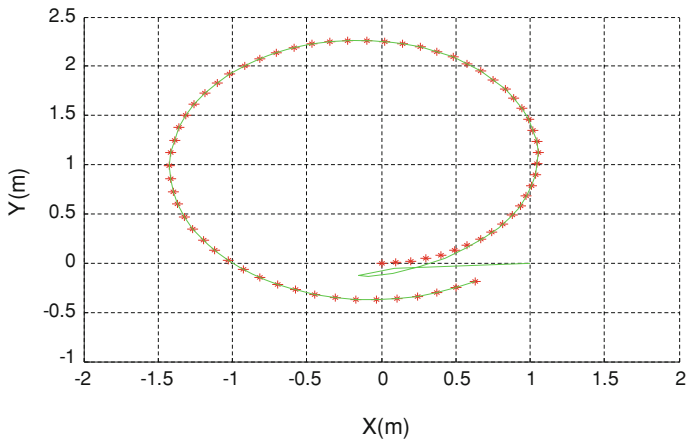
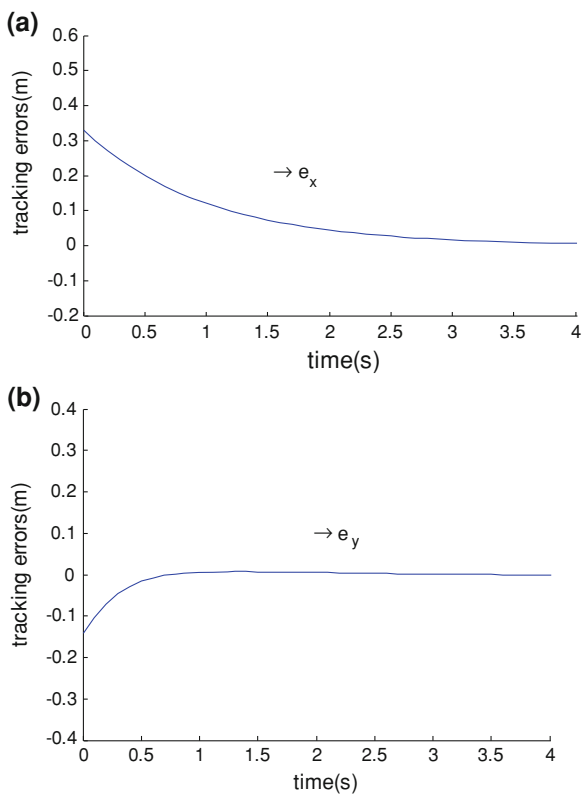


Fig. 13.5 Mobile robot trajectory tracking using particle filter (asterisk * denotes expectations of mobile robot, the green solid line represents the actual robot trajectory tracking)

Fig. 13.6 a Robot distance error e_x . **b** Robot distance error e_y



13.5 Conclusion

The chapter introduces a Divisional Resampling based on Particle Filter algorithm which combines to robot feedback control theory through the fusion of odometer and sonar sensors from collecting environmental information. And applied the chapter introduces a Divisional Resampling based on Particle Filter algorithms to mobile robot to localization and tracking. Simulation results show that based on the particle filter in the Divisional Resampling algorithm, application control theory, will be able to realize the mobile robot for dynamic target effective and accurate tracking. While the future work primarily concentrates in sensor information fusion simplified algorithm, improving the ability of anti-interference and the operation precision, as well as the particle filter algorithm efficiency, robustness and real-time and develop more important particle filter resampling algorithm.

References

1. Ahrens JH, Khalil HK (2007) Closed-loop behavior of a class of nonlinear systems under EKF-based control. *IEEE Trans Autom Control* 52:536–540
2. Rigatos GG (2010) Extended Kalman and Particle Filtering for sensor fusion in motion control of mobile robots. *Math Comput Simul* 81:590–607
3. Von Thrun S, Burgard W, Fox D (2004) *Probabilistic Robotics*
4. Arulampalam S, Maskell SR, Gordon NJ, Clapp T (2002) A tutorial on particle filters for on-line nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans Signal Process* 50:174–188
5. Hua Z (2009) Multi-sensor fusion in mobile robot localization of application. Master's thesis, Wuhan University of Technology
6. Xuan Z (2010) Based on improved particle filter for mobile robot localization. doi:10.1089/ind.2010.6.041
7. Hong Z (2002) Shapes in different environments, real-time tracking. Ph.D. thesis, Institute of Automation Chinese Academy of Sciences
8. Cincione C, Gurrieri GA (1997) A good overview of resampling and related methods [J]. *Soc Sci Comput Rev* 15(1):83–87
9. Wu BC (2006) Particle filter resampling algorithm and its application. Master thesis, Harbin Institute of Technology
10. Fang Z, Tong GF, Xu XH (2007) A robust and efficient mobile robot localization [J]. *Automatica Sinica* 33(1):48–53
11. Rigatos GG (2009) Particle Filtering for state estimation in nonlinear industrial systems. *IEEE Trans Instrum Meas* 58(11):3885–3900

Part II
Intelligence and Biometrics Technologies

Chapter 14

Sensor Placement Modes for Smartphone Based Pedestrian Dead Reckoning

Shahid Ayub, Xiaowei Zhou, Soroush Honary, Alireza Bahraminasab and Bahram Honary

Abstract These days, most of the smartphones come with integrated MEMS (Microelectromechanical systems) sensors like accelerometer, magnetometer, gyros etc. It has opened the ways to use them for location based applications. The big advantage of these sensor based navigation system is that they are environment independent in contrast to other existing positioning technologies. A lot of work is already done on fixed position sensor based systems where sensors are either attached to foot or belt. This work is focussed on developing a smartphone based pedestrian dead reckoning (PDR) system. The big issue with smartphone's sensor based PDR system is that the position of mobile is not deterministic in contrast to fixed position inertial measuring unit. In this work, three different modes of smartphone placement (Idle, Handheld and Listening) are investigated and accelerometer and magnetic sensors are used for step detection and heading determination respectively. Various step detection and stride length estimation methods are implemented and a comparison is given at the end.

Keywords Dead reckoning · Accelerometer · Heading · Step detection · Stride length

S. Ayub (✉) · X. Zhou · S. Honary · A. Bahraminasab · B. Honary
School of Computing and Communications, Lancaster University, Lancaster, UK
e-mail: s.ayub@lancs.ac.uk

X. Zhou
e-mail: xiaowei@hwcomms.com

S. Honary
e-mail: soroush@hwcomms.com

A. Bahraminasab
e-mail: a.bahraminasab@hwcomms.com

B. Honary
e-mail: b.honary@lancs.ac.uk

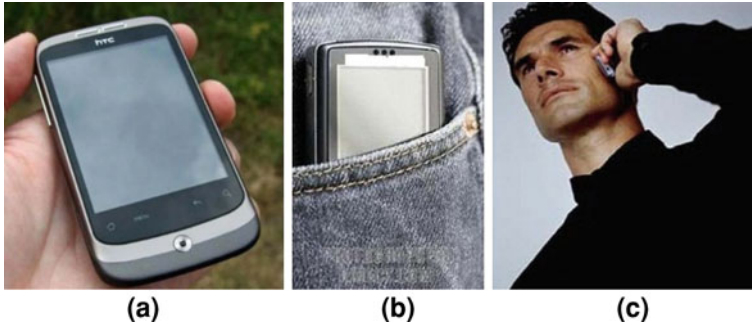


Fig. 14.1 Smartphone placement modes (a) handheld mode (b) idle mode (c) listening mode

14.1 Introduction

Positioning and navigation systems have gained a great success in location based services (LBS), such as personnel security, tracking of assets and people, intelligent guidance, location aware multimedia services etc. In recent years, due to rapid growth in MEMS sensors technology, such as low cost magnetic, barometer, accelerometer, gyros sensors, has opened the ways to use them for various location based services. Since inertial sensors can only provide relative information, so, the integration of inertial sensor based navigation with other positioning technologies is gaining importance to enhance the overall accuracy of the system. This paper presents an aided dead reckoning navigation system for mobile devices.

Dead Reckoning [1] is an estimation of the current position by using a pre-determined start point and then updating the position estimate through knowledge of the acceleration or speed over time and the direction that the person is facing using inertial sensors. Dead Reckoning System based on fixed IMU (Inertial Measuring Unit) [2–4] have been in existence for several years. In fixed position IMU based PDR, sensor block is either attached to shoe [2] or fitted around the belt [4], while in smartphone based systems the position of sensors is not deterministic. So, one has to be careful about the position of sensors as signal pattern may change in different modes. In this work three most common smartphone placement scenarios are discussed under various PDR algorithms as shown in Fig. 14.1. The main idea of PDR mechanization is to use accelerometer signals to detect steps, estimate step length and propagate position using a measured heading. Heading can be computed using gyroscopes or a digital compass.

The structure of this paper is as follows. Section (14.2) describes the fundamentals of Pedestrian Dead Reckoning using smartphones. In Sect. (14.2.1) modes selection is discussed. After that the Signal processing to detect the footsteps and a comparison of three stride length estimation approaches is given in Sects. (14.2.2 and 14.2.3), in Sect. (14.2.4) heading determination is discussed. In Sect. (14.3) Experimental results are shown. Finally, conclusion and future work is discussed in Sect. (14.4).

14.2 Smartphone Based PDR System

These days, most of the smartphones come with integrated sensors like accelerometer, magnetometer and gyros etc. These sensors provide useful information by monitoring the human walking behaviour. The acquired information is used by the Dead Reckoning algorithms to accurately compute the position. In general, there are two methods used to find the displacement by using accelerometer sensor signal, (1) Integration Method and (2) Signal Processing Method. A lot of earlier research is based on finding the displacement by double integration of the signal from the accelerometer [2, 3]. The problem with the integration method is that the error is accumulated in each step and any drift in acceleration results in worse calculation after a few steps. The one way to reduce this drift is the zero velocity updating (ZUPT) at each stance event. ZUPT algorithm performs zero velocity updates every time a step is detected. At foot stance, the velocity is known to be zero, so it is needed to correct the linear velocities obtained after integrating the accelerometer values. In this way, any drift in one step is not propagated further. The best scenario to use the ZUPT algorithm with high precision is when sensor unit is attached to foot to detect the stance event.

Another way to find the linear displacement is the foot step detection method [2] followed by stride length calculation. In this method, accelerometer signal is analysed for footstep detection. A specific pattern is repeated at each foot step. A linear displacement is calculated during each step by using stride length estimation methods. As in case of mobile phone the big issue is that placement of phone is not deterministic and can be varied with time in contrast to foot mounted IMU. So, integration method is not suitable for smartphone based PDR systems. In this work the most common scenarios of mobile placements are discussed. This work is aimed to determine the direction and distance travelled by a person in indoor environment for different smartphone placement modes. The process flow of PDR algorithm can be summarized as, at first basic filtering of accelerometer signal is done then mobile placement mode is selected and after that foot step detection is done. The stride length in each footstep is estimated and finally heading is determined. Figure 14.2 shows the block diagram of the algorithm flow of smartphone based PDR system, the following subsections describe the each block of the flow chart in detail.

14.2.1 Mode Filtering

In this work, the following mobile phone placement modes are investigated.

- (1) Idle Mode (Placed in trouser pocket)
- (2) Handheld Mode (Placed in hand)
- (3) Call listening mode (Placed near to ear)

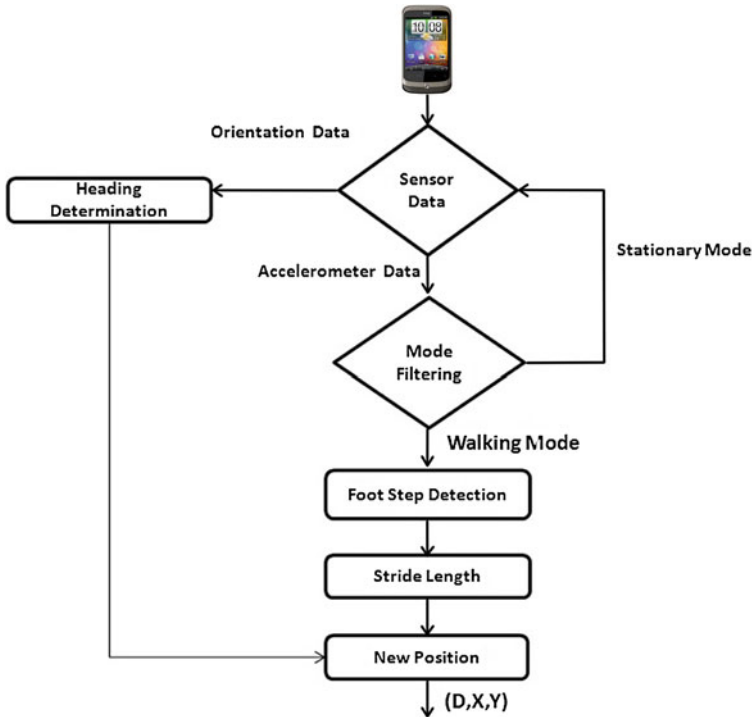


Fig. 14.2 PDR algorithm for smart phones

At first filtering is done to differentiate between the walking and non-walking modes of the mobile. One approach to differentiate between the two is to calculate the difference between the two consecutive windows of acceleration samples, if it is greater than some threshold level then samples are selected for foot step detection process.

$$T < \sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2} \quad (14.1)$$

Modes are selected by observing the average of acceleration along each axis. In idle mode, the average value along x and z axis is near to zero and along y axis it is near to g. In handheld mode, the average values of x and y axis are near to zero and z axis is near to g as shown in Fig. 14.3. In call listening mode, the average value along z axis is zero and along x axis it is near to 0.7 g and along y axis near to 0.7 g. So, by careful observation of accelerometer data, we can differentiate between one of these modes and any other mobile movement. If the observations follow the trend of any of the above three walking modes the data is selected for further processing to calculate the new position, otherwise discarded. The next step is to perform the footstep detection according to the specific mode. Once the data is selected for



Fig. 14.3 Smart phone’s sensors

specific mode the signal processing of the acceleration data is performed to make it suitable for footstep detection process. The signal obtained from the accelerometer sensors is first passed through the low pass filter to remove the fluctuations from the data and to make it smooth. The frequency of the data signal obtained from the sensor is 25 Hz. Interpolation is done to increase the rate of samples. After interpolation, resampling process is performed at 50 Hz. The resultant magnitude of acceleration is used further for step detection. By selecting the resultant value for foot step detection one can resolve the problem of tilting the mobile while walking, as resultant will remain same whether the mobile is tilted or not.

14.2.2 Footstep Detection

A zero crossing detector is implemented in this work. It has been observed from experiments that acceleration variance and peak-to-peak difference of acceleration have a trend to increase with walking frequency. Based on these observations, moving variance method is also implemented for comparison.

Zero Crossing Detector During walking mode a specific pattern of acceleration is repeated in each step. It has been observed by looking at this pattern that the signal crosses the zero level twice in each step as shown in Fig. 14.4a. So, one way to calculate the number of steps is to count the zero crossings and dividing it by two to calculate the number of footsteps. One condition that needs to be fulfilled is that the number of samples between two crossings should be within certain thresholds. If they are greater than maximum threshold or less than minimum

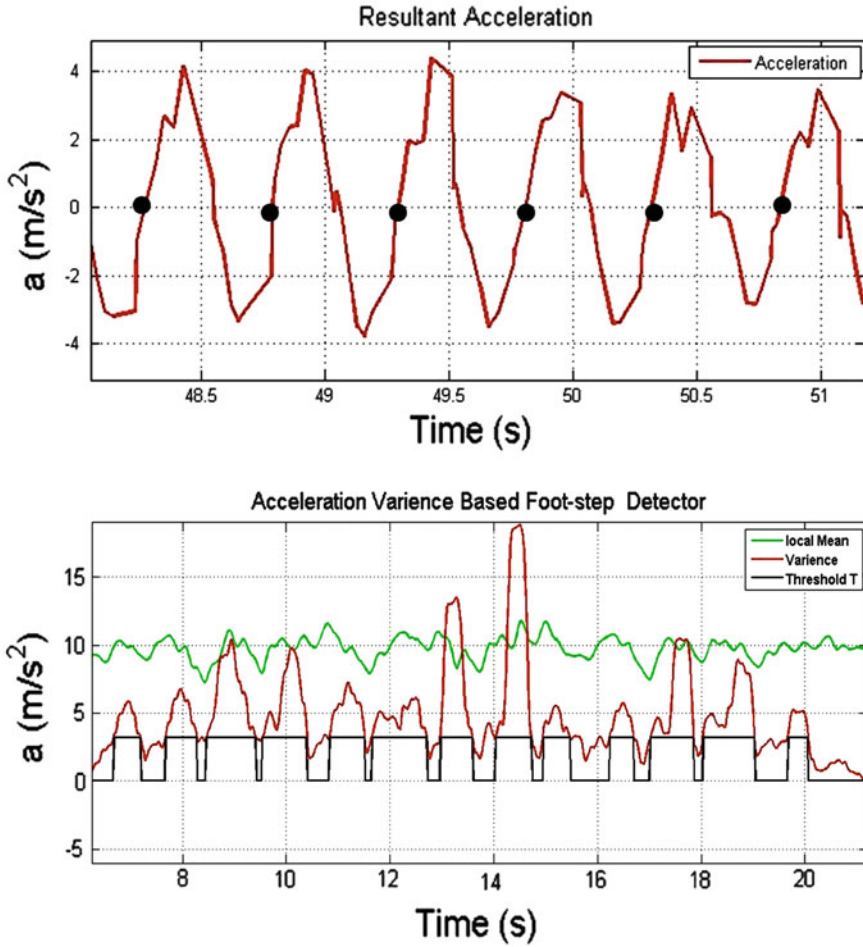


Fig. 14.4 Foot step detectors (a) zero crossing method (b) acceleration variance method

threshold, step is not counted. It has been observed from foot step detection process that number of steps have a trend to decrease with increasing walking frequency for a fixed distance.

Moving Variance Detector It has been also observed from experiments that acceleration variance has a trend to increase w.r.t the stride length, while keeping in view this thing the Moving Variance Filter [3] is implemented. At first the magnitude of every sample is calculated using x, y, z components of acceleration. By selecting the resultant magnitude of acceleration for processing, the problem of tilt can be solved because the pattern of resultant value will be same whether the mobile is tilted or not. Then local mean acceleration \bar{a}_n is calculated for each sample n of resultant acceleration, after that a moving variance filter is applied to make the foot activity more prominent.

$$\sigma_n^2 = \frac{1}{2 * w + 1} \sum_{j=n-w}^{n+w} (a_n - \bar{a}_n)^2$$

where w is the window around a sample n , σ^2 is the moving variance. The last step is to apply a variance threshold to detect a step. A step is detected in sample n when acceleration variance is above a certain threshold level. Figure 14.4b shows the results after applying variance filter and number of footsteps detected. The modes can also be differentiated from each other also by observing the acceleration signal. As in case of idle mode, the position of the phone is assumed to be in the pocket of trousers, only the accelerations of one leg are recorded. In this case, the number of steps taken is multiplied by two in contrast to handheld mode; the same is applied to listening mode.

14.2.3 Stride Length Measurement

After step detection, the stride length is determined to estimate the actual distance travelled. The one way to find [1] the step size is to keep the distance travelled during each step constant according to height and sex of the person. The static step size [1] can be calculated using $Stepsize = height * k$, where k is different for different sex. It is only a rough approximation, because the step size is not a constant value but related to walking speed and acceleration magnitude. It has been observed that as step frequency increases the number of steps decreases and peak acceleration difference increases. In human walking behaviour, a period becomes shorter, a stride becomes larger and vertical impact becomes bigger as the walking speed increases. To increase the accuracy, the step size should be determined continuously while the steps are taken. There are different methods for stride length calculation, but in most of them the accelerometer was attached to foot, which results in inaccurate step size than when it is put somewhere else. The following stride length calculation approaches are compared in this paper to get the better estimate of stride length.

1st Approach This approach is based on Weinberg algorithm [2] which states that the vertical bounce in an individual's step is directly correlated to that person's stride length. This bounce is calculated in the form the difference of the peaks at each step. Stride length is calculated using this filtered signal by using the following equation.

$$SL = k * \sqrt[4]{a_{max} - a_{min}}$$

Peak values are calculated in a window of size w around the sample corresponding to the step detection. k is constant and is determined experimentally or through calibration. Calculating the distance using only the peak values gives widely varying results when walking at different paces or different persons walk.

2nd Approach In [5], a simple solution for step size is determined. There is correlation between the maximum and minimum values, average of acceleration and Step length.

$$SL = k * \frac{1}{N} \sum_{i=1}^N a_i - a_{\min}/a_{\max} - a_{\min}$$

where N is the number of samples, a_{\max} and a_{\min} are the maximum and minimum acceleration in a specific window for step detection.

3rd Approach In [4] a relationship is proposed between mean of acceleration and stride length during a step.

$$SL = 0.98 * \sqrt[3]{\sum_{i=1}^N a_i/N}$$

14.2.4 Heading Determination

A navigation co-ordinate frame is the locally levelled geographic frame with its x-axis pointing east, y-axis pointing north and z-axis pointing up as shown in Fig. 14.3. Smartphones also provides an orientation vector that is used to detect the orientation of the phone in space. The orientation is given through three angles yaw, pitch and roll. When the axes of mobile phone are aligned with geographic frame, then all the three angles should be zero. In the case where the mobile is in idle or handheld mode, Azimuth value can be used to get the direction of movement. In listening mode, the direction can be determined approximately by shifting the azimuth value by 90°.

14.3 Experimental Results

In this work, three mobile placement modes are investigated for pedestrian navigation and a comparison of foot step detection and stride length estimation algorithms is presented. The tests were conducted by walking along an L shape corridor in a building recording the acceleration and orientation with the smartphone in each of the three modes. Figure 14.5 shows the actual path and trajectory estimated from experiment along with floor plan of the building where measurements took place, the red star is the start point, which is already known, blue star is the end point, and total length is 25 m. For each of three modes several recordings were taken and it has been observed that results of smartphone based PDR system are acceptable, though they are not as accurate as fixed position based systems.

Fig. 14.5 A trajectory estimated from experiment along with floor plan of the building where measurements took place, the *red star* is the start point, *blue star* is the end point and total length is 25 m

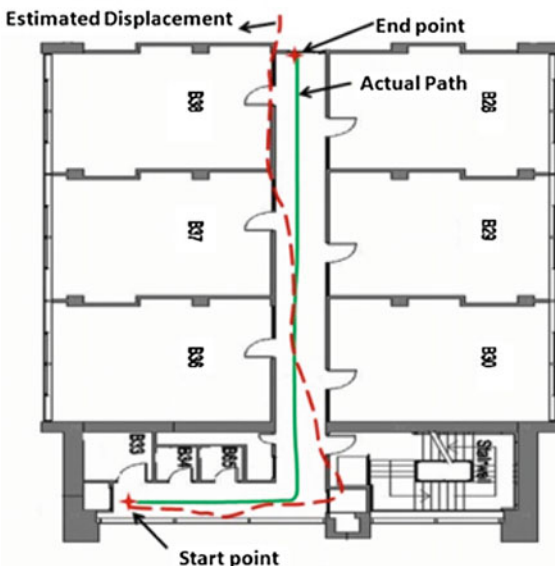


Table 14.1 Comparison of step detection and stride length estimation methods

Methods		Idle mode (%)	Handheld mode (%)	Call listening mode (%)
Stride length estimator	1st approach	5.24	6.54	14.54
	2nd approach	3.55	5.77	9.5
	3rd approach	6.82	7.6	15.3
Step detectors	Zero crossing	1.6	1.5	2.4
	Variance detector	3.5	4	5

Table (1) shows the percentage error between the actual distance and distance estimated using Stride length approaches for each mode Table 14.1. It has been observed that approach (2) for stride length calculation provides more accurate estimation than other approaches. Moreover it is observed from the results that zero crossing detector is more robust than moving variance step detection method. The results of step detection process show the percentage error between the actual number of steps and steps calculated using each detector.

14.4 Conclusion and Future Work

In this work three different mobile placement scenarios are discussed to make it ensure that smartphones can be used well for pedestrian dead reckoning systems. While knowing an initial point, an estimated path is measured which is almost

close to actual trajectory. In addition a comparison of step detection and stride length determination methods is performed to get better estimate. There are further tasks that can be extended to this work to make smartphone based PDR system more reliable. It has been observed that heading is affected by magnetic interference in indoor environments, in future a combination of gyros and magnetic sensors can be used for an accurate heading. Further this system can be made more robust by finding more reliable method of differentiating between different modes. In this work, a simple scenario of L shaped path is tested on a leveled surface; it can be extended to other scenarios like up and down stairs and running etc. Step detection method can also be made more precise by eliminating unnecessary components from acceleration signal.

References

1. Paul D (2008) *Groves, principles of GNSS, inertial and multi sensor integrated navigation systems*. Artech House, Boston
2. Jimenez AR, Seco F, Prieto C, Guevara J (2009) A comparison of pedestrian dead reckoning algorithms using a low cost MEMSIMU. In: *IEEE international symposium in intelligent signal processing*, Budapest, Hungary, August 2009
3. Skog I, Nilsson J-O, Handel P (2010) Evaluation of zero velocity detectors for foot-mounted inertial navigation systems. In: *2010 international conference on indoor positioning and indoor navigation (IPIN)*, Zurich, Switzerland
4. Kim JW, Jang HJ, Hwang D-H, Park C (2004) A Step, Stride and Heading Determination for the Pedestrian Navigation System. *J Glob Position Sys* 3(12):273–279
5. Scarlet J, Enhancing the performance of pedometers using a single accelerometer

Chapter 15

A New Methodology of Judging the Observability of the System

Xuehan Cheng and Yunxia Gao

Abstract The controllability and observability of a linear system are the two basic concepts of linear system, in order to judge the observability of linear system:

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ y(t) = Cx(t) \end{cases}$$
, this chapter bring forward a form of matrix decomposition through primary transform, and thus come up with a new methodology of judging the observability of linear system.

Keywords Primary transform · Matrix decomposition · Linear system observability

15.1 Introduction

The controllability and observability of a linear system are the two basic concepts of linear system, In order to judging the controllability and observability of linear system:

This work was supported by the National Natural Science Foundation of China (11001167).

X. Cheng (✉) · Y. Gao
Mathematics and Applied Mathematics School of Mathematics and Information of
Ludong University, Yantai, People's Republic of China
e-mail: chengxuehan@126.com

Y. Gao
e-mail: suiyuan19882007@126.com

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ y(t) = Cx(t) \end{cases} \quad (15.1)$$

(15.1), several different measures have been concluded in Essay [1], recently, Essay [2, 3] bring forward a form of matrix $\begin{pmatrix} A & B \\ C & \end{pmatrix}$ decomposition through primary transformation, through which could directly achieve some concepts such as infinite point, and then could judge the controllability and observability of linear system. This article is on the basic of primary transform of matrix and rises up a new form of correlation efficient decomposition to matrix $A \in C^{n \times n}$, $C \in C_{r_0}^{m \times n}$, through this decomposition can directly judge the observability of linear system (15.1), and then achieve a simple method of judging the observability of linear system through the proof of theorem.

Firstly, giving the lemma which this essay needed:

Lemma 1 [4] *For any matrix $A \in C^{m \times n}$, there are always a invertible matrix $P \in C_m^{m \times m}$, $Q \in C_n^{n \times n}$, satisfy $PAQ = \begin{pmatrix} I_{r_0} & 0 \\ 0 & 0 \end{pmatrix}$*

Lemma 2 [1] *The necessary and sufficient condition which the linear system*

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ y(t) = Cx(t) \end{cases} \text{ is observable is } \text{rank} \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix} = n$$

15.2 The Correlation Efficient Decomposition to Matrix and the Judgement of the Observability of Linear System

Theorem 1 *For any matrix $A \in C^{n \times n}$, $C \in C_{r_0}^{m \times n}$, there are always a invertible matrix $P \in C_n^{n \times n}$, $Q \in C_m^{m \times m}$, satisfy $QCP = \begin{pmatrix} I_{r_0} & 0 \\ 0 & 0 \end{pmatrix}$*

$$P^{-1}AP = \begin{pmatrix} A_1 & C_1 & 0 & \cdots & 0 \\ A_2 & 0 & C_2 & \cdots & 0 \\ \vdots & \vdots & & & \\ A_k & 0 & 0 & \cdots & C_k \\ A_{k+1} & 0 & 0 & \cdots & 0 \end{pmatrix} \quad (15.2)$$

or

$$P^{-1}AP = \begin{pmatrix} A_1 & C_1 & 0 & \cdots & 0 & 0 \\ A_2 & 0 & C_2 & \cdots & 0 & 0 \\ \cdots & & & & & 0 \\ A_k & 0 & 0 & \cdots & C_k & 0 \\ A_{k+1} & 0 & 0 & \cdots & 0 & 0 \\ A_{k+2} & 0 & 0 & \cdots & 0 & M \end{pmatrix} \tag{15.3}$$

In which, A_1 is the phalanx which the rank is r_0 , $C_s = \begin{pmatrix} I_{r_s} \\ O_{(r_{s-1}-r_s) \times r_s} \end{pmatrix}$, M is the phalanx which the rank is $n - \sum_{i=0}^k r_i$,

$$r_s = \text{rank} \begin{pmatrix} C \\ CA \\ \vdots \\ CA^s \end{pmatrix} - \text{rank} \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{s-1} \end{pmatrix} \tag{15.4}$$

Proof For any matrix $C \in C_{r_0}^{m \times n}$, there is a invertible matrix $Q_1 \in C_m^{m \times m}$, $P_1 \in C_n^{n \times n}$, satisfy $Q_1CP_1 = \begin{pmatrix} I_{r_0} & 0 \\ 0 & 0 \end{pmatrix}$ when $r_0 = n$ or $r_0 = 0$, let $P = P_1$, $Q = Q_1$, the decomposition was established. Otherwise, let $P_1^{-1}AP_1 = \begin{pmatrix} A_{11}^{(1)} & A_{12}^{(1)} \\ A_{21}^{(1)} & A_{22}^{(1)} \end{pmatrix}$, in which, $A_{12}^{(1)} \in C_{r_1}^{r_0 \times (n-r_0)}$, then follows the discussions below:

- (1) when $A_{12}^{(1)} = 0$, let $P = P_1$, $Q = Q_1$, then the decomposition (15.3) is established.
- (2) when $A_{12}^{(1)} \neq 0$, there is a invertible matrix M_1 which the rank is r_0 and N_1 which the rank is $m - r_0$, satisfy $M_1A_{12}^{(1)}N_1 = \begin{pmatrix} I_{r_1} & 0 \\ 0 & 0 \end{pmatrix}$. Let

$$N_1^{-1}A_{22}^{(1)}N_1 = \begin{pmatrix} A_{11}^{(2)} & A_{12}^{(2)} \\ A_{21}^{(2)} & A_{22}^{(2)} \end{pmatrix}, \quad \text{in} \quad \text{which,}$$

$A_{11}^{(2)} \in C^{r_1 \times r_1}$, $A_{22}^{(2)} \in C^{(m-r_0-r_1) \times (m-r_0-r_1)}$. Form matrix

$$P_2 = P_1 \begin{pmatrix} M^{-1} & \\ & N \end{pmatrix} \begin{pmatrix} I_{r_1} & 0 & 0 & 0 \\ 0 & I_{r_0-r_1} & 0 & 0 \\ A_{11}^{(2)} & 0 & I_{r_1} & 0 \\ A_{21}^{(2)} & 0 & 0 & I_{n-r_0-r_1} \end{pmatrix}, Q_2 = \text{diag}(M_1, I_{n-r_0})Q_1,$$

Then comes:

$$Q_2CP_2 = \begin{pmatrix} I_{r_0} & 0 \\ 0 & 0 \end{pmatrix}, P_2^{-1}AP_2 = \begin{pmatrix} * & I_{r_1} & 0 \\ * & 0 & 0 \\ * & 0 & A_{12}^{(2)} \\ * & 0 & A_{22}^{(2)} \end{pmatrix}$$

- (3) When $r_1 = n - r_0$, that's matrix $A_{12}^{(1)}$ has full column rank, let $P = P_2, Q = Q_2$, then the decomposition (15.2) was established.
- (4) when $A_{21}^{(2)} = 0$, let $P = P_2, Q = Q_2$, then the decomposition (15.3) was established.
- (5) when $A_{12}^{(2)} \neq 0$, follow step (2)–(4), making similar decomposition to matrix $A_{12}^{(2)}$, until $A_{12}^{(k+1)} = 0$ or $A_{12}^{(k+1)}$ not exist, and when $A_{12}^{(k+1)} = 0$, the decomposition (15.3) was established, when $A_{12}^{(k+1)}$ not exist, the decomposition (15.2) was established.

On the basic of Theorem 1 can directly achieve:

Inference 1 For any matrix $A \in C^{n \times n}, C \in C_{r_0}^{m \times n}$, there always

exist: $rank \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix} = \sum_{i=0}^k r_i$. Then on the basic of Lemma 2 can achieve:

Theorem 2 For any matrix $A \in C^{n \times n}, C \in C_{r_0}^{m \times n}$ which in the linear system (15.1), if there is a invertible matrix $P \in C_n^{n \times n}, Q \in C_m^{m \times m}$, satisfy formula (15.2), then the linear system (15.1) is observable; if it satisfy formula (15.39), that's phalanx M exist, then the linear system (15.1) is unobservable.

15.3 The Algorithm and Examples Which Judge the Observability of Linear System

15.3.1 The Algorithm Judge the Observability of Linear System

Based on the form of the matrix decomposition of Theorem 1, we can judge whether the system is observable through primary transform. Analysis the proof of Theorem 1 further, here we can give a more simple method of judging the observability of linear system (15.1):

The algorithm that judge the observability of linear system (15.1):

- (1) For any matrix $A \in C^{n \times n}, C \in C_{r_0}^{m \times n}$, making elementary column transformation to matrix $\begin{pmatrix} A \\ C \end{pmatrix}$ that can get matrix: $\begin{pmatrix} A_1^{(1)} \\ C_1^{(1)} \end{pmatrix} = \begin{pmatrix} A_2^{(1)} & A_3^{(1)} \\ C_2^{(1)} & 0 \end{pmatrix} \begin{matrix} n \\ m \end{matrix}$, in which, $C_2^{(1)} \in C_{r_0}^{m \times r_0}$, and at the same time, making row transformation to

matrix A (the elementary matrix corresponding to elementary row transformation is the inverse matrix of the one corresponding to elementary column transformation), then partitioning matrix $A_3^{(1)}$, let $A_3^{(1)} = \begin{pmatrix} C^{(1)} \\ A^{(1)} \end{pmatrix}$, in which $C^{(1)} \in C_{r_1}^{r_0 \times (n-r_0)}$, $A^{(1)} \in C^{(n-r_0) \times (n-r_0)}$.

- (2) By the same way, making elementary column transformation to matrix $\begin{pmatrix} A^{(1)} \\ C^{(1)} \end{pmatrix}$, transform $C^{(1)}$ into column echelon matrix $C^{(1)} = \begin{pmatrix} C_2^{(2)} & 0 \end{pmatrix}$, in which $C_2^{(2)} \in C_{r_1}^{r_0 \times r_1}$, and at the same time, making row transformation to matrix $A^{(1)}$ (the elementary matrix corresponding to elementary row transformation is the inverse matrix of the one corresponding to elementary column transformation), then get matrix $\begin{pmatrix} A_1^{(2)} \\ C_1^{(2)} \end{pmatrix} = \begin{pmatrix} A_2^{(2)} & A_3^{(2)} \\ C_2^{(2)} & 0 \end{pmatrix} \begin{matrix} n-r_0 \\ r_0 \end{matrix}$. Partitioning matrix $A_3^{(2)}$, let $A_3^{(2)} = \begin{pmatrix} C^{(2)} \\ A^{(2)} \end{pmatrix}$, in which, $C^{(2)} \in C_{r_2}^{r_1 \times (n-r_0-r_1)}$, $A^{(2)} \in C^{(n-r_0-r_1) \times (n-r_0-r_1)}$.
- (3) Follow the steps until $C^{(s)} = 0$ or $C^{(s)}$ is the full column rank matrix, then stop.
- (4) When $C^{(s)} = 0$, the system is unobservable; when $C^{(s)}$ is the full column rank matrix, the system is observable.

15.3.2 Example

The specific example is listed as follows:

Example 1 Condition coefficient matrix in the linear system are listed as:

$$A = \begin{pmatrix} 2 & 1 & 0 & 2 & 0 \\ 0 & 2 & 1 & 4 & 6 \\ 0 & 0 & 2 & 2 & 8 \\ 1 & 0 & 1 & -3 & 1 \\ 0 & 1 & 1 & 0 & -3 \end{pmatrix}, C = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

Making the judgement about the observability of linear system.

Solution Follow the given algorithm, making primary transformation to matrix A and C :

$$\begin{aligned}
 \begin{pmatrix} A \\ C \end{pmatrix} &= \begin{pmatrix} 2 & 1 & 0 & 2 & 0 \\ 0 & 2 & 1 & 4 & 6 \\ 0 & 0 & 2 & 2 & 8 \\ 1 & 0 & 1 & -3 & 1 \\ 0 & 1 & 1 & 0 & -3 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix} \xrightarrow{C_2-C_1} \begin{pmatrix} 2 & -1 & 0 & 2 & 0 \\ 0 & 2 & 1 & 4 & 6 \\ 0 & 0 & 2 & 2 & 8 \\ 1 & -1 & 1 & -3 & 1 \\ 0 & 1 & 1 & 0 & -3 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix} \\
 &\xrightarrow{r_1+r_2} \begin{pmatrix} 2 & 1 & 1 & 6 & 6 \\ 0 & 2 & 1 & 4 & 6 \\ 0 & 0 & 2 & 2 & 8 \\ 1 & -1 & 1 & -3 & 1 \\ 0 & 1 & 1 & 0 & -3 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix} \dots \xrightarrow{C_4-C_2} \begin{pmatrix} 2 & 1 & 0 & 7 & 14 \\ 0 & 2 & 1 & 4 & 14 \\ 0 & 0 & 2 & 2 & 8 \\ 1 & -1 & 1 & -2 & 1 \\ 0 & 1 & 0 & -1 & -3 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} \\
 &\xrightarrow{r_2+r_4} \begin{pmatrix} 2 & 1 & 0 & 7 & 14 \\ 1 & 1 & 2 & 2 & 15 \\ 0 & 0 & 2 & 2 & 8 \\ 1 & -1 & 1 & -2 & 1 \\ 0 & 1 & 0 & -1 & -3 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} A_1^{(1)} \\ C_1^{(1)} \end{pmatrix}
 \end{aligned}$$

partitioning matrix $\begin{pmatrix} A_1^{(1)} \\ C_1^{(1)} \end{pmatrix}$:

$$\begin{pmatrix} A_1^{(1)} \\ C_1^{(1)} \end{pmatrix} = \begin{pmatrix} 2 & 1 & 0 & 7 & 14 \\ 1 & 1 & 2 & 2 & 15 \\ 0 & 0 & 2 & 2 & 8 \\ 1 & -1 & 1 & -2 & 1 \\ 0 & 1 & 0 & -1 & -3 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} A_2^{(1)} & A_3^{(1)} \\ C_2^{(1)} & 0 \end{pmatrix}. \text{ Partitioning}$$

$$\text{matrix } A_3^{(1)}: A_3^{(1)} = \begin{pmatrix} 0 & 7 & 14 \\ 2 & 2 & 15 \\ 2 & 2 & 8 \\ 1 & -2 & 1 \\ 0 & -1 & -3 \end{pmatrix} = \begin{pmatrix} C^{(1)} \\ A^{(1)} \end{pmatrix}.$$

Then by the same way, making transformation to matrix $\begin{pmatrix} A^{(1)} \\ C^{(1)} \end{pmatrix}$:

$$\begin{pmatrix} A^{(1)} \\ C^{(1)} \end{pmatrix} = \begin{pmatrix} 2 & 2 & 8 \\ 1 & -2 & 1 \\ 0 & -1 & -3 \\ 0 & 7 & 14 \\ 2 & 2 & 15 \end{pmatrix} \xrightarrow{C_1 - C_2} \begin{pmatrix} 2 & 2 & 8 \\ -2 & 1 & 1 \\ -1 & 0 & -3 \\ 7 & 0 & 14 \\ 2 & 2 & 15 \end{pmatrix}$$

$$\xrightarrow{r_1 \leftrightarrow r_2} \begin{pmatrix} -2 & 1 & 1 \\ 2 & 2 & 8 \\ -1 & 0 & -3 \\ 7 & 0 & 14 \\ 2 & 2 & 15 \end{pmatrix} \dots \xrightarrow{C_3 - \frac{1}{2}C_2} \begin{pmatrix} -4 & 1 & -\frac{5}{2} \\ 2 & 2 & -7 \\ -1 & 0 & -1 \\ 7 & 0 & 0 \\ 2 & 2 & 0 \end{pmatrix}$$

$$\xrightarrow{r_2 + \frac{1}{2}r_3} \begin{pmatrix} -4 & 1 & -\frac{5}{2} \\ -\frac{7}{2} & 2 & -\frac{25}{2} \\ -1 & 0 & -1 \\ 7 & 0 & 0 \\ 2 & 2 & 0 \end{pmatrix} = \begin{pmatrix} A_1^{(2)} \\ C_1^{(2)} \end{pmatrix}. \text{ Partitioning matrix } \begin{pmatrix} A_1^{(2)} \\ C_1^{(2)} \end{pmatrix}:$$

$$\begin{pmatrix} A_1^{(2)} \\ C_1^{(2)} \end{pmatrix} = \begin{pmatrix} -4 & 1 & -\frac{5}{2} \\ -\frac{7}{2} & 2 & -\frac{25}{2} \\ -1 & 0 & -1 \\ 7 & 0 & 0 \\ 2 & 2 & 0 \end{pmatrix} = \begin{pmatrix} A_2^{(2)} & A_3^{(2)} \\ C_2^{(2)} & 0 \end{pmatrix}$$

Partitioning matrix $A_3^{(2)} : A_3^{(2)} = \begin{pmatrix} -\frac{5}{2} \\ -\frac{25}{2} \\ -1 \end{pmatrix} = \begin{pmatrix} C^{(2)} \\ A^{(2)} \end{pmatrix}.$

Because $C_1^{(3)}$ is the full column rank matrix, the system is observable.

Through this message, we can judge the observability of the system through primary transform of matrix, and through the process of this way, we can see that it is more better than the traditional way.

References

1. Horn RA, Johnson CR (1991) Topics in matrix analysis [M]. The Cambridge University, New York
2. Wei M, Wang Q, Cheng X (2010) Some new result for system decoupling and pole assignment problems [J]. Automatica 46:937-944
3. Wei M, Cheng X, Wang Q (2010) A Canonical decomposition of the right invertible system with application[J]. Slam J Matrix Anal Appl 31(4):1958-1981
4. Zheng D (1990) The theory of linear system (M). Tsinghua University Press, Beijing, (1933.3), (Chinese)

Chapter 16

Criteria Conditions for Generalized Strictly Diagonally Dominant Matrices

Yujing Liu and Li Guo

Abstract Several determination criterions for generalized strictly diagonally dominant matrix are put forward in this chapter, improving the existing correlative conclusions, extending the determination criterions for generalized strictly diagonally dominant matrix.

Keywords Generalized strictly diagonally dominant matrix · Sufficient condition · Nonzero element chain

16.1 Introduction

Generalized strictly diagonally dominant matrices is one of the particular matrices, and it has widespread use in mathematical modelling, mathematical physics and cybernetics. Thus, seeking the decision conditions of generalized strictly diagonally dominant matrices is of great importance, which has certain study in our country. Such as [1–4], Several practical decision conditions are given in the paper, and the judging question of generalized strictly diagonally dominant matrices is further extended.

Y. Liu (✉) · L. Guo

Department of Mathematics, Beihua University, Jilin, 132013, Jilin, China

e-mail: liuyujing3512@163.com

16.2 Definition and Lemma

Here $C^{n \times n}$ denotes the set of n complex square matrices, $A = (a_{ij}) \in C^{n \times n}$, $R_i(A) = \sum_{j \neq i} |a_{ij}|$, $S_i(A) = \sum_{j \neq i} |a_{ji}|$, $i, j \in N = \{1, 2, \dots, n\}$. If $|a_{ii}| > \sum_{j \neq i} |a_{ij}| \quad \forall i \in N$, then A is called strictly diagonally dominant matrices, denote $A \in D$. If there is Positive diagonal matrix X , such that $AX \in D$, then A is called generalized strictly diagonally dominant matrices, denote $A \in D^*$.

This chapter introduces the following notation

$M(A)$ is a comparison matrix of A ,

$$N_1 = \{i \in N | 0 < 2|a_{ii}| \leq \sum_{j \neq i} (|a_{ij}| + |a_{ji}|)\},$$

$$N_2 = \{i \in N | 2|a_{ii}| > \sum_{j \neq i} (|a_{ij}| + |a_{ji}|)\},$$

$$r = \max_{i \in N_2} \left\{ \frac{\sum_{j \neq i} (|a_{ij}| + |a_{ji}|)}{2|a_{ii}|} \right\},$$

$$\delta_i = \frac{\sum_{j \in N_1} (|a_{ij}| + |a_{ji}|) + r \sum_{\substack{j \in N_2 \\ j \neq i}} (|a_{ij}| + |a_{ji}|)}{2|a_{ii}|}, \quad \forall i \in N_2$$

Lemma 1[1] Let $A = (a_{ij}) \in C^{n \times n}$, $B = M(A) + M^T(A)$, if $B \in D^*$, then $A \in D^*$.

Lemma 2[1] Let $A = (a_{ij}) \in C^{n \times n}$, if $i \in N$, we have

$$2|a_{ii}| \geq \sum_{j \neq i} (|a_{ij}| + |a_{ji}|),$$

and when $\bar{J}(A) = \{i \in N | 2|a_{ii}| > \sum_{j \neq i} (|a_{ij}| + |a_{ji}|)\}$, satisfying $\forall i \in N/\bar{J}(A)$, we have nonzero element chain $a_{ir_1}, a_{r_1 r_2}, \dots, a_{r_k t}, i \neq r_1 \neq r_2 \neq \dots \neq r_k \neq t$, such that $t \in \bar{J}(A)$, then $A \in D^*$.

Proof Let $B = M(A) + M^T(A)$, then we know by conditions of lemma B is nonzero element chain diagonally dominant matrix, then $B \in D^*$, and by lemma 1 we get $A \in D^*$.

16.3 Main Result

Theorem 1 Let $A = (a_{ij}) \in C^{n \times n}$, if $i \in N_1$, we have

$$2|a_{ii}| > \sum_{\substack{j \in N_1 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) + \sum_{j \in N_2} (|a_{ij}| + |a_{ji}|) \delta_j, \tag{16.1}$$

then $A \in D^*$.

Proof Let

$$\varepsilon_i \equiv \frac{2|a_{ii}| - \sum_{\substack{j \in N_1 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) - \sum_{j \in N_2} (|a_{ij}| + |a_{ji}|) \delta_j}{\sum_{j \in N_2} (|a_{ij}| + |a_{ji}|)}, \quad \forall i \in N_1 \quad (16.2)$$

when $\sum_{j \in N_2} (|a_{ij}| + |a_{ji}|) = 0$, and denote $\varepsilon_i = +\infty$. By (16.1) we know $\forall i \in N_1, \varepsilon_i > 0$. take a positive number $\varepsilon (\neq +\infty)$, sufficiently small, such that

$$0 < \varepsilon < \min_{i \in N_1} \varepsilon_i \leq +\infty \quad (16.3)$$

and take a $X = \text{diag}\{x_1, x_2, \dots, x_n\}$, in which $x_i = \begin{cases} 1, & i \in N_1, \\ \delta_i + \varepsilon, & i \in N_2. \end{cases}$ X is a plus dominant matrix, clearly. Let $B = M(A) + M^T(A)$, then $D = BX = (M(A) + M^T(A))X = (d_{ij})_{n \times n}$, satisfy to $\forall i \in N_1$, if $\sum_{j \in N_2} (|a_{ij}| + |a_{ji}|) = 0$, That is $\forall j \in N_2, |a_{ij}| + |a_{ji}| = 0$. By (16.1) we obtain

$$\begin{aligned} \sum_{j \neq i} |d_{ij}| &= \sum_{\substack{j \in N_1 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) + \sum_{j \in N_2} (|a_{ij}| + |a_{ji}|) (\delta_j + \varepsilon) \\ &= \sum_{\substack{j \in N_1 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) < 2|a_{ii}| = |b_{ii}| = |d_{ii}|. \end{aligned}$$

If $\sum_{j \in N_2} (|a_{ij}| + |a_{ji}|) \neq 0$, by (16.2) and (16.3), we obtain

$$\begin{aligned} \sum_{j \neq i} |d_{ij}| &= \sum_{\substack{j \in N_1 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) + \sum_{j \in N_2} (|a_{ij}| + |a_{ji}|) (\delta_j + \varepsilon) \\ &< \left(\sum_{\substack{j \in N_1 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) + \sum_{j \in N_2} (|a_{ij}| + |a_{ji}|) \delta_j \right) + \varepsilon_i \sum_{j \in N_2} (|a_{ij}| + |a_{ji}|) \\ &= 2|a_{ii}| = |b_{ii}| = |d_{ii}|. \end{aligned}$$

To $\forall i \in N_2, 2|a_{ii}| > \sum_{j \neq i} (|a_{ij}| + |a_{ji}|)$, therefore

$$2|a_{ii}| - \sum_{j \neq i} (|a_{ij}| + |a_{ji}|) > 0 \quad (16.4)$$

and because of $i \in N_2$, $\delta_j < r < 1$, then

$$\begin{aligned} & \sum_{j \in N_1} (|a_{ij}| + |a_{ji}|) + \sum_{\substack{j \in N_2 \\ j \in i}} (|a_{ij}| + |a_{ji}|) \delta_j \\ & < \sum_{j \in N_1} (|a_{ij}| + |a_{ji}|) + r \sum_{\substack{j \in N_2 \\ j \in i}} (|a_{ij}| + |a_{ji}|) \end{aligned} \tag{16.5}$$

thus, we have

$$\begin{aligned} & \left(\sum_{j \in N_1} |a_{ij}| + |a_{ji}| \right) + \sum_{\substack{j \in N_2 \\ j \in i}} (|a_{ij}| + |a_{ji}|) \delta_j \\ & \quad - \sum_{j \in N_1} (|a_{ij}| + |a_{ji}|) + r \sum_{\substack{j \in N_2 \\ j \in i}} (|a_{ij}| + |a_{ji}|) \\ \varepsilon < 0 < & \frac{\sum_{j \in N_2} (|a_{ij}| + |a_{ji}|) \delta_j}{2|a_{ii}| - \sum_{\substack{j \in N_2 \\ j \in i}} (|a_{ij}| + |a_{ji}|)} \end{aligned} \tag{16.6}$$

So $\forall i \in N_2$, from the condition set δ_i and (16.6), we get

$$\begin{aligned} & |d_{ii}| - \sum_{j \neq i} |d_{ij}| \\ & > \left(\sum_{j \in N_1} (|a_{ij}| + |a_{ji}|) + \sum_{\substack{j \in N_2 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) \delta_j \right) - \left(\sum_{j \in N_1} (|a_{ij}| + |a_{ji}|) + r \sum_{\substack{j \in N_2 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) \right) \\ & + 2\delta_j |a_{ii}| - \left(\sum_{j \in N_1} (|a_{ij}| + |a_{ji}|) + \sum_{\substack{j \in N_2 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) \delta_j \right) \\ & = 2\delta_j |a_{ii}| - \left(\sum_{j \in N_1} (|a_{ij}| + |a_{ji}|) + r \sum_{\substack{j \in N_2 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) \right) = 0. \end{aligned}$$

To sum up, $\forall i \in N, |d_{ii}| > \sum_{j \neq i} |d_{ij}|$.

That is $B \in D^*$, from Lemma 1, we know $A \in D^*$.

Theorem 2 Let $A = (a_{ij}) \in C^{n \times n}$, and $\forall i \in N_1$, we have

$$2|a_{ii}| \geq \sum_{\substack{j \in N_1 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) + \sum_{j \in N_2} (|a_{ij}| + |a_{ji}|) \delta_j.$$

If A satisfies one of the requirements below, $A \in D^*$.

- (i) $J(A) = \{t \in N_1 | 2|a_{tt}| > \sum_{\substack{j \in N_1 \\ j \neq t}} (|a_{tj}| + |a_{jt}|) + \sum_{j \in N_2} (|a_{tj}| + |a_{jt}|) \delta_j\} \neq \emptyset$.

And $\forall i \in N_1/J(A)$, have nonzero element chain $a_{ir_1}, a_{r_1r_2}, \dots, a_{r_k t}$, such that $t \in J(A)$.

(ii) $\forall s \in N_2$, exist $k \in N_2, s \neq k$, such that $a_{sk} \neq 0$, and $\forall i \in N_1/J(A)$.

Have nonzero element chain $a_{ir_1}, a_{r_1r_2}, \dots, a_{r_k t}$, such that $t \in N_2 \cup J(A)$.

Proof (1) If A meets (i)'s condition.

Let $N_1/J(A) = \{i_1, i_2, \dots, i_k\}$ not loss generality, we take $i = i_1 \in N_1/J(A)$.and by (i) have nonzero element chain $a_{ir_1}, a_{r_1r_2}, \dots, a_{r_k t}$, such that $t \in J(A)$.That is

$$2|a_{tt}| > \sum_{\substack{j \in N_1 \\ j \neq t}} (|a_{tj}| + |a_{jt}|) + \sum_{j \in N_2} (|a_{tj}| + |a_{jt}|) \delta_j,$$

$$\text{Now let } 1 > d_t > \frac{\sum_{\substack{j \in N_1 \\ j \neq t}} (|a_{tj}| + |a_{jt}|) + \sum_{j \in N_2} (|a_{tj}| + |a_{jt}|) \delta_j}{2|a_{tt}|} > 0.$$

Multiply matrix $A^{(1)}$ by positive diagonal matrix $D = (1, 1, \dots, d_t, \dots, 1)$, from right, we get $A^{(1)} = (a_{ij}^{(1)})$.

$$2|a_{tt}^{(1)}| = 2|a_{tt}|d_t > \sum_{\substack{j \in N_1 \\ j \neq t}} (|a_{tj}| + |a_{jt}|) + \sum_{j \in N_2} (|a_{tj}| + |a_{jt}|) \delta_j,$$

And $d_t < 1$, $a_{r_k t} \neq 0$, then $a_{r_k t}^{(1)} = a_{r_k t} d_t < a_{r_k t}$.

$$\sum_{\substack{j \in N_1 \\ j \neq r_k}} (|a_{r_k j}^{(1)}| + |a_{j r_k}^{(1)}|) < \sum_{\substack{j \in N_1 \\ j \neq r_k}} (|a_{r_k j}| + |a_{j r_k}|),$$

$$\sum_{j \in N_2} (|a_{r_k j}^{(1)}| + |a_{j r_k}^{(1)}|) < \sum_{j \in N_2} (|a_{r_k j}| + |a_{j r_k}|).$$

We obtain

$$\begin{aligned} 2|a_{r_k r_k}^{(1)}| &= 2|a_{r_k r_k}| \geq \sum_{\substack{j \in N_1 \\ j \neq r_k}} (|a_{r_k j}| + |a_{j r_k}|) + \sum_{j \in N_2} (|a_{r_k j}| + |a_{j r_k}|) \delta_j \\ &> \sum_{\substack{j \in N_1 \\ j \neq r_k}} (|a_{r_k j}^{(1)}| + |a_{j r_k}^{(1)}|) + \sum_{j \in N_2} (|a_{r_k j}^{(1)}| + |a_{j r_k}^{(1)}|) \delta_j. \end{aligned}$$

And let

$$1 > d_{r_k} > \frac{\sum_{\substack{j \in N_1 \\ j \neq r_k}} (|a_{r_k j}^{(1)}| + |a_{j r_k}^{(1)}|) + \sum_{j \in N_2} (|a_{r_k j}^{(1)}| + |a_{j r_k}^{(1)}|) \delta_j}{2|a_{r_k r_k}^{(1)}|} > 0.$$

Multiply matrix $A^{(1)}$ by positive diagonal matrix $D_1 = \text{diag}(1, 1, \dots, d_{r_k}, \dots, 1)$, from right, we get $A^{(2)} = (a_{ij}^{(2)})$. By analogy, using right multiply $k + 1$ positive diagonal matrix, we have,

$$2|a_{r_{k-1}r_{k-1}}^{(2)}| > \sum_{\substack{j \in N_1 \\ j \neq r_{k-1}}} (|a_{r_{k-1}j}^{(2)}| + |a_{jr_{k-1}}^{(2)}|) + \sum_{j \in N_2} (|a_{r_{k-1}j}^{(2)}| + |a_{jr_{k-1}}^{(2)}|) \delta_j,$$

$$2|a_{i_1 i_1}^{(k+1)}| > \sum_{\substack{j \in N_1 \\ j \neq i_1}} (|a_{i_1 j}^{(k+1)}| + |a_{j i_1}^{(k+1)}|) + \sum_{j \in N_2} (|a_{i_1 j}^{(k+1)}| + |a_{j i_1}^{(k+1)}|) \delta_j.$$

Line i_2, i_3, \dots, i_k which can be adjusted similarly to the column, can form “strictly diagonally dominant”, just as the above situation. Clearly, after the above adjustment, the other lines of the acquired matrix still remain “strictly diagonally dominant”. That is, the above whole conversions for column adjustment equals to a series of D, D_1, D_2, \dots, D_k , make $A^{(k+1)} = ADD_1 D_2 \dots D_k = (a_{ij}^{(k+1)})$, meet $\forall i \in N_1$, we have,

$$2|a_{ii}^{(k+1)}| > \sum_{\substack{j \in N_1 \\ j \neq i}} (|a_{ij}^{(k+1)}| + |a_{ji}^{(k+1)}|) + \sum_{j \in N_2} (|a_{ij}^{(k+1)}| + |a_{ji}^{(k+1)}|) \delta_j.$$

thereupon, by theorem 1, we obtain $A^{(k+1)} \in D^*$, furthermore $A \in D^*$.

(2) If A meets (ii)’s condition.

Let $X = \{x_1, x_2, \dots, x_n\}$, in which $x_i = \begin{cases} 1 & i \in N_1, \\ \delta_i & i \in N_2. \end{cases}$ Obviously, X is positive diagonal matrix. Let $B = (M(A) + M^T(A))X = (b_{ij})_{n \times n}$, then $b_{ij} = x_j(|a_{ij}| + |a_{ji}|)$. to $\forall i \in N_2$, we get $j \in N_2, i \neq j$ let $a_{ij} + a_{ji} \neq 0$. so

$$|b_{ii}| = 2\delta_i |a_{ii}| > \sum_{j \in N_1} (|a_{ij}| + |a_{ji}|) + \sum_{\substack{j \in N_2 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) \delta_j = \sum_{j \neq i} (|b_{ij}| + |b_{ji}|);$$

To $\forall i \in J(A)$, we get

$$|b_{ii}| = 2|a_{ii}| > \sum_{\substack{j \in N_1 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) + \sum_{j \in N_2} (|a_{ij}| + |a_{ji}|) \delta_j = \sum_{j \neq i} |b_{ij}|;$$

To $\forall i \in N_1/J(A)$, we get

$$|b_{ii}| = 2|a_{ii}| = \sum_{\substack{j \in N_1 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) + \sum_{j \in N_2} (|a_{ij}| + |a_{ji}|) \delta_j = \sum_{j \neq i} |b_{ij}|.$$

Since AX doesn’t change A ’s nature of nonzero element chain, from situation (ii) we’ll have nonzero element chain $b_{i_1 r_1}, b_{r_1 r_2}, \dots, b_{r_k t}$, let $t \in N_2$. That is

t satisfies $|b_{tt}| > \sum_{r_k \neq t} (|b_{r_k t}| + |b_{tr_k}|)$, by following the lemma 2, we get $B \in D$, thus $A \in D^*$.

Theorem 3 Let $A = (a_{ij}) \in C^{n \times n}$, if $\forall i \in N_1$, we get

$$2|a_{ii}| = \sum_{\substack{j \in N_1 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) + \sum_{j \in N_2} (|a_{ij}| + |a_{ji}|) \delta_j$$

Also, we have nonzero element chain $a_{ir_1}, a_{r_1 r_2}, \dots, a_{r_t}$, so that $r_k, t \in N_2$, then $A \in D^*$.

Proof Let $X = \{x_1, x_2, \dots, x_n\}$, in which $x_i = \begin{cases} 1 & i \in N_1, \\ \delta_i & i \in N_2. \end{cases}$ Obviously, X is positive diagonal matrix. Let $B = (M(A) + M^T(A))X = (b_{ij})_{n \times n}$, so $b_{ij} = x_j(|a_{ij}| + |a_{ji}|)$ to $\forall i \in N_2$, if $\sum_{j \neq i} (|a_{ij}| + |a_{ji}|) = 0$, then $a_{ij} = 0$, $a_{ji} = 0$, $j \in N$, $i \neq j$ we get

$$\begin{aligned} |b_{ii}| = 2\delta_i |a_{ii}| > 0 &= \sum_{j \in N_1} (|a_{ij}| + |a_{ji}|) + \sum_{\substack{j \in N_2 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) \delta_j \\ &= \sum_{j \neq i} (|a_{ij}| + |a_{ji}|); \end{aligned}$$

If $\sum_{j \neq i} (|a_{ij}| + |a_{ji}|) \neq 0$, and $\sum_{\substack{j \in N_2 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) \neq 0$, so

$$\begin{aligned} |b_{ii}| = 2\delta_i |a_{ii}| &= \sum_{j \in N_1} (|a_{ij}| + |a_{ji}|) + r \cdot \sum_{\substack{j \in N_2 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) \\ &> \sum_{j \in N_1} (|a_{ij}| + |a_{ji}|) + \sum_{\substack{j \in N_2 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) \delta_j \\ &= \sum_{j \neq i} (|b_{ij}| + |b_{ji}|); \end{aligned}$$

If $\sum_{j \neq i} (|a_{ij}| + |a_{ji}|) \neq 0$, but $\sum_{\substack{j \in N_2 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) = 0$, so

$$\begin{aligned} |b_{ii}| = 2\delta_i |a_{ii}| &= \sum_{j \in N_1} (|a_{ij}| + |a_{ji}|) + r \cdot \sum_{\substack{j \in N_2 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) \\ &= \sum_{j \in N_1} (|a_{ij}| + |a_{ji}|) = \sum_{j \neq i} (|b_{ij}| + |b_{ji}|). \end{aligned}$$

To $\forall i \in N_1$, because

$$\begin{aligned} |b_{ii}| &= 2|a_{ii}| = \sum_{\substack{j \in N_1 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) + \sum_{j \in N_2} (|a_{ij}| + |a_{ji}|)\delta_j \\ &= \sum_{j \neq i} (|b_{ij}| + |b_{ji}|) \end{aligned}$$

Form the known conditions, we know there is nonzero element chain $b_{ir_1}, b_{r_1r_2}, \dots, b_{r_k t}$, let $r_k, t \in N_2$, thereby $\forall i \in N_1$, there is nonzero element chain $b_{ir_1}, b_{r_1r_2}, \dots, b_{r_k t}$, let $r_k \in N_2$. then from the preceding proof, we know $|b_{r_k r_k}| > \sum_{t \neq r_k} |b_{r_k t}|$. That is, $\bar{J}(B) = \{i \in N \mid |b_{ii}| > \sum_{i \neq j} |b_{ij}|\} \neq \emptyset$ and $\forall i \in N/\bar{J}(B)$

there is nonzero element chain $b_{ir_1}, b_{r_1r_2}, \dots, b_{r_{k-1}r_k}$, making $r_k \in \bar{J}(B)$.

Since B satisfies lemma 2, so $B \in D^*$, thus $A \in D^*$.

Inference 1 Let $A = (a_{ij}) \in C^{n \times n}$ is a irreducible matrix. If $i \in N_1$, we have

$$2|a_{ii}| \geq \sum_{\substack{j \in N_1 \\ j \neq i}} (|a_{ij}| + |a_{ji}|) + \sum_{j \in N_2} (|a_{ij}| + |a_{ji}|)\delta_j$$

then $A \in D^*$.

References

1. Guo X, Gao Y (1999) Some criteria method for generalized strictly diagonally dominant matrices and nonsingular M-matrix. Numer Math J Chin Univ 21(2):189–192
2. He A, Huang R (2006) Some criteria method for generalized strictly diagonally dominant matrices. Acta Mathematicae Applicatae Sinica 19(2):401–406
3. Gao Y (1982) Criteria conditions for generalized diagonally dominant matrices and nonsingular matrix. J Northeast Norm Univ 14(3):23–28
4. Berman A, Plemmons RJ (1979) Nonnegative matrices in the mathematical sciences. Academic Press, New York

Chapter 17

Nonlinear Retarded Integral Inequalities for Discontinuous Functions and Its Applications

Wu-Sheng Wang, Zizun Li and Anmin Tang

Abstract It is well-known that integral inequality for continuous function is an important tool for studying the existence, uniqueness, boundedness, stability and other qualitative properties of solutions of differential equations and integral equations. The integral inequality for discontinuous function is an important tool for studying impulsive differential equations as well. To study the estimations of solution of nonlinear retarded impulsive integral equation, firstly retarded integral inequalities including the nonlinear composite function of discontinuous function are established, next the estimations of the unknown function of the integral inequalities are given by the methods of replacement, enlargement, differential, integral, segmentation, mathematical induction. Finally, the estimations obtained here are used to give the estimation of the solution of a class of nonlinear impulsive differential equation.

Keywords Retarded integral inequality · Discontinuous function · Estimation

17.1 Introduction

Gronwall-Bellman type inequality which furnishes explicit bounds on unknown function have become an important tool in the study of the existence, boundedness, stability and other qualitative properties of solutions of differential and integral

W.-S. Wang (✉) · A. Tang
Department of Mathematics, Hechi University, Yizhou, 546300, Guangxi, People's
Republic of China
e-mail: wang4896@163.com

Z. Li
School of Mathematics and Computer Science, Guilin University of Electronic
Technology, Guilin, 541004, Guangxi, People's Republic of China

equations. Some results related to Gronwall-Bellman type inequality can be found in [1–4]. In recent years much attention has been given to the analogous inequalities and their applications for discontinuous functions, some recent works can be found in [5–8] and some references therein. In 2010, Li et al. [4] obtained the explicit bound to the unknown function of the following inequalities.

$$\begin{aligned}
 u^2(t) \leq & k(t) + 2 \int_0^{\alpha(t)} [M_1 f_1(t, s)u(s) + N_1 g_1(t, s)u^2(s)]ds \\
 & + 2 \int_0^t [M_2 f_2(t, s)u(s) + N_2 g_2(t, s)u^2(s)]ds, t \geq 0
 \end{aligned}$$

On the basis of the above inequality, we establish a new class of Gronwall-Bellman type inequality for discontinuous function, this result furnish a handy tool for the study of the conditions of boundedness, stability by Lyapunov, practical stability by Chetaev for the solutions of impulsive differential and integro-differential systems.

17.2 Conclusion

Throughout this paper, \mathbb{R} denotes the set of real number, $t_0 \geq 0$ is given number. $\mathbb{R}_+ := (0, \infty)$, $I_i := [t_{i-1}, t_i)$, $i = 1, 2, \dots$.

17.2.1 A. Conclusion 1

Theorem 1 *Let us consider a nonnegative piecewise continuous function $u(t)$ at $t \geq t_0 \geq 0$, with the first kind of discontinuity at the points $t_i(t_0 < t_1 < t_2 \dots, \lim_{i \rightarrow \infty} t_i = \infty)$, which satisfies the retarded integral inequality for discontinuous function*

$$\begin{aligned}
 u^m(t) \leq & k(t) + 2 \int_{\alpha(t_0)}^{\alpha(t)} [M_1 f_1(t, s)u^{\frac{m}{2}}(s) + N_1 g_1(t, s)u^m(s)]ds \\
 & + 2 \int_{t_0}^t [M_2 f_2(t, s)u^{\frac{m}{2}} + N_2 g_2(t, s)u^m(s)]ds + \sum_{t_0 < t_i < t} \beta_i u(t_i - 0) \quad (17.1)
 \end{aligned}$$

where $\beta_i \geq 0, m > 0, M_i \geq 0, N_i \geq 0, i = 1, 2$ are given constants, $k : \mathbb{R}_+ \rightarrow (0, \infty)$ is a continuous and nondecreasing function, $\alpha \in C(\mathbb{R}_+^2, \mathbb{R}_+)$ is a nondecreasing function with $\alpha(t) \leq t, \alpha(t_i) = t_i, i = 0, 1, 2, \dots, \lim_{t \rightarrow \infty} \alpha(t) = \infty, f_i, g_i \in C(\mathbb{R}_+^2, \mathbb{R}_+)$ are nondecreasing on $t, \partial_t f_i(t, s), \partial_t g_i(t, s) \in C(\mathbb{R}_+^2, \mathbb{R}_+), i = 1, 2$. Then the function $u(t)$ will satisfy the estimation

$$u(t) \leq \sqrt[m]{k(t)} \left[e^{\int_{t_i}^t R_i(s) ds} \left(c_i + \int_{t_i}^t Q_i(s) e^{-\int_{t_i}^s R_i(\tau) d\tau} ds \right) \right]^{\frac{2}{m}} \tag{17.2}$$

$\forall t \in [t_i, t_{i+1}], \quad i = 0, 1, 2, \dots,$

where

$$R_j(s) := N_1 g_1(s, \alpha(s)) \alpha'(s) + N_2 g_2(s, s) + \int_{\alpha(t_k)}^{\alpha(s)} N_1 \partial_s g_1(s, \tau) d\tau + \int_{t_k}^s N_2 \partial_s g_2(s, \tau) d\tau \tag{17.3}$$

$$Q_j(s) := M_1 \tilde{f}_1(s, \alpha(s)) \alpha'(s) + M_2 \tilde{f}_2(s, s) + \int_{\alpha(t_k)}^{\alpha(s)} M_1 \partial_s \tilde{f}_1(s, \tau) d\tau + \int_{t_k}^s M_2 \partial_s \tilde{f}_2(s, \tau) d\tau \tag{17.4}$$

$$j = 0, 1, 2, \dots, i - 1, s \in [t_k, t_{k+1}), k = i, s \in [t_i, t),$$

and

$$\tilde{f}_1(t, s) = \frac{f_1(t, s)}{\sqrt{k(s)}} \tilde{f}_2(t, s) = \frac{f_2(t, s)}{\sqrt{k(s)}} \tag{17.5}$$

$$c_0 = 1 \tag{17.6}$$

$$c_i = \frac{\beta_i}{\frac{m-1}{m}(t_i)} \left[e^{\int_{t_{i-1}}^{t_i} R_{i-1}(s) ds} \left(c_{i-1} + \int_{t_{i-1}}^i Q_{i-1}(s) e^{-\int_{t_{i-1}}^s R_{i-1}(\tau) d\tau} ds \right) \right]^{\frac{2}{m}} + \left[\int_{t_{i-1}}^i R_{i-1}(s) ds \left(c_{i-1} + \int_{t_{i-1}}^i Q_{i-1}(s) e^{-\int_{t_{i-1}}^s R_{i-1}(\tau) d\tau} ds \right) \right]^2 \quad i = 1, 2, \dots \tag{17.7}$$

Proof Taking into account the inequality of (17.1), we get

$$\begin{aligned} \frac{u^m(t)}{k(t)} &\leq 1 + 2 \int_{\alpha(t_0)}^{\alpha(t)} \left[M_1 f_1(t, s) \frac{u^{\frac{m}{2}}(s)}{k(s)} + N_1 g_1(t, s) \frac{u^m(s)}{k(s)} \right] ds \\ + 2 \int_{t_0}^t \left[M_2 f_2(t, s) \frac{u^{\frac{m}{2}}(s)}{k(s)} + N_2 g_2(t, s) \frac{u^m(s)}{k(s)} \right] ds &+ \sum_{t_0 < t_i < t} \beta_i \frac{u(t_i - 0)}{k(t_i)}, \forall t \geq t_0. \end{aligned} \quad (17.8)$$

Let

$$W(t) := \frac{u^m(t)}{k(t)}, \quad (17.9)$$

from (17.5) and (17.8), we have

$$\begin{aligned} W(t) &\leq 1 + 2 \int_{\alpha(t_0)}^{\alpha(t)} \left[M_1 \frac{f_1(t, s)}{\sqrt{k(s)}} W^{\frac{1}{2}}(s) + N_1 g_1(t, s) W(s) \right] ds \\ &+ 2 \int_{t_0}^t \left[M_2 \frac{f_2(t, s)^{\frac{1}{2}}}{\sqrt{k(s)}} (s) + N_2 g_2(t, s) W(s) \right] ds + \sum_{t_0 < t_i < t} \beta_i \frac{W^{\frac{1}{m}}(t_i - 0)}{k^{\frac{m-1}{m}}(t_i)} \\ &= 1 + 2 \int_{\alpha(t_0)}^{\alpha(t)} \left[M_1 \tilde{f}_1(t, s) W^{\frac{1}{2}}(s) + N_1 g_1(t, s) W(s) \right] ds \\ &+ 2 \int_{t_0}^t \left[M_2 \tilde{f}_2(t, s)^{\frac{1}{2}}(s) + N_2 g_2(t, s) W(s) \right] ds + \sum_{t_0 < t_i < t} \beta_i \frac{W^{\frac{1}{m}}(t_i - 0)}{k^{\frac{m-1}{m}}(t_i)}, \forall t \geq t_0 \end{aligned} \quad (17.10)$$

Denote $v(t)$ by

$$\begin{aligned} v(t) &:= 2 \int_{\alpha(t_0)}^{\alpha(t)} \left[M_1 \tilde{f}_1(t, s) W^{\frac{1}{2}}(s) + N_1 g_1(t, s) W(s) \right] ds \\ &+ 2 \int_{t_0}^t \left[M_2 \tilde{f}_2(t, s)^{\frac{1}{2}}(s) + N_2 g_2(t, s) W(s) \right] ds \end{aligned} \quad (17.11)$$

In the following, we shall prove the estimation (17.2).

Firstly, we consider the case $t \in I_1$, by (17.11), we have

$$W(t) \leq 1 + v(t), \quad W^{\frac{1}{2}}(t) \leq \sqrt{1 + v(t)} \quad (17.12)$$

By the assumptions on f_i , g_i and α , we see that $v(t)$ is nondecreasing on \mathbb{R}_+ . Hence, from (17.11) and (17.12), we have

$$\begin{aligned}
 v'(t) &= 2\alpha'(t) \left[M_1 \tilde{f}_1(t, \alpha(t)) W^{\frac{1}{2}}(\alpha(t)) + N_1 g_1(t, \alpha(t)) W(\alpha(t)) \right] \\
 &\quad + 2 \left[M_2 \tilde{f}_2(t, t) W^{\frac{1}{2}}(t) + N_2 g_2(t, t) W(t) \right] \\
 &\quad + 2 \int_{\alpha(t_0)}^{\alpha(t)} \left[M_1 \partial_t \tilde{f}_1(t, s) W^{\frac{1}{2}}(s) + N_1 \partial_t g_1(t, s) W(s) \right] ds \\
 &\quad + 2 \int_{t_0}^t \left[M_2 \partial_t \tilde{f}_2(t, s) W^{\frac{1}{2}}(s) + N_2 \partial_t g_2(t, s) W(s) \right] ds \\
 &\leq 2\sqrt{1 + v(t)} [M_1 \tilde{f}_1(t, \alpha(t)) \alpha'(t) + M_2 \tilde{f}_2(t, t) + \int_{\alpha(t_0)}^{\alpha(t)} M_1 \partial_t \tilde{f}_1(t, s) ds] \\
 &\quad + \int_{t_0}^t M_2 \partial_t \tilde{f}_2(t, s) ds + 2(1 + v(t)) [N_1 g_1(t, \alpha(t)) \alpha'(t) \\
 &\quad + N_2 g_2(t, t) + \int_{\alpha(t_0)}^{\alpha(t)} N_1 \partial_t g_1(t, s) ds + \int_{t_0}^t N_2 \partial_t g_2(t, s) ds] \tag{17.13}
 \end{aligned}$$

By the definition of $R_j(t)$ in (17.3) and $Q_j(t)$ in (17.4), from (17.13), we obtain

$$v'(t) \leq 2Q_0(t) \sqrt{1 + v(t)} + 2R_0(t)(1 + v(t)),$$

i.e.

$$\frac{v'(t)}{2\sqrt{1 + v(t)}} \leq Q_0(t) + R_0(t) \sqrt{1 + v(t)},$$

Or equivalently

$$\frac{[1 + v(t)]'}{2\sqrt{1 + v(t)}} \leq Q_0(t) + R_0(t) \sqrt{1 + v(t)}, \tag{17.14}$$

from (17.14), we obtain

$$\frac{d(\sqrt{1 + v(t)})}{dt} \leq Q_0(t) + R_0(t) \sqrt{1 + v(t)}. \tag{17.15}$$

From (17.15), for all $t \in I_1$, we obtain

$$\sqrt{1 + v(t)} \leq e^{\int_{t_0}^t R_0(s) ds} \left(1 + \int_{t_0}^t Q_0(s) e^{-\int_{t_0}^s R_0(\tau) d\tau} ds \right). \tag{17.16}$$

By (17.12), we have

$$W(t) \leq \left[e^{\int_{t_0}^t R_0(s) ds} \left(1 + \int_{t_0}^t Q_0(s) e^{-\int_{t_0}^s R_0(\tau) d\tau} ds \right) \right]^2 \tag{17.17}$$

By (17.9), from (17.17), we get $u(t) \leq \sqrt[m]{k(t)}$
 $\left[e^{\int_{t_0}^t R_0(s) ds} \left(1 + \int_{t_0}^t Q_0(s) e^{-\int_{t_0}^s R_0(\tau) d\tau} ds \right) \right]^{\frac{2}{m}}$, Implying that (17.2) is true for $t \in I_1$.

Next, we consider $t \in I_2 = [t_1, t_2)$. Using the hypotheses on f_i, g_i and α , from (17.10), we have

$$\begin{aligned} W(t) &\leq 1 + 2 \int_{\alpha(t_0)}^{\alpha(t_1)} [M_1 \tilde{f}_1(t, s) W^{\frac{1}{2}}(s) + N_1 g_1(t, s) W(s)] ds \\ &\quad + 2 \int_{t_0}^{t_1} [M_2 \tilde{f}_2(t, s) W^{\frac{1}{2}}(s) + N_2 g_2(t, s) W(s)] ds \\ &\quad + \beta_1 \frac{W_m^{\frac{1}{m}}(t_1 - 0)}{k^{\frac{m-1}{m}}(t_1)} + 2 \int_{\alpha(t_1)}^{\alpha(t)} [M_1 \tilde{f}_1(t, s) W^{\frac{1}{2}}(s) + N_1 g_1(t, s) W(s)] ds \\ &\quad + 2 \int_{t_1}^t [M_2 \tilde{f}_2(t, s) W^{\frac{1}{2}}(s) + N_2 g_2(t, s) W(s)] ds \\ &\leq 1 + 2 \int_{\alpha(t_0)}^{\alpha(t_1)} [M_1 \tilde{f}_1(t_1, s) W^{\frac{1}{2}}(s) + N_1 g_1(t_1, s) W(s)] ds \\ &\quad + 2 \int_{t_0}^{t_1} [M_2 \tilde{f}_2(t_1, s) W^{\frac{1}{2}}(s) + N_2 g_2(t_1, s) W(s)] ds \\ &\quad + \beta_1 \frac{W_m^{\frac{1}{m}}(t_1 - 0)}{k^{\frac{m-1}{m}}(t_1)} + 2 \int_{\alpha(t_1)}^{\alpha(t)} [M_1 \tilde{f}_1(t, s) W^{\frac{1}{2}}(s) + N_1 g_1(t, s) W(s)] ds \end{aligned}$$

$$\begin{aligned}
 &+ 2 \int_{t_1}^t [M_2 \tilde{f}_2(t, s) W^{\frac{1}{2}}(s) + N_2 g_2(t, s) W(s)] ds \\
 &\leq \left[e^{\int_{t_0}^{t_1} R_0(s) ds} \left(1 + \int_{t_0}^{t_1} Q_0(s) e^{-\int_{t_0}^s R_0(\tau) d\tau} ds \right) \right]^2 \\
 &+ \frac{\beta_1}{k^{\frac{m-1}{m}}(t_1)} \left[e^{\int_{t_0}^{t_1} R_0(s) ds} \left(1 + \int_{t_0}^{t_1} Q_0(s) e^{-\int_{t_0}^s R_0(\tau) d\tau} ds \right) \right]^{\frac{2}{m}} \\
 &+ 2 \int_{\alpha(t_1)}^{\alpha(t)} [M_1 \tilde{f}_1(t, s) W^{\frac{1}{2}}(s) + N_1 g_1(t, s) W(s)] ds \\
 &+ 2 \int_{t_1}^t [M_2 \tilde{f}_2(t, s) W^{\frac{1}{2}}(s) + N_2 g_2(t, s) W(s)] ds. \tag{17.18}
 \end{aligned}$$

Denote $z(t)$ by

$$\begin{aligned}
 z(t) &= 2 \int_{\alpha(t_1)}^{\alpha(t)} [M_1 \tilde{f}_1(t, s) W^{\frac{1}{2}}(s) + N_1 g_1(t, s) W(s)] ds \\
 &+ 2 \int_{t_1}^t [M_2 \tilde{f}_2(t, s) W^{\frac{1}{2}}(s) + N_2 g_2(t, s) W(s)] ds. \tag{17.19}
 \end{aligned}$$

From (17.7) and (17.19), (17.18) can be written as $W(t) \leq c_1 + z(t)$, $W^{\frac{1}{2}} \leq \sqrt{c_1 + z(t)}$.

Differentiating $z(t)$, we get

$$\begin{aligned}
 z'(t) &= 2\alpha'(t) [M_1 \tilde{f}_1(t, \alpha(t)) W^{\frac{1}{2}}(\alpha(t)) + N_1 g_1(t, \alpha(t)) W(\alpha(t))] \\
 &+ 2 [M_2 \tilde{f}_2(t, t) W^{\frac{1}{2}}(t) + N_2 g_2(t, t) W(t)] \\
 &+ 2 \int_{\alpha(t_1)}^{\alpha(t)} [M_1 \partial_t \tilde{f}_1(t, s) W^{\frac{1}{2}} + N_1 \partial_t g_1(t, s) W(s)] ds \\
 &+ 2 \int_{t_1}^t [M_2 \partial_t \tilde{f}_2(t, s) W^{\frac{1}{2}} + N_2 \partial_t g_2(t, s) W(s)] ds
 \end{aligned}$$

$$\begin{aligned}
 &\leq 2\sqrt{c_1 + z(t)} \left[M_1 \tilde{f}_1(t, \alpha(t))\alpha'(t) + M_2 \tilde{f}_2(t, t) + \int_{\alpha(t_1)}^{\alpha(t)} M_1 \partial_t \tilde{f}_1(t, s) \right. \\
 &\quad \left. + \int_{t_1}^t M_2 \partial_t \tilde{f}_2(t, s) \right] + 2(c_1 + z(t)) [N_1 g_1(t, \alpha(t))\alpha'(t) + N_2 g_2(t, t) \\
 &\quad + \int_{\alpha(t_1)}^{\alpha(t)} N_1 \partial_t g_1(t, s) ds + \int_{t_1}^t N_2 \partial_t g_2(t, s) ds]. \tag{17.20}
 \end{aligned}$$

By the definition of $R_j(t)$ (17.3) and $Q_j(t)$ (17.4), from (17.20), we obtain $z'(t) \leq 2Q_1(t)\sqrt{c_1 + z(t)} + 2R_1(t)(c_1 + z(t))$.

Similar the proof of procedure $t \in [t_0, t_1)$, we can deduce that $u(t) \leq \sqrt[m]{k(t)} \left[e^{\int_{t_1}^t R_1(s) ds} \left(c_1 + \int_{t_1}^t Q_1(s) e^{-\int_{t_1}^s R_1(\tau) d\tau} ds \right) \right]^{\frac{2}{m}}$, for all $t \in [t_1, t_2)$, it implies that (17.2) is true for $t \in [t_1, t_2)$.

In a similar way, for $t \in I_i = [t_i, t_{i+1})$, we can deduce that $u(t) \leq \sqrt[m]{k(t)} \left[e^{\int_{t_i}^t R_i(s) ds} \left(c_i + \int_{t_i}^t Q_i(s) e^{-\int_{t_i}^s R_i(\tau) d\tau} ds \right) \right]^{\frac{2}{m}}$, for all $t \in [t_i, t_{i+1})$. This completes the proof.

Remark 1 (1) When $m = 2, \beta_i = 0$, Theorem 2.1 reduces to Theorem 2.1 of Li et al. [4].

(2) When $k(t) = c, m = 1, M_1 = M_2 = N_1 = 0, N_2 = \frac{1}{2}, g(t, s) = v(s)$, Theorem 2.1 reduces to Theorem 1 of Samoilenko and Perestyuk [5].

17.2.2 B. Conclusion 2

Theorem 2 *Let us suppose that a nonnegative piecewise Continuous function $u(t)$ at $t \geq t_0 \geq 0$, with the first kind of discontinuity at the points $t_i \left(t_0 < t_1 < t_2 \dots \lim_{i \rightarrow \infty} t_i = \infty \right)$, which satisfies the integral inequality for discontinuous function*

$$\begin{aligned}
 u^m(t) &\leq k(t) + \frac{m}{m-n} \int_{a(t_0)}^{a(t)} [M_1 f_1(t, s) u^n(s) + N_1 g_1(t, s) u^n(s) w(u(s))] ds \\
 &\quad + \frac{m}{m-n} \int_{t_0}^t [M_2 f_2(t, s) u^n + N_2 g_2(t, s) u^n(s) w(u(s))] ds + \sum_{t_0, t_i} \beta_i u^m(t_i - 0), \tag{17.21}
 \end{aligned}$$

where $\beta_i \geq 0, m > n > 0, M_i \geq 0, N_i \geq 0, i = 1, 2$ are given constants, $k : \mathbb{R}_+ \rightarrow (0, \infty)$ is a continuous and nondecreasing function $f_i, g_i \in C(\mathbb{R}_+^2, \mathbb{R}_+)$, are nonincreasing on $t, \partial_t f_i(t, s), \partial_t g_i(t, s) \in C(\mathbb{R}_+^2, \mathbb{R}_+), i = 1, 2, a \in C^1(\mathbb{R}_+^2, \mathbb{R}_+)$ is a nondecreasing function with $\alpha(t) \leq t, \alpha(t_i) = t_i, \lim_{t \rightarrow \infty} \alpha(t) = \infty$, function $w(s)$ satisfies the following class ζ : (1) w is nondecreasing; (2) $w : \mathbb{R}_+ \rightarrow \mathbb{R}_+, w(0) = 0$; (3) $w(\alpha\beta) \leq w(\alpha)w(\beta)$. Then $\forall t \geq t_0$ the function $u(t)$ will satisfy the estimation

$$u(t) \leq \sqrt[m]{k(t)} \left\{ \Phi_i^{-1} \left[\Phi_i \left(e_i + \int_{\alpha(t_i)}^{\alpha(t)} M_1 \hat{f}_1(t, s) ds + \int_{t_i}^t M_2 \hat{f}_2(t, s) ds \right) + \int_{\alpha(t_i)}^{\alpha(t)} N_1 \hat{g}_1(t, s) w(\sqrt[m]{k(s)}) ds + \int_{t_i}^t N_2 \hat{g}_2(t, s) w(\sqrt[m]{k(s)}) ds \right] \right\}^{\frac{1}{m-n}} \tag{17.22}$$

For all $t \in [t_i, t_{i+1}), i = 0, 1, 2, \dots$, where

$$\Phi_i(t) = \int_{t_i}^t \frac{ds}{w(s^{\frac{1}{m-n}})} \lim_{t \rightarrow \infty} \Phi_i(t) = \infty, i = 0, 1, 2, \dots \tag{2.23}$$

$$\hat{f}_1(t, s) = \frac{f_1(t, s)}{\sqrt[m]{k^{m-n}(s)}}, \hat{f}_2(t, s) = \frac{f_2(t, s)}{\sqrt[m]{k^{m-n}(s)}},$$

$$\hat{g}_1(t, s) = \frac{g_1(t, s)}{\sqrt[m]{k^{m-n}(s)}}, \hat{g}_2(t, s) = \frac{g_2(t, s)}{\sqrt[m]{k^{m-n}(s)}},$$

$$e_0 = 1,$$

$$e_i = (1 + \beta_i) \sqrt[m]{k(t)} \left\{ \Phi_{i-1}^{-1} \left[\Phi_{i-1} \left(e_{i-1} + \int_{\alpha(t_{i-1})}^{\alpha(t_i)} M_1 \hat{f}_1(t, s) ds + \int_{\alpha(t_{i-1})}^{\alpha(t_i)} N_1 \hat{g}_1(t, s) w(\sqrt[m]{k(s)}) ds + \int_{t_{i-1}}^{t_i} M_2 \hat{f}_2(t, s) ds + \int_{t_{i-1}}^{t_i} N_2 \hat{g}_2(t, s) w(\sqrt[m]{k(s)}) ds \right) \right] \right\}^{\frac{m}{m-n}}, i = 0, 1, 2, \dots$$

17.2.3 C. Conclusion 3

In this section, we will show that our results are useful in proving the boundedness of solutions of impulsive differential system. We consider an impulsive system as follows:

$$\begin{aligned} \frac{dx^n(t)}{dt} &= F(t, x(t), x(\alpha(t))) \quad t \neq t_i \\ \Delta x|_{t=t_i} &= I_i(x) \\ x(t_0^+) &= x_0 \end{aligned} \tag{17.24}$$

where $x \in \mathbb{R}^k$, $F \in C(\mathbb{R}^{2k+1}, \mathbb{R}^k)$, k is a given natural number, $I_i(x) \in C(\mathbb{R}^k, \mathbb{R}^k)$, $\alpha \in C^1(\mathbb{R}^k, \mathbb{R}^k)$ is a nondecreasing function with $\alpha(t) \leq t$, $t \geq t_0 \geq 0$, $\alpha(t_i) = t_i$, $t_{i-1} < t_i, \forall i = 1, 2, \dots, \lim_{i \rightarrow \infty} t_i = \infty$. Let us assume that F, I_i satisfy the following conditions:

$$\begin{aligned} (a) \quad \|F(t, x)\| &\leq f_1(t)\|x(t)\|^n + f_2(t)\|x(\alpha(t))\|^n + g_1(t)\|x(t)\|^n w(\|x(t)\|) \\ &\quad + g_2(t)\|x(\alpha(t))\|^n w(\|x(\alpha(t))\|), \end{aligned}$$

(b) $\|I_i(x)\| \leq \beta_i \|x\|^m$, where $\beta_i \geq 0$ are constants, $i = 1, 2, \dots$, $f_1, f_2, g_1, g_2 \in C(\mathbb{R}_+, \mathbb{R}_+)$, $w \in C(\mathbb{R}_+, \mathbb{R}_+)$ is a nondecreasing function with $w(t) > 0$ for $t > 0$.

Corollary 1 Under assumptions of the conditions (a) and (b), all solutions $x(t)$ of the system (17.24) have the estimation

$$\begin{aligned} \|x(t)\| &\leq \left\{ \Phi_i^{-1} \left[\Phi_i \left(e_i + \frac{m-n}{m} \left(\int_{t_i}^t f_1(s) ds + \int_{\alpha(t_i)}^{\alpha(t)} \frac{f_2(\alpha^{-1}(s))}{\alpha'(\alpha^{-1}(s))} ds \right) \right. \right. \\ &\quad \left. \left. + \frac{m-n}{m} \left(\int_{t_i}^{t_i} g_1(s) ds + \int_{\alpha(t_i)}^{\alpha(t)} \frac{g_2(\alpha^{-1}(s))}{\alpha'(\alpha^{-1}(s))} ds \right) \right] \right\}^{\frac{1}{m-n}} \end{aligned} \tag{17.25}$$

for all $t \in [t_i, t_{i+1})$, $i = 0, 1, 2, \dots$, where $\Phi_i(t), i = 0, 1, 2, \dots$, are defined by (2.23), $e_0 = x_0$,

$$\begin{aligned} e_j &= (1 + \beta_j) \left\{ \Phi_{j-1}^{-1} \left[\Phi_{j-1} \left(e_{j-1} + \frac{m-n}{m} \left(\int_{t_{j-1}}^{t_j} f_1(s) ds + \int_{\alpha(t_{j-1})}^{\alpha(t_j)} \frac{f_2(\alpha^{-1}(s))}{\alpha'(\alpha^{-1}(s))} ds \right) \right) \right. \right. \\ &\quad \left. \left. + \frac{m-n}{m} \left(\int_{t_{j-1}}^{t_j} g_1(s) ds + \int_{\alpha(t_{j-1})}^{\alpha(t_j)} \frac{g_2(\alpha^{-1}(s))}{\alpha'(\alpha^{-1}(s))} ds \right) \right] \right\}^{\frac{1}{m-n}}, \quad j = 1, 2, \dots, i. \end{aligned}$$

Acknowledgments This work is supported by the Natural Science Foundation of Guangxi Autonomous Region (0991265) and the Key Project of Hechi University (2009YAZ-N001)

References

1. Bellman R (1943) The stability of solutions of linear differential equations. J Duke Math 10:643–647

2. Agarwal RP, Deng S, Zhang W (2005) Generalization of a retarded Gronwall-like inequality and its applications. *Appl Math Comput* 165:599–612
3. Cheung WS (2006) Some new nonlinear inequalities and applications to boundary value problems. *Nonlinear Anal* 64:2112–2128
4. Li L, Meng F, He L (2010) Some generalized integral inequalities and their applications. *J Math Anal Appl* 372:339–349
5. Samoilenko AM, Perestyuk N (1987) *Differential equations with impulse effect*. Visha Shkola, Kyiv
6. Borysenko SD, Iovane G, Giordano P (2005) Investigations of the properties motion for essential nonlinear systems perturbed by impulses on some hypersurfaces. *Nonlinear Anal* 62:345–363
7. Gallo A, Piccirillo AM (2009) About some new generalizations of Bellman-Bihari results for integro-functional inequalities with discontinuous functions and applications. *Nonlinear Anal* 71:2276–2287
8. Wang WS, Zhou X (2010) A generalized Gronwall-Bellman-Ou-Iang type inequality for discontinuous functions and applications to BVP. *Appl Math Comput* 216:3335–3342

Chapter 18

Oscillatory and Asymptotic Behavior of a Second Order Nonlinear Differential Equation with Perturbation

Li Gao, Quanxin Zhang and Xia Song

Abstract In this paper, a class of second-order nonlinear differential equation with perturbation is studied. By using the generalized Riccati transformation, the integral averaging technique and the method of classification, new oscillatory criteria and asymptotic behavior are obtained for all solutions of the equations, which generalize and improve some known results.

Keywords Oscillatory criterion · Asymptotic behavior · Differential equation with perturbation

18.1 Introduction

As is well known, the comparison and separation theory of zeros distribution for second order homogeneous linear differential equations established by Sturm lay a foundation of oscillation theory for differential equations. During one and a half century, oscillation theory of differential equations has developed quickly and played an important role in qualitative theory of differential equations and theory of boundary value problem. The study of oscillation theory plays an important role in physical sciences and technology; for example, the oscillation of building and machine, electromagnetic vibration in radio technology and optical science, self-excited vibration in control system, sound vibration, beam vibration in synchrotron accelerator, the vibration sparked for burning rocket engine, the complicated

L. Gao (✉) · Q. Zhang · X. Song
Department of Mathematics and Information Science, Binzhou University, 256603,
Shandong, People's Republic of China
e-mail: gaolibzxy@163.com

oscillation in chemical reaction, etc. All the different phenomena can be unified into oscillation theory through an oscillation equation. There are many books on oscillation theory, we choose to refer to [1, 2]. Firstly, we give the oscillatory definition of a solution of a differential equation.

Definition 1 $x(t)$ is a solution of a differential equation, if $x(t)$ is not the eventually zero solution, and existing a sequence $\{t_i\}$, $\lim_{i \rightarrow \infty} t_i = \infty$, such that $x(t_i) = 0$. Such a solution is said to be oscillatory and otherwise it is said to be nonoscillatory. A nonoscillatory solution $x(t)$ is called weakly oscillatory if $x'(t)$ changes sign for arbitrary large t . See [1, 3].

Definition 2 An equation is called oscillatory if all its solutions are oscillatory.

The oscillatory and asymptotic behavior of second-order nonlinear differential equations have been widely applied in research of a lossless high-speed computer network and physical sciences. In this paper, we study the oscillatory and asymptotic behavior of solutions for a class of second-order nonlinear differential equation with perturbation

$$(a(t)\psi(x(t))x'(t))' + Q(t, x(t)) = P(t, x(t), x'(t)), \tag{18.1}$$

where we let

- (A1) $a : [t_0, +\infty) \rightarrow \mathbb{R}$, $\mathbb{R} = (-\infty, +\infty)$ is positive continuously differentiable function;
- (A2) $\psi : \mathbb{R} \rightarrow \mathbb{R}$ is continuously differentiable function and $\psi(u) > 0$ for $u \neq 0$;
- (A3) $Q : [t_0, +\infty) \times \mathbb{R} \rightarrow \mathbb{R}$ is continuous, and there exists a continuous function $q(t)$ and continuously differentiable function $f(x)$, where $q : [t_0, +\infty) \times \mathbb{R} \rightarrow \mathbb{R}, f : \mathbb{R} \rightarrow \mathbb{R}, uf(u) > 0, f : \mathbb{R} \rightarrow \mathbb{R}$, and $f'(u) > 0$ for $u \neq 0$, such that $\frac{Q(t,x)}{f(x)} \geq q(t)$ for $x \neq 0$;
- (A4) $P \in ([t_0, +\infty) \times \mathbb{R}^2 \rightarrow \mathbb{R})$, and there exists $p(t) \in ([t_0, +\infty) \rightarrow \mathbb{R})$ which is continuous, such that $\frac{P(t, x(t), x'(t))}{f(x)} \leq p(t)$ for $x \neq 0$.

It is easy to see that (18.1) can be transformed into

$$(a(t)\psi(x(t))x'(t))' + p(t)x'(t) + q(t)f(x(t)) = 0, \tag{E1}$$

where $Q(t, x(t)) = q(t)f(x(t)), R(t, x(t), x'(t)) = -p(t)x'(t)$.

In (E1), if $a(t) = 1, \psi(u) = 1$, then (E1) can be transformed to

$$x''(t) + p(t)x'(t) + q(t)f(x(t)) = 0. \tag{E2}$$

In (E1), if $p(t) = 0$, then (E1) is transformed to

$$(a(t)\psi(x(t))x'(t))' + q(t)f(x(t)) = 0. \tag{E3}$$

In (E2), if $p(t) = 0, f(x) = x$, then (E2) becomes

$$x''(t) + q(t)x(t) = 0. \tag{E4}$$

Equation (E4) has many oscillatory criteria. One of the most well-known criteria is Wintner’s oscillatory criterion [4]. It states that the linear equation(E4) is oscillatory if

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_{t_0}^t \int_{t_0}^s q(t) dx ds = \infty.$$

In 1992, Cecchi and Marini [3] extended and improved this criterion to (E3). In 2000, Rogovchenko [5] and in 2008, Ohriska [6], the oscillation result for (E3) with delay variable had been discussed. In 2008, Cakmak [7] discussed the oscillatory criterion of (E2) with damping, which extended and improved the Wintner’s result. Among the other papers in the study of (E1), we refer to [8–11]. In 2010, Zhang and Wang [12] discussed the oscillatory behavior of (18.1) under some conditions by using the generalized Riccati transformation, the integral averaging technique and the method of classification. This article is the continuation of [12], we continue to discuss the oscillatory and asymptotic behavior of (18.1) by using the generalized Riccati transformation, the integral averaging technique and the method of classification. We established two oscillatory and three asymptotic behavior criteria which develop and generalize the known results.

With respect to their asymptotic behavior, all solutions of (18.1) can be divided into the following four classes:

- $S^+ = \{x = x(t) \text{ solution of (18.1): there exists } t_x \geq t_0 \text{ such that } x(t)x'(t) \geq 0 \text{ for } t \geq t_x\}$;
- $S^- = \{x = x(t) \text{ solution of (18.1): there exists } t_x \geq t_0 \text{ such that } x(t)x'(t) < 0 \text{ for } t \geq t_x\}$;
- $S^O = \{x = x(t) \text{ solution of (18.1): there exists } t_n, t_n \rightarrow +\infty, \text{ such that } x(t_n) = 0\}$;
- $S^{WO} = \{x = x(t) \text{ solution of (18.1): } x(t) \neq 0 \text{ for sufficiently large } t \text{ and for all } t_x > t_0 \text{ there exist } t_{x_1} > t_x, t_{x_2} > t_x \text{ such that } x'(t_{x_1})x'(t_{x_2}) < 0\}$.

With very simple arguments we can prove that S^+, S^-, S^O, S^{WO} are mutually disjoint. By the above definitions, it turns out that solutions in the class S^+ are either eventually positive nondecreasing or negative nonincreasing, solutions in the class S^- are either eventually positive decreasing or negative increasing, solutions in the class S^O are oscillatory, and finally, solutions in the class S^{WO} are weakly oscillatory.

18.2 Oscillatory Behavior

Lemma 1 (I) If

$$\limsup_{t \rightarrow +\infty} \int_{t_0}^t [q(s) - p(s)] ds = +\infty, \tag{18.2}$$

then $S^+ = \emptyset$ for (18.1);

(II) If

$$\lim_{t \rightarrow +\infty} \int_{t_0}^t [q(s) - p(s)] ds = +\infty, \tag{18.3}$$

then $S^{WO} = \emptyset$ for (18.1).

Proof (I) Suppose that (18.1) has a solution $x(t) \in S^+$. There is no loss of generality in assuming that there exists $t_1 \geq t_0$ such that $x(t) > 0, x'(t) \geq 0$ for all $t \geq t_1$. The proof is similar such that $x(t) < 0, x'(t) \leq 0$ for all $t \geq t_1$. Consider the function $W(t) = \frac{a(t)\psi(x(t))x'(t)}{f(x(t))}, t \geq t_1$.

Based on (18.1), we have

$$\begin{aligned} W'(t) &= -\frac{Q(t, x(t))}{f(x(t))} + \frac{P(t, x(t), x'(t))}{f(x(t))} - a(t)\psi(x(t))f'(x(t))\frac{x'^2(t)}{f^2(x(t))} \\ &\leq -q(t) + p(t) - a(t)\psi(x(t))f'(x(t))\frac{x'^2(t)}{f^2(x(t))} \\ &\leq -q(t) + p(t), \end{aligned}$$

and consequently, for all $t \geq t_1$,

$$\frac{a(t)\psi(x(t))x'(t)}{f(x(t))} \leq \frac{a(t_1)\psi(x(t_1))x'(t_1)}{f(x(t_1))} - \int_{t_1}^t [q(s) - p(s)] ds. \tag{18.4}$$

From (18.2), we obtain $\liminf_{t \rightarrow +\infty} \frac{a(t)\psi(x(t))x'(t)}{f(x(t))} = -\infty$, which contradicts the assumption $x(t) > 0, x'(t) \geq 0$ for all large t .

(II) Suppose that (18.1) has a solution $x(t) \in S^{WO}$. There is no loss of generality in assuming that there exists $t_1 \geq t_0$ such that $x(t) > 0$ for all $t \geq t_1$. The proof is similar such that $x(t) < 0$ for all $t \geq t_1$. Since for all $t_\alpha \geq t_0$ there exist $t_{\alpha_1} > t_\alpha, t_{\alpha_2} > t_\alpha$ such that $x'(t_{\alpha_1})x'(t_{\alpha_2}) < 0$. Proceeding as in the proof of (I), we obtain (18.4). From (18.3), we obtain $x'(t) < 0$ for all large t , which gives a contradiction since $x'(t_{\alpha_1})x'(t_{\alpha_2}) < 0$. The proof is complete.

Remark 1 If $R(t, x(t), x'(t)) = 0, Q(t, x(t)) = q(t)f(x(t))$ is satisfied, from Lemma 1(II), we can obtain Theorem 2(a) in [3].

Remark 2 Obviously, if assumption (18.3) and (18.2) are satisfied, then we can

$$\begin{aligned} &\text{obtain } S^+ \\ &= S^{WO} = \emptyset \end{aligned}$$

Lemma 2 If $\frac{f(u)}{\psi(u)}$ is strongly sublinear, i.e., the function $\frac{\psi(u)}{f(u)}$ is locally integrable on $(0, \varepsilon)$ and $(-\varepsilon, 0)$ for some $\varepsilon > 0$, that is

$$\int_0^\varepsilon \frac{\psi(u)}{f(u)} du < +\infty, \int_{-\varepsilon}^0 \frac{\psi(u)}{f(u)} du > -\infty, \tag{18.5}$$

and if

$$\limsup_{t \rightarrow +\infty} \int_T^t \frac{1}{a(s)} \int_T^s [q(\tau) - p(\tau)] d\tau ds = +\infty, \tag{18.6}$$

for all $T \geq t_0$, then $S^- = \emptyset$ for (18.1).

Proof Suppose that (18.1) has a solution $x(t) \in S^-$. There is no loss of generality in assuming that there exists $t_1 \geq t_0$ such that $x(t) > 0, x'(t) < 0$ for all $t \geq t_1$. The proof is similar such that $x(t) < 0, x'(t) > 0$ for all $t \geq t_1$. From (18.4) we obtain

$$\frac{a(t)\psi(x(t))x'(t)}{f(x(t))} \leq - \int_{t_1}^t [q(s) - p(s)] ds, t \geq t_1,$$

i.e.,

$$\frac{\psi(x(t))x'(t)}{f(x(t))} \leq - \frac{1}{a(t)} \int_{t_1}^t [q(s) - p(s)] ds, t \geq t_1,$$

Thus, we have

$$\int_{t_1}^t \frac{\psi(x(s))x'(s)}{f(x(s))} ds \leq - \int_{t_1}^t \frac{1}{a(s)} \int_{t_1}^s [q(\tau) - p(\tau)] d\tau ds$$

i.e.,

$$\int_{x(t)}^{x(t_1)} \frac{\psi(u)}{f(u)} du \geq \int_{t_1}^t \frac{1}{a(s)} \int_{t_1}^s [q(\tau) - p(\tau)] d\tau ds,$$

which, because of (18.6), implies

$$\limsup_{t \rightarrow +\infty} \int_{x(t)}^{x(t_1)} \frac{\psi(u)}{f(u)} du = +\infty.$$

This contradicts condition (18.5). The proof is complete.

Theorem 1 If assumptions (18.3), (18.5) and (18.6) are satisfied, then (18.1) is oscillatory.

Proof From the proof of Lemma 1 and Lemma 2 it follows that $S^+ = S^{WO} = \emptyset$ for (18.1). Thus, for every solution $x(t)$ of (18.1), we have $x(t) \in S^O$, i.e., (18.1) is oscillatory.

Theorem 2 Let condition (18.3) hold and if

$$\lim_{t \rightarrow +\infty} \int_{t_0}^t \frac{1}{a(s)} ds = +\infty \quad (18.7)$$

is satisfied, then (18.1) is oscillatory.

Proof From Lemma 1 it follows that $S^+ = S^{WO} = \emptyset$ for (18.1). Then in order to complete the proof, it suffices to show that $S^- = \emptyset$ for (18.1). Let $x(t)$ be a solution of class S^- of (18.1). There is no loss of generality in assuming that there exists $t_1 \geq t_0$ such that $x(t) > 0$, $x'(t) < 0$ for all $t \geq t_1$. It follows from (18.3) that there exists $t_2 \geq t_1$ such that

$$\int_{t_1}^{t_2} [q(s) - p(s)] ds = 0, \quad \int_{t_2}^t [q(s) - p(s)] ds \geq 0, \quad t \geq t_2$$

Integrating (18.1) from t_2 to t , we have

$$\begin{aligned} & a(t)\psi(x(t))x'(t) \\ &= a(t_2)\psi(x(t_2))x'(t_2) + \int_{t_2}^t P(s, x(s), x'(s)) ds - \int_{t_2}^t Q(s, x(s)) ds \\ &\leq a(t_2)\psi(x(t_2))x'(t_2) + \int_{t_2}^t p(s)f(x(s)) ds - \int_{t_2}^t q(s)f(x(s)) ds \\ &= a(t_2)\psi(x(t_2))x'(t_2) + \int_{t_2}^t [p(s) - q(s)]f(x(s)) ds \\ &= a(t_2)\psi(x(t_2))x'(t_2) + f(x(t)) \int_{t_2}^t [p(s) - q(s)] ds \\ &\quad + \int_{t_2}^t f'(x(s))x'(s) \int_{t_2}^s [q(\tau) - p(\tau)] d\tau ds \\ &\leq a(t_2)\psi(x(t_2))x'(t_2) = k(k < 0) \end{aligned}$$

Consequently, for all $t \geq t_2$, we have

$$\int_{x(t_2)}^{x(t)} \psi(u) du \leq k \int_{t_2}^t \frac{1}{a(s)} ds,$$

which, because of (18.7) and $k < 0$, implies

$$\lim_{t \rightarrow +\infty} \int_{x(t)}^{x(t_2)} \psi(u) du = +\infty,$$

and so a contradiction since $\lim_{t \rightarrow +\infty} x(t)$ exists and finite and ψ is continuous. The proof is complete.

Remark 3 If $R(t, x(t), x'(t)) = 0$, $Q(t, x(t)) = q(t)f(x(t))$ is satisfied, from Theorem 2, we can obtain the first result of Theorem 3 in [3].

18.3 Asymptotic Behavior

Theorem 3 *If assumption (18.3) is satisfied, then all nonoscillatory solutions of (18.1) can be divided into the following two classes:*

$$A_C : x(t) \rightarrow C \neq 0, t \rightarrow +\infty, \quad A_0 : x(t) \rightarrow 0, \quad t \rightarrow +\infty.$$

Proof Let $x(t)$ be a nonoscillatory solution of (18.1). Then $x(t) \notin S^0$, from Lemma 1 it follows that $S^+ = S^{W0} = \emptyset$ for (18.1). Thus, we have $x(t) \in S^-$.

If $x(t)$ is eventually positive, then there exists $t_1 \geq t_0$ such that $x(t) > 0, x'(t) < 0$ for all $t \geq t_1$. Since $x'(t) < 0$ for all $t \geq t_1$, we obtain $x(t)$ is decreasing and has lower bounded, so $\lim_{t \rightarrow +\infty} x(t)$ exists and $\lim_{t \rightarrow +\infty} x(t) = C \geq 0$, where C is a constant.

If $x(t)$ is eventually negative, then $x'(t) > 0$ for $t \geq t_1$. So $x(t)$ is increasing and has upper bounded, then $\lim_{t \rightarrow +\infty} x(t)$ exists and $\lim_{t \rightarrow +\infty} x(t) = C \leq 0$, where C is a constant.

From the above, every nonoscillatory solution of (18.1) should belong to the class of either A_C or A_0 . The proof is complete.

Theorem 4 *If condition (18.3) holds, and (18.1) has a nonoscillatory solution $x(t)$ in the class A_C (i.e., $\lim_{t \rightarrow +\infty} x(t) = C \neq 0$), then*

$$\int_T^{+\infty} \frac{1}{a(s)} \int_T^s [q(\tau) - p(\tau)] d\tau ds < +\infty, \tag{18.8}$$

for sufficiently large $T \geq t_0$.

Proof Let $x(t)$ be a nonoscillatory solution of type A_C of (18.1). Without loss of generality, we assume that $C > 0$, hence, $x(t)$ is eventually positive. As in the proof of Theorem 3, then there exists $T \geq t_0$, such that $x(t) > 0, x'(t) < 0$ for all $t \geq T$. As in the proof of Lemma 1, we obtain (18.4), i.e.,

$$\frac{a(t)\psi(x(t))x'(t)}{f(x(t))} \leq M - \int_T^t [q(s) - p(s)]ds,$$

for sufficiently large $T \geq t_0$, where $M = \frac{a(T)\psi(x(T))x'(T)}{f(x(T))}$. Hence

$$\frac{\psi(x(t))x'(t)}{f(x(t))} \leq -\frac{1}{a(t)} \int_T^t [q(s) - p(s)]ds.$$

Integrating the above inequality from T to t , then

$$\int_{x(T)}^{x(t)} \frac{\psi(u)}{f(u)} du \leq - \int_T^t \frac{1}{a(s)} \int_T^s [q(\tau) - p(\tau)]d\tau ds.$$

Letting $t \rightarrow +\infty$. Then

$$\int_{x(T)}^C \frac{\psi(u)}{f(u)} du \leq - \int_T^{+\infty} \frac{1}{a(s)} \int_T^s [q(\tau) - p(\tau)]d\tau ds.$$

Noting that $x(T) > C > 0, \frac{\psi(u)}{f(u)} > 0$, we have

$$\int_T^{+\infty} \frac{1}{a(s)} \int_T^s [q(\tau) - p(\tau)]d\tau ds \leq \int_C^{x(T)} \frac{\psi(u)}{f(u)} du < +\infty,$$

then (18.8) holds. The proof is similar for $C < 0$. The proof is complete.

Theorem 5 If conditions (18.3) and (18.5) hold, and if (18.1) has a nonoscillatory solution $x(t)$ in the class A_0 (i.e. $\lim_{t \rightarrow +\infty} x(t) = 0$), then (18.8) holds for sufficiently large $T \geq t_0$.

Proof Let $x(t)$ be a nonoscillatory solution of type A_0 of (18.1). Without loss of generality, we may assume that $x(t)$ is eventually positive. As in the proof of Theorem 4, we have

$$\int_{x(T)}^{x(t)} \frac{\psi(u)}{f(u)} du \leq - \int_T^t \frac{1}{a(s)} \int_T^s [q(\tau) - p(\tau)]d\tau ds.$$

Letting $t \rightarrow +\infty$, from $x(t) \rightarrow 0, x(T) > 0$ and condition (5). Then

$$\int_T^{+\infty} \frac{1}{a(s)} \int_T^s [q(\tau) - p(\tau)]d\tau ds \leq \int_0^{x(T)} \frac{\psi(u)}{f(u)} du < +\infty.$$

Thus (18.8) holds. For $x(t)$ is eventually negative, the proof is similar. The proof is complete.

Remark 4 From the above three theorems (Theorem3–5), we obtain Theorem 1 in this paper.

18.4 Example

Example Consider the equation

$$\left(\frac{1}{2t^{\frac{2}{3}}} x^{-\frac{2}{5}}(t) x'(t) \right)' + \frac{1}{t^{\frac{2}{3}}} x^{\frac{1}{5}}(t) + \frac{1}{4(1+t)^{\frac{2}{3}}} x^{\frac{1}{5}}(t) [x'(t)]^2 = 0 (t > 0), \quad (18.9)$$

where $a(t) = \frac{1}{2}t^{-\frac{1}{3}}$, $\psi(u) = u^{-\frac{2}{5}}$, $q(t) = t^{-\frac{1}{3}}$, $p(t) = 0$, $f(u) = u^{\frac{1}{5}}$. It is easy to prove that (18.9) satisfies all conditions of Theorem 1, so we can prove that (18.9) is oscillatory.

References

1. Ladde GS, Lakshmikantham V, Zhang BG (1987) Oscillation theory of differential equations with deviating arguments. Marcel Dekker, New York
2. Yan J (1992) Oscillation theory of ordinary differential equations. Shanxi education press, Taiyuan
3. Cecchi M, Marini M (1992) Oscillatory and nonoscillatory behavior of a second order functional differential equation. Rocky Mt J Math 22:1259–1276
4. Wintner A (1949) A criterion of oscillatory stability. Quart Appl Math 7:115–117
5. Rogovchenko YV (2000) On oscillation of a second order nonlinear delay differential equation. Funkcial Ekvac 43:1–29
6. Ohriska J (2008) Oscillation of second order linear delay differential equations. Cent Eur J Math 6(3):439–452
7. Cakmak D (2008) Oscillation criteria for nonlinear second order differential equations with damping. Ukr Math J 60(5):799–809
8. Zhang Q, Yan J (2004) Oscillatory behavior of a second order nonlinear differential equation with damping. J Sys Sci Math Sci 24:296–302 (in Chinese)
9. Yan J, Zhang Q (1993) Oscillatory theorems for second order nonlinear differential equations with damping. J Sys Sci Math Sci 13(3):276–278 (in Chinese)
10. Yan J (1986) Oscillation theorems for second order linear differential equations with damping. Proc Amer Math Soc 98:276–282
11. Philos CG, Purnaras IK (1992) On the oscillation of second order nonlinear differential equations. Arch Math 59:260–271
12. Zhang Q, Wang L (2010) Oscillatory behavior of solutions for a class of second order nonlinear differential equation with perturbation. Acta Appl Math 110:885–893

Chapter 19

A Expanded Binary Tree Method of Automatic River Coding and Algorithm

Ji-qi Deng, Huang-ling Gu and Xiao-qing Luo

Abstract As the tree morphological characteristics of river-net, binary tree code cannot reflect the tributary and main bifurcation of river, the binary tree coding method is expanded, which some problem of the tributary and main bifurcation of river and the topological structure is solved and the usability of river coding is also improved, by discussing the key technology and coding method of the reaches and nodes between the reaches and establishing the topological structure of river. Generating geometric network with the network model of ArcGIS Geodatabase, the expanded river coding method is applied to Xiangjiang River basin, and the platform of rivers visualization and automation coding has been realized by using ArcGIS Engine and C#, so that the pollution sources tracking has a hydrological data foundation.

Keywords River · Binary tree coding · Topological structure · ArcGIS Engine

19.1 Introduction

River is a complex network, according with the component of network rules. It has important significance to automatically river coding by setting up river network in accordance with the network rules, which is also the river management in actual work needs. River coding is an important foundation work in hydrology management information construction, which is the basic premise of hydrological management information. So far mainly rivers coding method has two kinds,

J. Deng (✉) · H. Gu · X. Luo
School of Geosciences and Info-Physics, Central South University,
Changsha 410083, China

which one is generated simulated river network by digital elevation model data, then completed rivers coding through the generated network, such as the Yu and Chen Hunhe River coding of Liaohe River basin [1], and the Ren and Liu irrigation river coding of Huaihe River basin [2]. But because the accuracy of digital elevation model data is not high and the generation algorithm of simulated river has defects, it makes many difficulties of the preparation and application of river coding. The other kind of river coding method is to generate binary tree structure of river, establish tree hierarchies' relationship of river, and then encode by traversing binary tree in-depth, such as the Li et al. coding of Lhasa River Basin [3]. But this coding method need divide the basin above all, and the river coding could not differentiate the tributary from mainstream and the integrity, besides, the application process cannot locate the position of river by river coding all alone for the whole basin.

Currently, the heavy metal pollution of Xiangjiang River basin is very serious. In order to analyses, evaluate and forecast the heavy metal pollution effectively, the rivers hydrological features data is indispensable to track pollution. Only with the hydrological data support, it can quickly judge the pollution location, so, firstly it needs to resolve the complex topology relationships between nasty and branch, and therefore the river coding is the basic of pollution analysis critical. According to the hydrological characteristics of Xiangjiang River basin, a expanded binary tree coding method of river network is put forward, and a set of automatic coding system has been developed based on the coding method, which is applied to Xiangjiang River successfully and laid a foundation of tracking model of the basin pollution.

19.2 The Analysis of River Coding Method

Currently river coding method basically has two kinds. The first kind is identifying the river features, for example the river flow, valley area, important degree, etc., and it is mainly used in river conservancy, river management; the second kind is used in basin conservancy environmental protection and focused on stream topologies, which is used to distinguish the upstream and downstream relationship as to judge the influence range of river pollution quickly. The binary tree method of river coding is a new approach of river coding. It puts the binary tree structure in the area of computer data structure into the river coding, Compared with other river coding method, binary tree coding can realize efficient topological operation, and can be expanded to make binary tree structure by increasing the nodes for irregular binary tree river. To the river in Fig. 19.1 of left side, the binary tree coding can be gotten by established the binary tree structure of river, as shown in Fig. 19.1 of right side, and the root node's value is one, which the coding rules fulfill the formula that is $Left = 2 * Parent$, $Right = 2 * Parent + 1$, so the results in Fig. 19.1 can be available. In other words, starting from any node, the upstream node can be queried according to the formula $Left = 2 * Parent$,

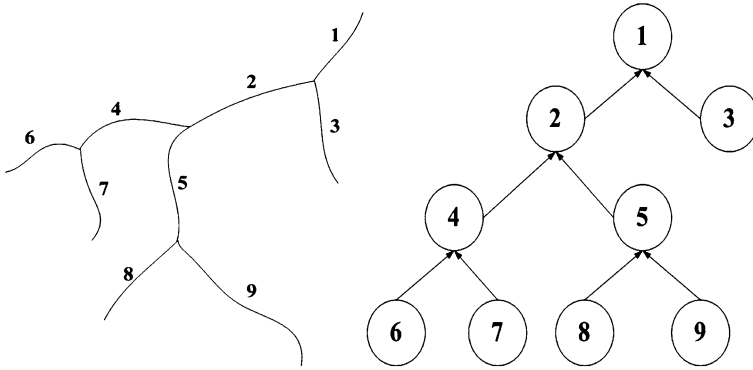


Fig. 19.1 The binary tree of river network

Right = $2 * \text{Parent} + 1$, similarly, starting from any node, the downstream node can be queried according to the formula $\text{Parent} = \lceil (\text{Left or Right}) / 2 \rceil$. This binary tree structure has simple, clear and efficient mathematical relation, so as to increase the efficiency of the river coding and traversing.

The length of Xiangjiang River basin is 856 km, and the drainage area is 94,600 km², and there are 1,300 tributaries along Xiangjiang River, whose main tributaries are Xiao River, Chunling River, Lei River, Mi River, Zhen River and Lian River etc. Above Yongzhou of Xiangjiang River is upstream and below Hengshan Mountain is downstream, and the flow is from down to up on the map, finally into the Dongting Lake, so the traversing of Xiangjiang River is from downstream to upstream through the Dongting Lake as a start point. For the characteristics of long mainstream and multi-branch, and expressing the river better, the six-place coding method is applied in level one of river. Due to multi-reach of Xiangjiang River, as shown in Fig. 19.2, hence only adopting binary tree structure coding could not accurate, a expanded binary tree coding method is put forward, namely, branching stream is embodied in the binary tree of river, and then the node of binary tree is expanded to keep the overall structure and the river's levels, coded with the upstream directly, meanwhile, in order to identify the relationship between the upstream and downstream of bifurcation, a expanded code is added in the coding process to reflect the bifurcate situation, that is importing "A and B", then reaching its downstream coding in front of the expanded code.

The expanded binary tree coding method chooses binary tree code for the basic coding method, with the coding units of reach, and combining the numbers with letters. According to the generated river network to filter coding, identifying the mainstream and tributaries, then the mainstream and tributaries adopted hierarchical mode is coded by certain coding rules. Starting form the Dongting Lake, along the main river from downstream to upstream, each level of river uses six-place coding, which the top four is to differentiate river and the hind two as the

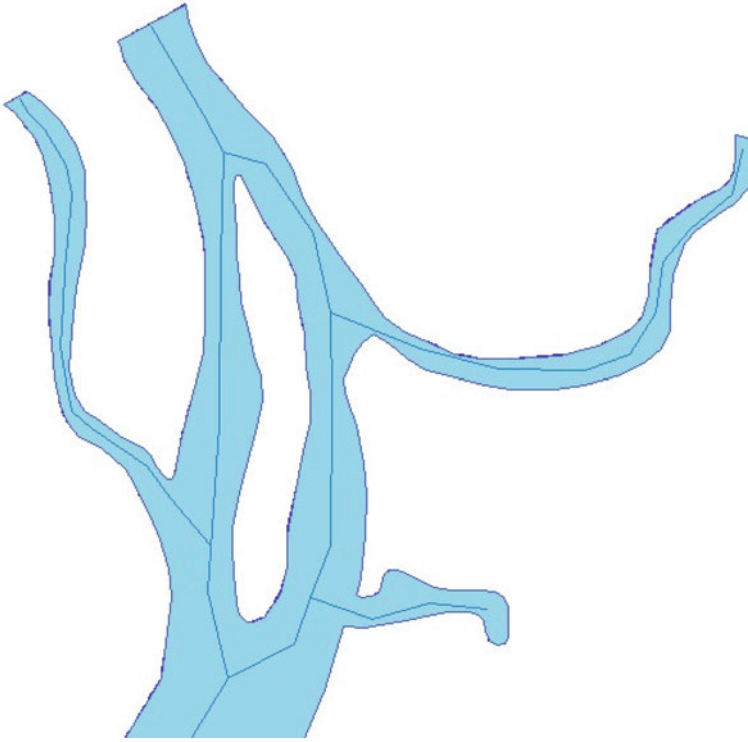


Fig. 19.2 The bifurcation conditions of Xiangjiang River basin

expanded code is to reflect the bifurcation situation, such as “A and B”, then the secondary and level three rivers adopt various river coding stack. With this method, the upstream and downstream can be analyzed by river code data, also, the mainstream and tributaries can be distinguished.

19.3 The Design and Implementation of River Coding Algorithm

19.3.1 The Process of River Coding

The river network data of the Xiangjiang River basin is adopted, whose original data is polygon data, and the main river which describes the whole river network of Xiangjiang River is extracted through the centre line of the polygon data. The river-net structure uses the network model of ArcGIS Geodatabase, generating geometric network from extracting linear elements. Before the river coding, it

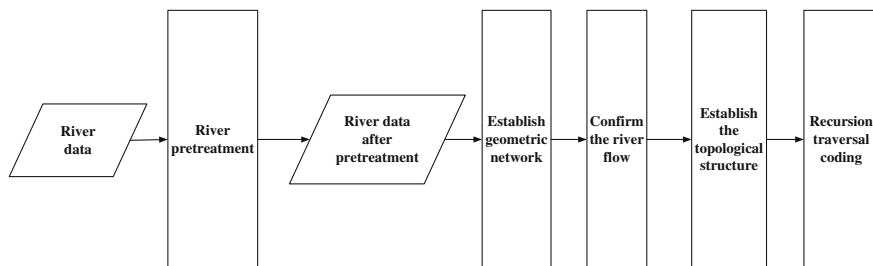


Fig. 19.3 River coding flow chart

needs to judge the flow, then identify the river-net topology and establish complete the data structure of river-net topology. Finally, the coding rules are in accordance with the recursive coding, as the coding process is shown in Fig. 19.3.

19.3.2 The Judgment of Upstream and Downstream

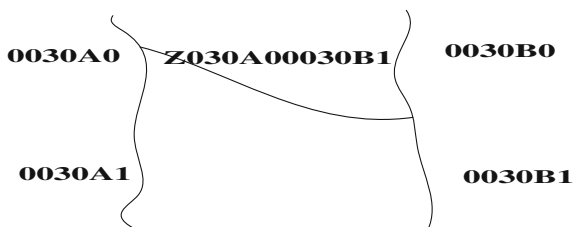
The angle between the mainstream and tributaries can reflect the rivers' connection relationship in "quantity". In the tree river system, mainstream and tributaries are almost intersected into less than 90° , but in the particular geological conditions, it also can appear individual angle equal or more than 90° . According to the angle between the mainstream and tributaries can judge the flow [4].

The flow can be confirmed by calculating the angle between each reach at the same node, thus, the mainstream can be gotten through certain flow, and the main tributary flowed to the mainstream can be determined accordingly.

19.3.3 The Steps of River Coding

The river coding should establish the tree hierarchies relations, which needs extract the mainstream and tributaries. The mainstream can be identified according the principle of length priority and 180° approximation [4]. The specific river coding method is searching from downstream to upstream beginning from the Dongting Lake, if encountering the bifurcate, then comparing the length and the angle between the river, so the mainstream can be determined by the river of the longer length and more close to 180° , according to the method to trace the source of the mainstream along the upstream, so the whole mainstream can be determined finally. Then the tributary rivers in different levels can be gained through traversing each branch node in mainstream according with the 180° in principle. With the method above can build the whole basin tree hierarchies

Fig. 19.4 “H”-form river coding method



relations. The river can be coded based on the tree hierarchies relations, and the specific coding rule is shown as follows:

- (1) Starting from the Dongting Lake, searching from downstream to upstream, the code have six-place, which the first four is to differentiate the reach and the first five used to appear the bifurcate situation extended.
- (2) If the level one of river that has been coded already has lower river, then the level two of river uses twelve-place code from downstream to upstream, which the former six places stands for the first-degree code and the posterior six places is for the corresponding level code.
- (3) The tributary rivers are coded according the river into the mainstream, namely, after the code of mainstream adds a “1”, then the corresponding coding adopts the step (2) above.
- (4) If the rivers are “H” shape river, as the Fig. 19.4 shows, the left and right side of rivers are coded considering not the “connected” river, then the “connected” river can be included into the edge river of “import” (the left “edge river” of Fig. 19.4), coding the “connected” river which the first assignment is “Z” and the code of the upstream close to another “edge river” (the right “edge river” of Fig. 19.4) which remitting the “connected” river is after the first assignment. As shown in Fig. 19.4, the rivers at right side code are 0030B0, 0030B1; the rivers at left side are 0030A0, 0030A1; the “connected” river code is Z030A00030B0.
- (5) With the same coding rule, the river can be filtered coding until that the entire rivers have code.

19.3.4 The Data Structure of River Network Topology

The river coding is based on the topology relation, but the actual river-net topology structure is very complex, which a river has several reaches and each node has several reaches. The river funneling point node and the river reach edge can describe the topology structure of river network. The structure can be organized the nodes and edges, so that each node knows its connected edge and each edge knows its specific information of start and end nodes. The definition of the structure of node and edge are shown in Tables 19.1 and 19.2.

Table 19.1 Edge data structure

Property	Describe
EdgeID	The ID of Edge
StartNode	The start point of Edge
EndNode	The end point of Edge
Direction	The sign of positive direction searched
ReverseDirection	The sign of reverse direction searched
EdgeCode	The code of edge
Length	The length of edge

19.3.5 Recursive Code

The binary tree river coding is a recursive process, which the upstream reach should be found by the node of downstream every time and then the mainstream and branch should be divided and coded until that all the rivers are traversed. The recursive function is shown as follows:

```
private string RiverCoding(pNode startNode, string preCode)
{
    foreach(startNode in _nodes)//Traversing the nodes
    {
        if(startNode.EdgeCol.Count == 1) startNode. _EdgeCol[0]. _EdgeCode =
        "001000";//If the current node has only one edge, the edge can be coded directly
        foreach(pEdge edge in startNode.EdgeCol)//Choose any one of the interlinked
        edges with the node to search
        {
            ArrayList tempNodeCollection = new ArrayList();//store the node
            ArrayList tempEdgeCollection = new ArrayList();//store the edge
            tempNodeCollection.Add(startNode.NodeID);//Record the ID of node as
            the starting point
            pEdge tempEdge = edge;//Current searched edge as the current edge
            tempEdgeCollection.Add(tempEdge.EdgeID);//Record the ID of current
            edge
            //Judge whether the current edge tempEdge has already been searched at
            positive and reverse direction, if it has been traversed, then the search can be given
            up
            if(this.CompareCurEdgeInEdges(tempEdge))
            {
                tempNodeCollection = null;
                tempEdgeCollection = null;
                continue;
            }
            pNode SNode = _nodes[edge.EndPoint];//The end of the current
            edge as the current point
```


Table 19.2 Node data structure

Property	Describe
NodeID	The ID of node
HasProcess	The searched sign
PreNodeID	The precursor node
EdgeCollection	Connected sideline
EdgeNodeAngleCollection	The angle of node and connected edge
NodeCode	The code of node

```

tempNodeCollection.Add(edge.EndPoint);
SNode.Code = tempEdge. _EdgeCode + "1000";
RiverCoding(SNode);
}
}
return;
}

```

19.3.6 Algorithm

For the expanded binary tree coding method above, an automatic coding system has been developed by using C# language with the basic of ArcGIS Engine and the platform of Microsoft.NET, and the river code is realized automatically, and the code of Xiangjiang River basin has been achieved through the system, so it is verified the effectiveness of this river coding method. The coding result of Xiangjiang River basin is shown in Fig. 19.5.

From Fig. 19.5, the expanded binary tree code desirable the internal structure of river characteristic, in the coding method, the river network well reflect the partial and whole relationships between the mainstream and tributary.

19.4 Conclusion

As the foundation of binary tree coding system and the complex characters of bifurcation rivers, it puts forward a feasible expanded binary tree coding method, establishing the river topology structure using the geometric network model of ArcGIS Geodatabase, with the length priority and 180° approximation principle to identify the mainstream and tributaries, and distinguish the mainstream from the code, so the bifurcation and special river can be solved better by obtaining the definition coding rule and the whole river coding properly shows the river valley features well. Finally developing automatic coding system, the code of Xiangjiang

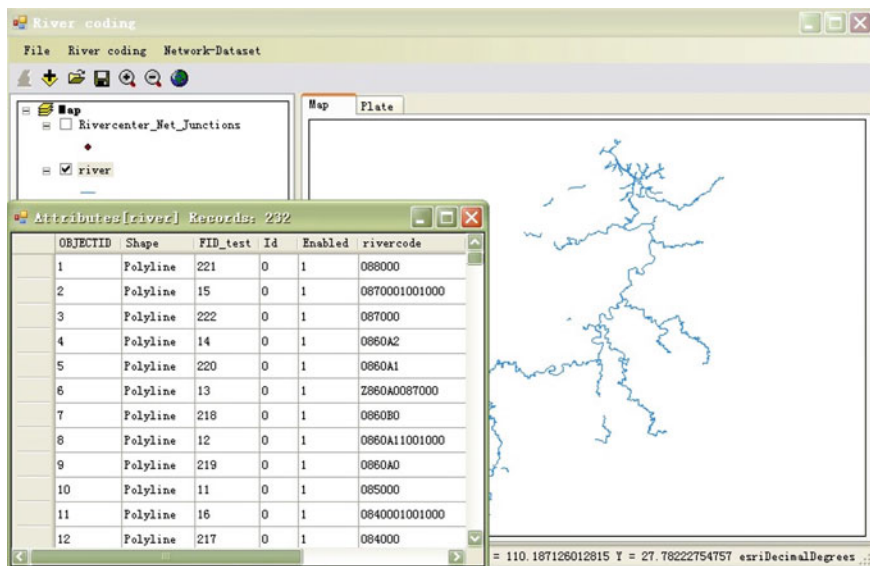


Fig. 19.5 The coding result of reach coding system

River has been realized through the system; meanwhile, it certifies the effectiveness of the coding method. But because the flow is determined by the length priority and 180° approximation principle, it is not necessarily corrected in practice, so this coding method can only roughly description of river structure.

Acknowledgments We thank the bureau of Hunan Provincial Environment Protection.They provide the data and requirements of front-line workers. Otherwise, the chapter is supported by Sub-topics of major national science and technology projects (2009ZX07212-001-06) and Natural Science Foundation of Hunan Province (07JJ6074).

References

1. Yu M, Chen X (2009) River coding river coding based on DEM [J]. *Yangtze River* 40(24):36–38
2. Ren L, Liu X (1999) Application of digital elevation model to topological evaluation of drainage system [J]. *Adv Water Sci* 10(2):129–134
3. Li T, Wang G, Liu J (2006) Drainage network codification method for digital watershed model [J]. *Adv Water Sci* 17(5):658–664
4. C Zhao (2004) Research on Horton code and graphic generalization of catchment [D]. Wuhan University, Wuhan
5. Wang G, Wu B, Li T (2007) Digital yellow river model [J]. *J Hydro-environ Res* 1(1):1–11
6. Li J, Cao G, Yu X (2010) Overview of river coding technology at home and abroad [J]. *Water Res Inf* 2:25–30

7. Shao Y, Li J (2005) The topology analysis application of the data model of geodatabase and the geometric network [J]. *Eng Surv Mapp* 14(1):17–19
8. Mitia B, Andrej V (1995) Watershed coding of large river basins [C]. *Model Manag Sustain Rasin-Scale Water Res Sys* 347–351
9. Josef F, Thomas H (2009) Coding of watershed and river hierarchy to support GIS-based hydrological analyses at different scales [J]. *Comput Geosci* 35(3):688–695

Chapter 20

Domain-Specific Ontology Mapping Based on Common Property Collection

Bo Jiang, Yunzhao Cheng and Jiale Wang

Abstract In order to improve the quality of ontology mapping, it's critical for researchers to take important information of ontology into account. In this paper, we propose a novel approach for ontology mapping by comparing the common properties of concepts in different ontologies. Our approach is to find the corresponding concepts based on the statistical information of the common properties. The property collections of concepts are obtained by defining a domain-specific property corpus, which contains all the properties most widely accepted. This enables us easily to detect common information for mapping. The experiment shows the efficiency and effectiveness of the proposed approach is satisfactory.

Keywords Ontology mapping · Property collection · Property corpus

20.1 Introduction

Ontologies serve as a mean for information sharing and capture the semantics of a domain of interest. However, building a single, unifying model of concepts and definitions is neither efficient nor practical. A more practical assumption to reach the

B. Jiang (✉) · Y. Cheng · J. Wang
School of Computer Science&Information Engineering, Zhejiang Gongshang University,
Hangzhou, 310018, Zhejiang, China
e-mail: nancybjjiang@zju.edu.cn

Y. Cheng
e-mail: chengyunzhao20@163.com

J. Wang
e-mail: wjl8026@mail.zjgsu.edu.cn

purpose of interoperation between multiple, heterogeneous of ontologies is based on the communication between ontologies through ontology mapping approaches.

Ontology mapping is the process of linking corresponding terms from different ontologies. The most common request is to find the concepts similar to each other.

The tasks of ontology mapping are commonly performed manually by domain experts. The manual task could be time-consuming and inefficient. In order to make computer programs could solve the mapping works automatically, different kinds of matching approaches have been proposed [1] Property is one of the most important kinds of information for a concept. In this paper, the properties are taken as the basic proofs for ontology mapping.

In this paper, we propose a new approach to solve the mapping problem between two ontologies with less calculation and higher accuracy by comparing the properties of concepts.

20.2 Related Work

The existing ontology mapping approaches can be classified according to different semantic levels.

Literal level: The most commonly used method here is edit distance calculation [4]. It compares the outlooks of the words but pays little attention to the semantic meanings.

Semantic level: The methods in this level mainly focus on the calculation of the semantic similarity. WordNet is widely used as a common document to find the synonyms. Cosine Similarity Measure, another widely used non-Euclidean distance measure of similarity between two vectors by finding the angle between them. It's a very good idea to use this approach combined with common ontologies or a corpus in this domain. In [5] they take the corpus as a bridge for ontology mapping.

Structure level: In this level the hierarchies of the ontologies are considered. With the consideration of the hierarchy, the relatedness among concepts could be taken into account.

Some other kinds of solutions, models and frameworks are proposed in [7].

In [11], the author listed 17 rules for ontology mapping, and the rules have been widely adopted in applications. Different approaches may take several of them as their comparing mechanism.

A similar idea is also appeared in [5] by Chin pang Cheng. In that paper, they took industry documentation as a bridge for ontology mapping, and many existing methods were adopted.

20.3 Overview of Mapping Approach

The mapping approach presented in this paper is based on one of the 17 rules [11]: if two concepts with same properties then they are similar. We use the reference engine as our basic tool.

First step: Building the property corpus before mapping work. To compare two concepts from different ontologies, we select the most common properties of them which could perfectly specify the two concepts. Some principles are adopted to build the property corpus. First, the properties must be widely accepted, like birthday, ID number etc. Second, in order to specify one concept, the property must be very professional in those special fields. Taking teacher for example, the teacher number should be the best choice. With the concerns of different property names may present the same thing due to different zones of contexts. The property corpus should be well organized. The synonyms of the properties should be considered and be listed in the corpus.

Second step: Making mapping rules. This step determines how to find two similar concepts based on the property corpus. The property collection only contains the properties included in the property corpus.

20.4 The Implementation of Ontology Mapping

Here are several steps to implement the mapping approach. First we give the definitions of the terms which will be used. Second, the mapping rules will be made. Finally, a mapping demonstration will be performed step by step.

20.4.1 Definitions

Table 20.1

Table 20.1 Term definitions

Symbol	Definition
O_i	Ontology i
$C_{i,j}$	The concept node j of ontology O_i
$Collection(C_{i,j})$	The property collection of $C_{i,j}$, all the properties should from property corpus.
$Range(C_{i,j})$	According to the $Collection(C_{i,j})$ we could determine $Range(C_{i,j})$
IE_i	Inference engine i

20.4.2 Mapping Rules

Rule1: if $Collection(C1) = Collection(C2) \rightarrow Range(C1) = Range(C2)$

Rule2: if $Collection(C1) \subseteq Collection(C2) \rightarrow Range(C1) > Range(C2)$ then visit $C1$'s children get the collection of them and compare the their ranges with $Range(C2)$

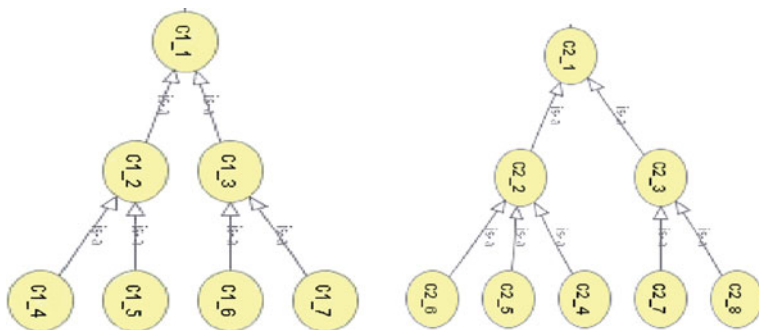


Fig. 20.1 Two ontologies for demonstration

Rule3: if Collection (C1) \supseteq Collection (C2) \rightarrow Range (C1) < Range (C2)

Rule4: if Range (C1) = Range (C2) \rightarrow C1 is similar to C2

Rule5: if Range (C1) > Range (C2) \rightarrow C2 is the child of C1. If C2 is also the child of C2's parents', removed it from its parents'.

In this paper we just want to find the concepts with the same range, namely pairs like this: C1 and C2 with Range (C1) = Range (C2).

20.4.3 Mapping Steps

We built two ontologies O1 and O2 with protégé in order to demonstrate how those rules listed will be used. Figure 20.1 shows the structures of the two ontologies. The properties are not performed. The property corpus is ready before mapping work start.

Here is the demonstration of two ontologies. The ontology O1 is the source ontology on the left side, the other is target ontology O2.

20.4.3.1 Inference Engine

At least two same inference engines are needed. These engines should support basic rules in ontology. In this paper we use Jena embedded in our develop tools, and take OWL as the default ontology language.

According to R10 in 17 Rules [11], concepts are similar, if the properties of the concepts are similar. As we emphasized above, the properties of the concepts will be compared in the collections must from the property corpus, and the rest of them will be ignored (Fig. 20.2).

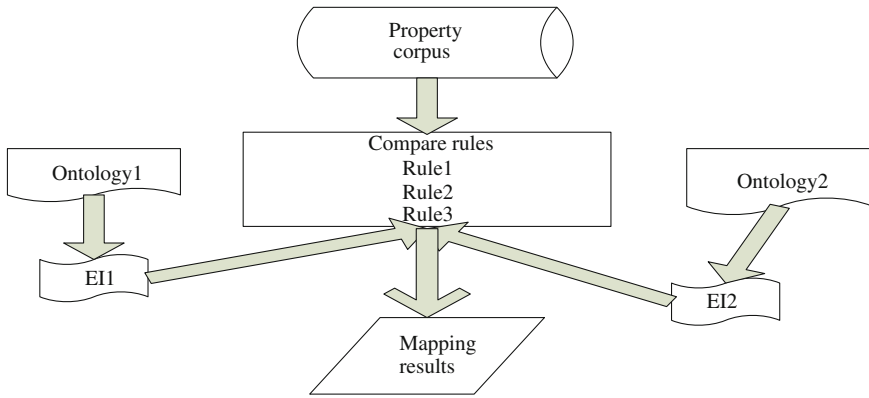


Fig. 20.2 The structure of algorithm

20.4.3.2 Ontology Mapping Process

In this paper we map O1 to O2, so we need to find the concepts in O1 which related or similar to the concepts in O2. In order to make our mapping approach clear and simple, we find C1-4 in O1 similar to C2-6 in O2 before we start. And then we will show how the algorithm works during the mapping procedure. We know that C1-4 similar to C2-6. So we get Collection (C1-4) = Collection (C2-6) \rightarrow Range (C1-4) = Range (C2-6) according to the mapping rules.

In order to reduce the mapping workloads, we proposed a new method to visit the nodes of two ontology trees that is in opposite visiting direction. First, two ontologies O1 and O2 will be taken as the inputs of two inference engines. To O1, the engine visits the nodes of it from bottom to top. That means C1_4 will be first visited so Collection (C1-4) and Range (C1-4) will be got. In contrast, the engine visits O2 from top to bottom, that means C2-1 will be first visited, and Collection (C2-1) and Range (C2-1) will be got. With the help of opposite directions visiting method the efficiency of the algorithm will be remarkable improved.

Here are the specific steps in mapping procedure with EI1 deal with O1 and EI2 deal with O2:

- a. In EI1, it visits C1-4 first, and gets Collection (C1-4).
- b. At the same time, EI2 will visit C2-1, and get Collection (C2-1).
- c. As we already know that C1-4 is similar to C2-6, we could get Collection (C2-1) \subseteq Collection (C1-4) \rightarrow Range (C2-1) $>$ Range (C1-4). Mark C1-4 is the child of C2-1.
- d. In EI2, it visits the children of C2-1, namely C2-2 and C2-3, and get Collection (C2-2) and Collection (C2-3).
- e. Compare Collection (C1-4) with Collection (C2-2) and Collection (C2-3). According to the mapping result, we know Collection (C2-2) \subseteq Collection (C1-4) \rightarrow Range (C2-2) $>$ Range (C1-4). To C2-3, if Range (C2-3) $>$ Range (C1-4) we visit its children C2-7 and C2-8.

- f. Mark C1-4 is the child of C2-2, and remove it from C2-1.
- g. Visit C2-2's children, and get Collection (C2-6), Collection (C2-6), and Collection (C2-6).
- h. Compare Collection (C1-4) with Collection (C2-4), Collection (C2-5) and Collection (C2-6). We could get Collection (C1-4) = Collection (C2-1) \rightarrow Range (C1-4) = Range (C2-6).
- i. According to the rules listed above we get the result C1-4 similar to C2-6 with Range (C1-4) = Range (C2-6).

With the same method, we could get all the mapping information between the two ontologies.

Notes: When we compare Collection (C2_2) and Collection (C2-3) with Collection (C1-4). If Collection (C1-4) \neq Collection (C2-3) we don't have to compare C2-3 and its children with C1-4. In this case, the workload will be reduced. In this paper we only focus on finding two concepts with same range.

20.5 Analysis

The mapping approach presented in this paper not focuses on the similarity calculation, but collecting the matching information through the properties of concepts. We only use properties from corpus, and prove the similarity of the two concepts according to R17 [11].

One of the most important tasks in this paper is the building of property corpus. The property corpus could be organized from general to particular, and different restrictions could be implemented due to different requirements.

In this paper we conduct the ontology mapping work by narrowing the width of concepts from general to particular.

In this paper, the work of property collection comparison could be conducted in different nodes on the web, and as the algorithm go deep most of the branches could be ignored, and we just go to the specific one which the most suitable concept exist. To compare two or more concepts' properties, these tasks could be processed in parallel ways, and it will take full advantage of the computing power in internet.

20.6 Experiment

Both of the ontologies are about the organizations of teachers. The first step is to build the property corpus or use the corpus which is already existed. In this experiment the property corpus contains properties like Teaching_ID (Teacher_ID), Higher_Professor_ID, Professor_ID and University_ID.

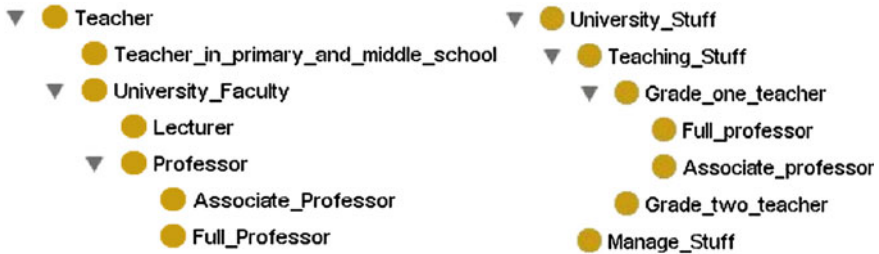


Fig. 20.3 Source ontology O1 and target ontology O2

Fig. 20.4 Part of ontology O1

```

<owl:Class rdf:about="#Professor">
  <rdfs:subClassOf rdf:resource="#University_Faculty"/>
</owl:Class>
<owl:DatatypeProperty rdf:ID="University_ID">
  <rdfs:domain rdf:resource="#University_Faculty"/>
  <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="Higher_Professor_ID">
  <rdfs:domain rdf:resource="#Full_Professor"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="Professor_ID">
  <rdfs:domain rdf:resource="#Professor"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="Teacher_ID">
  <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
  <rdfs:domain rdf:resource="#Teacher"/>
</owl:DatatypeProperty>

```

Fig. 20.5 Part of ontology O2

```

<owl:Class rdf:about="#Grade_one_teacher">
  <rdfs:subClassOf rdf:resource="#Teaching_Staff"/>
</owl:Class>
<owl:DatatypeProperty rdf:ID="University_ID">
  <rdfs:domain rdf:resource="#University_Staff"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="Teaching_ID">
  <rdfs:domain rdf:resource="#Teaching_Staff"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="Higher_Professor_ID">
  <rdfs:domain rdf:resource="#Full_professor"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="Professor_ID">
  <rdfs:domain rdf:resource="#Grade_one_teacher"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="Stuff_ID">
  <rdfs:domain rdf:resource="#University_Staff"/>
  <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
</owl:DatatypeProperty>

```

Here are the demonstrations of two ontologies, O1 is on the left side and the other is O2 (Figs 20.3, 20.4 and 20.5).

Here are parts of two ontologies

Following the steps in last section, we carried out the experiment with the two ontologies above. O1 is the source ontology, and the target ontology is O2. Both of them are from same domain. The concepts of O1 are visited from bottom to top and the O2 on the contrary. EI1 deals with O1 and EI2 deals with O2.

Table 20.2 Mapping results

Source concept similar to->	Concept in targets ontology
Full_Professor	Full_professor
Associate Professor	Associate professor
Professor	Grade_one_teacher
Lecturer	Grade_two_teacher
University_Faculty	Teaching_Staff

We get the collection and range relations between the concepts in ontologies:

- a. Collection (Full_Professor) = Collection (Full_professor) = {Higher_Professor_ID, Professor_ID, Teacher_ID, University_ID}
- b. Collection (Associate Professor) = Collection (Associate professor)
- c. Collection (Professor) = Collection (Grade_one_teacher)
- d. Collection (Lecturer) = Collection (Grade_two_teacher)
- e. Collection (University_Faculty) = Collection (Teaching_Staff)

According to the mapping rules, we get the mapping results: Table 20.2

20.7 Discussion

In this experiment, we conducted our mapping work through the property comparison. When we use the other existing methods of semantic similarity calculating to carry out the mapping experiment the results were Full_Professor to Full_professor and Associate Professor and Associate professor.

So the results of this experiment are more completed and objective compare to other methods.

20.8 Conclusions

The mapping approach presented in this paper based on the comparison of properties of concepts, the first job should be done before mapping work is the building of property corpus. Several principles have been listed to build it, and different restrictions and principles should be adopted according to specific application.

In the mapping procedures, we use two inference engines to visit two ontologies in opposite directions to improve the efficiency of algorithm. Different rules and assertions in ontology could be used are not considered.

Finally, different branches could be processed in parallel; the method will take full advantage of computing power on the web.

Acknowledgments A Project Supported by Scientific Research Fund of Zhejiang Provincial Education Department (No.Y201018008).

References

1. Rahm E, Bernstein PA (2001) A survey of approaches to automatic schema matching. VLDB 10:334–350
2. Shvaiko P (2004) A classification of schema-based matching approaches. Technical report, Informatica e Telecomunicazioni, University of Trento
3. Hvaiko P, Euzenat J (2004) A survey of schema-based matching approaches. Technical report, Informatica e Telecomunicazioni, University of Trento
4. Ristad ES, Yianilos PN (1996) learning string edit distance, Research Report CS-TR-532-96 Oct 1996 Revised Oct 1997
5. Cheng CP, Lau GT, Pan J, Kincho H Law domain-specific ontology mapping by corpus-based semantic similarity. Engineering informatics group department of civil environmental engineering Stanford University, Stanford, CA 94035, USA
6. Pei M, Nakayama K, Hara T. Constructing a global ontology by concept mapping using wikipedia thesaurus. Graduate school of Info. Science and Tech, Osaka University
7. Kalfoglu Y, Schorlemmer M Information-flow-based ontology mapping. Advanced knowledge technologies (AKT) Department of Electronics and Computer Science, University of Southampton
8. Pan R, Ding Z, Yu Y, Peng Y. A bayesian network approach to ontology mapping. Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County
9. Maedche A, Motik B, Silva N, Volz R (2002) MAFRA-A mapping framework for distributed ontologies. Knowledge engineering and knowledge management: ontologies and the semantic web. Volume 2473/2202
10. Besana P, Robertson D (2008) Probabilistic dialogue models for dynamic ontology mapping. Centre for intelligent systems and applications, University of Edinburgh
11. Ehrig M, Sure Y (2004) Ontology mapping-an integrated approach. Apr.2004

Chapter 21

Efficient Mobile Electronic Full Declaration System Based on Smart Device

Xianyu Bao, Qing Lu, Bo Zhao, Weimin Zheng and Yang Wang

Abstract Smart device is an effective tool to support mobile electronic full declaration. This chapter described the hardware realization of smart device based on Intel's Atom processor. According to the characteristics of the electronic full declaration services at Inspection and Quarantine, smart device is improved from the angles of information interaction, client software and system security. Then a general mobile electronic full declaration system is proposed based on smart device. In addition, a concrete application is illustrated with an example of the declaration services at Inspection and Quarantine for exported commodities. Application results validate the proposed system.

Keywords Mobile electronic full declaration services · Commodities declaration · Intel's atom processor · Smart device

21.1 Introduction

In recent years, electronic declaration services have been widely popularized by national inspection and quarantine agent. That has brought great benefits for both enterprises and the inspection and quarantine authorities (IQA). On one hand,

X. Bao (✉) · Q. Lu · W. Zheng · Y. Wang
Shenzhen Academy of Inspection and Quarantine, Shenzhen 518010, China
e-mail: baoxianyu@163.com

X. Bao · W. Zheng
Department of Computer Science and Technology, Tsinghua University, Beijing
100084, China

B. Zhao
Liaoning Entry-Exit Inspection and Quarantine Bureau, Dalian 116001, China

enterprises could declare and revise online to save time; on the other hand, the IQA could improve the level of management and enforcement efficiency, and realize a win-win situation. But with the development of electronic full declaration services, the IQA aims to gradually realize the information full declaration through strengthening all aspects of import and export commodities, including planting and breeding, inspection, production, processing, transportation, customs clearance, warehousing, repackaging, marketing, and so on. Therefore, the higher demand is required to strengthen the current electronic declaration services. It is a very challenging task to develop a mobile electronic full declaration system, which can track commodity locations anywhere and anytime, in order to replace the traditional fixed electronic declaration way. Since 2009, Shenzhen Academy of Inspection and Quarantine has cooperated with Tsinghua University to develop a series of mobile electronic full declaration system, such as mobile Internet devices [1], application system of fruits and vegetables supplied to Hong Kong [2], and virtual gate for Shenzhen ports based on RF technology [3].

To further improve information level of the electronic declaration and realize full declaration in all aspects for import and export commodities, RFID technology and mobile Internet technology should be used. Additionally, smart device based on Intel X86 architecture and Atom processor technology provide effective ways to achieve the developed system.

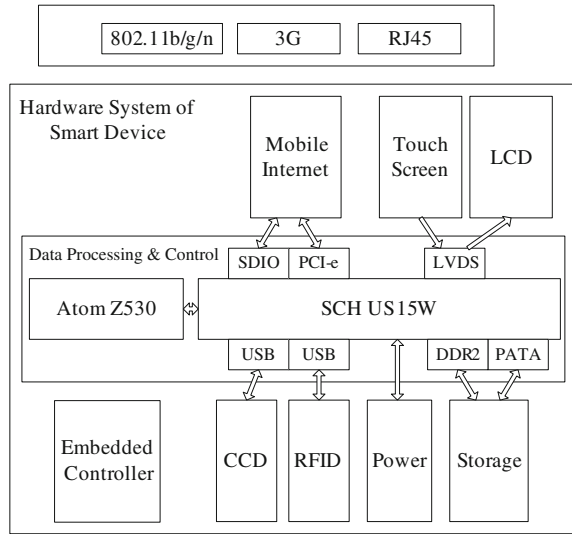
In the light of the above research background, this chapter developed the hardware system of smart device, which has the advantages of high integration, fast speed, Internet access anywhere and anytime, small size, and light weight. Thus it is especially suitable for full declaration for import and export commodities. Information interaction, client software and system security taken into consideration, a general mobile electronic full declaration system is developed based on smart device.

21.2 Paper Preparation

As shown in Fig. 21.1, the hardware system of smart device consists of data processing module based on Atom Z530 [4], peripheral interface control module based on SCH US15 W [5], storage, power management, LVDS input/output, RFID data acquisition, and mobile Internet. The main advantage of the system is that two Intel chips are adopted, which formed X86 architecture.

The system has been widely used in all aspects of the Shenzhen export commodities supplied chain. The specific work principle as follows: data processing and control module mainly completes real-time data computation and peripheral interface circuit control. In the planting and breeding demonstration base, food information is written into the RFID tags of commodity packages by using the touch screen input module and the RFID data acquisition module of the system. At the same time, information is stored into the local system and transferred to the Shenzhen ports' database by mobile Internet module (802.11 a/b/g, 3G, or RJ45).

Fig. 21.1 The hardware system of smart device



When staffs supervise in the factory, they can fill out inspection records through the system. If needed, photos are taken by CCD camera for evidence. When food is being exported, the port staffs use the same system to collect RFID tags information, which is on-the-spot compared with the port database data. If succeeded, the food could be exported to Hong Kong.

The developed system adopts the design methods of Tablet PC and replaces the previous keyboard and mouse (can be external) with full touch screen. Windows XP can be supported in the developed system, which means that it is compatible with most application software. Another advantage is that the RFID data acquisition module supports RFID readers using many kinds of frequency band, such as Low Frequency (LF), Medium Frequency (MF), and Ultra High Frequency (UHF).

21.3 The Function Module Partition and Hardware Design

21.3.1 The Data Process and Control

The module is the core of the hardware system of smart device. Take the demand of low power consumption, computation speed, peripheral logic, protection, storage, mobile Internet communication into comprehensive consideration, we choose Atom Z530 (441 balls, 13 × 14 mm) and bridge-chip SCH US15 W (1,249 balls, 22 × 22 mm). The two chips can communicate each other with high-speed 400/533MT/s FSB.

ATOM Z530 is used for real-time data processing. It adopted 45 nm high-K metal grid transistor manufacturing process and advanced power management technology. Its frequency can be up to 1,600 MHz and thermal design power (TDP) can be low to 2 W (mainstream laptop processor's TDP is 35 W) [6]. SCH US15 W integrated graphics chip has 3D graphics processing and high-definition video hardware decoding of 1080i and 720p. It also integrates most of I/O interfaces of the PC and handheld devices, such as PCI-e, USB, SDIO, LVDS, PATA, and so on.

21.3.2 The Memory

The memory can be divided into inner memory and hard disk. The voltage of inner memory module is 1.8 V, and memory volume can be up to 2 GB DDR2 533 MT/s. When the motherboard power is on, it can load the data from hard disk, CD or other external memory's data through the system bus and open up temporary data buffers used to store temporary data. It realizes fast data exchange between the memory and bridge-chip through its 64 SM_DQ data signal pins and 8 SM_DQS strobe signal pins.

The hard disk uses the parallel PATA interface standard. The transmission speed reaches up to 133 MB/s. This chapter selects the SST85LD1008 M SSD produced by Silicon. The 8 GB disk space can be extended to 64 GB. In order to reduce power consumption, disk was directly mounted to the motherboard with LBGA format.

21.3.3 The LVDS

Low Voltage Differential Signaling (LVDS) [7] is a kind of low-voltage differential signal. It utilized very low voltage swing (about 350 mV) to realize hundreds Mbit/s data transfer rate at one or more pairs of differential line. Due to the low voltage and low current, it has characteristics of low power consumption, low error, low crosstalk and low radiation.

LVDS input/output module is made up of the touch screen input circuit, the video control circuit and the LCD display circuit. Among them, the touch screen input circuit would perceive the information of the resistive touch screen generated by touch position coordinates (X RIGHT, X LEFT, Y UP, Y DOWN), then output to the bridge-chip by TSC2007IPW converter, in order to achieve the positioning signals and instructions implementation. The video control circuit uses four pairs of differential data signal line LA_D [0:3], and a pair of differential clock line LA CLK signal together to achieve the simultaneous transmission of video data. LCD display is done by utilizing LVDS cable to link LVDS with LCD interface.

21.3.4 RFID Data Acquisition

The RFID data acquisition module works as follows. The first step is to read out the contents of RFID tag and output to LCD display. The second step is to write the input data from touch screen into the RFID tag. The module uses two ADF7020^[9] which is half-duplex RF communication chip to realize this function. The chips are both connected to SCH US15 W by mini USB interface. Its work modes and frequency allocation are controlled by external control circuit. In the receive mode, the signals are processed by selection, amplification, demodulation and baseband processing. In the send mode, the baseband signals are sent through the carrier signal after modulation and amplification processing.

21.3.5 Mobile Internet

A major feature of the developed handset is to connect to the Internet anytime and anywhere. The reason is that the module supports both wireless and wired Internet connection, including built-in 802.11, built-in 3G, and RJ45 Ethernet interface. All data collected by the RFID data acquisition module will be transmitted in time through the mobile Internet to the Shenzhen port supervision database, in order to realize data-recording and port inspection.

21.3.6 Power Management

The developed handset consists of processor, bridge-chip, inner memory, hard disk, touch screen, LCD, PCI-e, SDIO, USB, and other digital and analog circuits. In order to make these functional modules with different power requirements, the refinement needs and power conversion efficiency of each module are considered. The developed handset utilized near 20 groups power supply which range from 0.9 to 5 V, shown in Fig. 21.2.

21.3.7 Embedded Controller

Embedded controller (EC) is a unique part of smart device. The EC module uses IT8502E-J chip produced by ITE to control power-on time of the motherboard. Meanwhile it provides different power management policies for energy-saving in the S3, S4 or S5 state with the help of bridge-chip. In addition, EC module can control the charge and discharge time of battery, fan speed varies and other functions by SMBus. It is worth to mention that most tasks of EC are cooperated with the BIOS code.

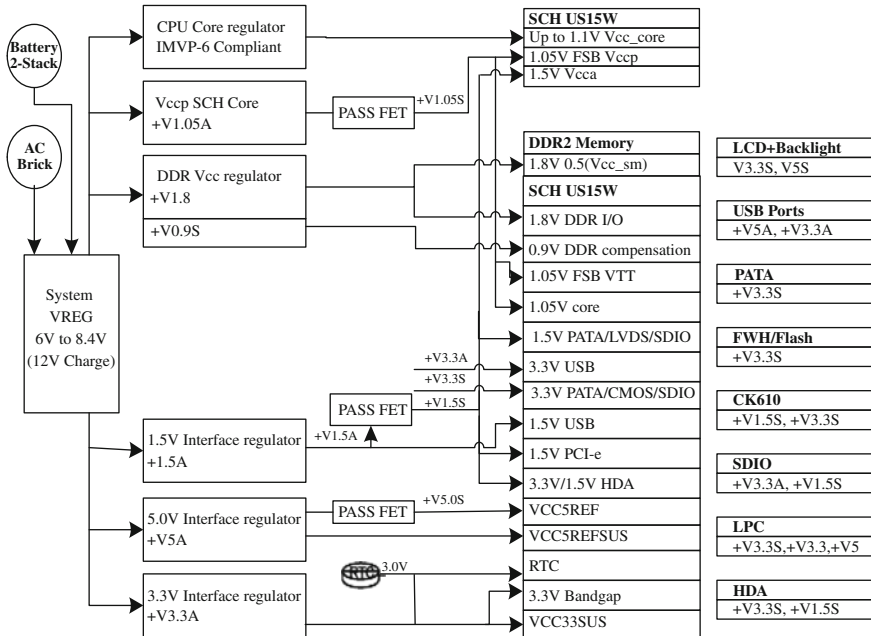


Fig. 21.2 Power management

21.4 The Mobile Electronic Full Declaration System Based on Smart Device

This section builds a general mobile electronic full declaration system based on smart device. The system could help import and export enterprises easily complete the services interaction with the IQA, as shown in Fig. 21.3. The system can be divided into services module and control module. The control module manages and maintains the declaration database, and the services module exchanges the information with the IQA staffs. Nowadays, the function of the services module includes declaration services, supervision services and certificate of origin services.

As smart device is a mobile electronic declaration terminal, the compatibility with the existing electronic declaration database must be taken into consideration. Therefore smart device have to be improved from the angles of information interaction, client software and system security. Information interaction is that, in the process of receive, transmit, transfer and management, the IQA staffs communicate with the declaration customers by means of voice, text, video on smart device. Client software is developed based on the Java platform. It can exchange data with the declaration database and allow that clients login the system and do declaration anywhere and anytime. In addition, in case of the growing network

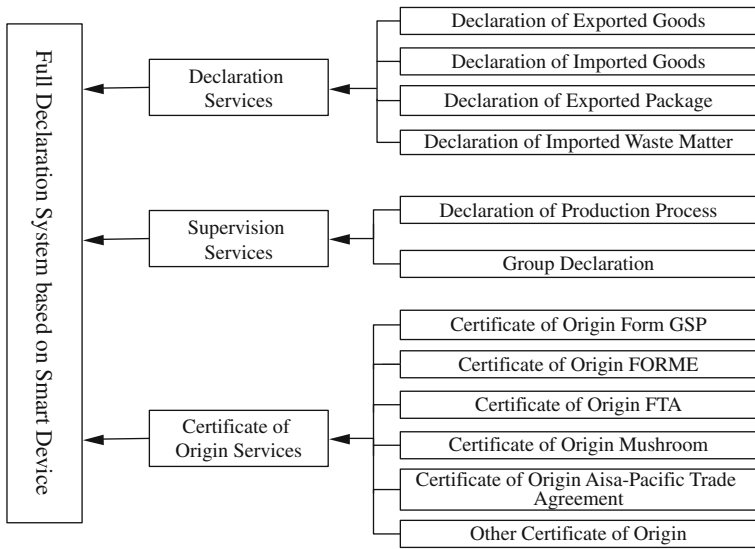


Fig. 21.3 The mobile electronic full declaration system based on smart device

security threats, the security of the electronic full declaration system is an important problem. To solve the problem, we use the mobile access network and network security barrier between the public network and the inner network. Meanwhile advanced ID authentication, access management and end to end VPN technology are used to enhance the system security.

21.5 The Applications of Declaration for Exported Commodities

Take the declaration services for exported commodities for example, as shown in Fig. 21.4, enterprises login electronic full declaration system based on smart device. Firstly, various communication parameters and company’s basic information are set by default contents. When a document is created, the default information could help reduce time of data input and improve work efficiency. Secondly, the declaration module for exported commodities can automatically complete the design of the exported electronic declaration form, and generate electronic message to wait to be sent. Once the enterprise’s authentication passes, enterprises can access to the network of the IQA. By utilizing the data exchange platform, they can send the electronic declaration E-mail to the related authority. Thirdly, the communication unit receives the data from the data exchange platform and automatically verifies the data. If verification passes, the checking authorities

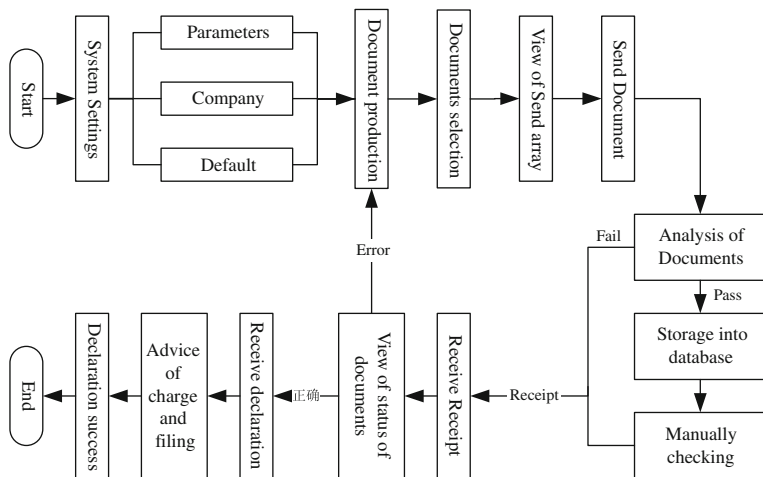


Fig. 21.4 The flow diagram of full declaration services for exported commodities

manually review the data by Computer Management System of the IQA. Lastly, enterprises receive the receipt on smart device. When the error information is received, they need to modify the error and resent the data, in order to complete the declaration.

As the declaration terminal is not constrained by time and space, the workspace of the enterprises’ staff can be theoretically extended to any space. No matter at home or on the way to company, the staff can declare online anywhere and anytime. Thus time and space are reasonably and effectively used. At the same time, the terminal greatly enhances the efficiency of the IQA.

21.6 Summary

In this chapter, a low-power, high-performance smart device based on Atom Z530 and SCH US15 W was developed for mobile electronic full declaration. The hardware design of data process and control, memory, power, RFID data acquisition, mobile Internet is given. Then a general mobile electronic full declaration system is proposed by improving information interaction, client software and system security. In the end, the applications of the declaration for exported commodities demonstrate the effectiveness of the proposed system.

Acknowledgments This chapter is supported by the National 863 Program of China (2008AA04A109), General Administration of Quality Supervision, Inspection and Quarantine’s Special Funds for Scientific Research on Public Causes (200810327), Shenzhen—Hong Kong Creative Funding (08LH-05), and the Doctor Science Research Fund of Shenzhen Entry-Exit Inspection and Quarantine Bureau of China (SZ2009103).

References

1. Bao X, Lu Q, Zheng W (2010) Design of embedded mobile internet device based on atom. CACIS 2010, USTC Press, pp 91–95
2. Lu Q, Wang X, Liu S (2009) Research on RFID supervision technology for vegetables supplied to Hong Kong. J Plant Quar 23:32–34
3. Lu Q, Ding X, Chen X (2010) Design and realization of virtual gate system based on RF technology. J Mod Electron Tech 31:139–144
4. Intel corporation: Intel ATOM Processor Z5×× Series. Rev.005, (2008) <http://download.intel.com/design/processor/specupdt/319536.pdf>.
5. Intel corporation: Intel System Controller Hub (SCH). Rev.2.0, (2008) <http://download.intel.com/design/processor/specupdt/364236.pdf>
6. Gerosa G, Curtis S, Addeo M (2009) A sub-2 low power IA processor for mobile internet device in 45 nm high-k metal gate CMOS. IEEE J Solid-State Circuits 44:73–82
7. Chen M, Jose SM, Michael N (2005) Low-voltage low-power LVDS drivers. IEEE J Solid-State Circuits 40:472–479

Chapter 22

The Predicate System Based on Schweizer–Sklar t-Norm and Its Completeness

Li Qiao-Yan and Cheng Tao

Abstract The aim of this chapter is the partial axiomatization for first-order predicate calculus formal system based on Schweizer–Sklar t-norm. By introducing the universal quantifier and existential quantifier, a predicate calculus formal deductive system $\forall UL^*$ based on Schweizer–Sklar t-norm according to propositional calculus formal deductive system UL^* based on Schweizer–Sklar t-norm is built up, moreover, the completeness of system $\forall UL^*$ are proved. So it shows that the semantic and syntactic of system $\forall UL^*$ are harmony.

keywords Schweizer–Sklar t-norm · Predicate calculus formal system · Approximate reasoning

22.1 Introduction

In recent years, the basic research of fuzzy logic and fuzzy reasoning is growing actively day by day, especially for the logic based on t-norm and its residua (see [1–11]). Some well-known logic systems have been built up, such as the basic logic (BL) [1, 3] introduced by Hajek; the monoidal t-norm based logic (MTL) [2] introduced by Esteva and Godo; a formal deductive system L^* introduced by Wang (see [7–11]); Universal logic was proposed by Huacan He [12–17], and so on.

One of the most important and central problems of fuzzy logic is the proper definition of logical operations. Because the t-norm reflects the property of the

L. Qiao-Yan (✉) · C. Tao (✉)
School of Science, Xi'an Polytechnic University, Xi'an 710048, China
e-mail: liqiaoyan@xpu.edu.cn
e-mail:

logic connective “AND” to a large extent, so t-norm and its according to R-implication are selected to describe the logical “AND” and “IMPLICATION” respectively. However, in order to describe the flexibility in human’s logical reasoning, choosing some parametric t-norms may be more reasonable. In fact, Klement and Navara [18] had studied the fuzzy logic system based on the parametric t-norm, Frank t-norm. Wu [19] and Wang [20] studied the parametric Kleene system. Moreover, the residual implications induced by the Schweizer–Sklar t-norm were studied by Whalen [21], where the parameter p is connected with the strength of the interaction among fuzzy rules. In [22], Zhang proposed the propositional calculus formal deductive system UL^* based on Schweizer–Sklar t-norm. In [23], the first-order predicate calculus formal system $\forall UL^*$ based on Schweizer–Sklar t-norm is given and the soundness of system are given. In this chapter, based on the work in [22, 23], we give the completeness of system $\forall UL^*$.

The chapter is organized as follows. After this introduction, Sect. 22.2 we will introduce the predicate calculus formal deductive system $\forall UL^*$ based on Schweizer–Sklar t-norm. In Sect. 22.3 the completeness of system $\forall UL^*$ will be proved. The final section offers the conclusion.

22.2 Predicate Formal System $\forall UL^*$

In order to build first-order predicate formal deductive system based on Schweizer–Sklar t-norm, we give the first-order predicate language as following:

First-order language J consists of symbols set and generation rules:

The symbols set of J consist of as following:

- (1) Object variables: $x, y, z, x_1, y_1, z_1, x_2, y_2, z_2, \dots$;
- (2) Object constants: a, b, c, a_1, b_1, c_1 , Truth constants: $\bar{0}, \bar{1}$;
- (3) Predicate symbols: $P, Q, R, P_1, Q_1, R_1, \dots$;
- (4) Connectives: $\&, \rightarrow, \Delta, -$;
- (5) Quantifiers: \forall (universal quantifier), \exists (existential quantifier);
- (6) Auxiliary symbols: $(,), \cdot$.

The symbols in (1)–(3) are called non-logical symbols of language J . The object variables and object constants of J are called terms. The set of all object constants is denoted by $\text{Var}(J)$. The set of all object variables is denoted by $\text{Const}(J)$. The set of all terms is denoted by $\text{Term}(J)$. If P is n -ary predicate symbol, t_1, t_2, \dots, t_n are terms, then $P(t_1, t_2, \dots, t_n)$ is called atomic formula.

The formula set of J is generated by the following three rules in finite times:

- (i) If P is atomic formula, then $P \in J$;
- (ii) If $P, Q \in J$, then $P \& Q, P \rightarrow Q, \Delta P \in J, -P \in J$;
- (iii) If $P \in J$, and $x \in \text{Var}(J)$, then $(\forall x)P, (\exists x)P \in J$.

The formulas of J can be denoted by $\varphi, \phi, \psi, \varphi_1, \phi_1, \psi_1, \dots$. Further connectives are defined as following:

$$\begin{aligned} \varphi \wedge \psi \text{ is } \varphi \&\psi (\varphi \rightarrow \psi), \quad \varphi \vee \psi \text{ is } ((\varphi \rightarrow \psi) \rightarrow \psi) \wedge (\psi \rightarrow \varphi) \rightarrow \varphi, \\ \neg \varphi \text{ is } \varphi \rightarrow \bar{0}, \quad \varphi \equiv \psi \text{ is } (\varphi \rightarrow \psi) \&\psi (\psi \rightarrow \varphi). \end{aligned}$$

Definition 1 The axioms and deduction rules of predicate formal system $\forall\text{UL}^*$ as following:

(i)The following formulas are axioms of $\forall\text{UL}^*$:

- (UL1) $(\varphi \rightarrow \psi) \rightarrow ((\psi \rightarrow \chi) \rightarrow (\varphi \rightarrow \chi))$
- (UL2) $(\varphi \&\psi) \rightarrow \varphi$
- (UL3) $(\varphi \&\psi) \rightarrow (\psi \&\varphi)$
- (UL4) $(\varphi \wedge \psi) \rightarrow \varphi$
- (UL5) $(\varphi \wedge \psi) \rightarrow (\psi \rightarrow \varphi)$
- (UL6) $((\varphi \&\psi) \rightarrow \chi) \rightarrow (\varphi \rightarrow (\psi \rightarrow \chi))$
- (UL7) $((\varphi \rightarrow \psi) \rightarrow \chi) \rightarrow (((\psi \rightarrow \varphi) \rightarrow \chi) \rightarrow \chi)$
- (UL8) $(\varphi \rightarrow (\psi \rightarrow \chi)) \rightarrow ((\varphi \&\psi) \rightarrow \chi)$
- (UL9) $((\varphi \rightarrow \psi) \rightarrow \chi) \rightarrow (((\psi \rightarrow \varphi) \rightarrow \chi) \rightarrow \chi)$
- (UL10) $\bar{0} \rightarrow \varphi$
- (UL11) $(-\ - \phi) \equiv \phi$
- (UL12) $\Delta\varphi \vee \neg\Delta\varphi$
- (UL13) $\Delta(\varphi \vee \psi) \rightarrow (\Delta\varphi \vee \Delta\psi)$
- (UL14) $\Delta\varphi \rightarrow \varphi$
- (UL15) $\Delta\varphi \rightarrow \Delta\Delta\varphi$
- (UL16) $\Delta(\varphi \rightarrow \psi) \rightarrow (\Delta\varphi \rightarrow \Delta\psi)$
- (UL17) $\Delta(\phi \rightarrow \psi) \rightarrow \Delta(\neg\psi \rightarrow \neg\phi)$
- (UL18) $(\forall x)\varphi(x) \rightarrow \varphi(t)$ (t substitutable for x in $\varphi(x)$)
- (UL19) $\varphi(t) \rightarrow (\exists x)\varphi(x)$ (t substitutable for x in $\varphi(x)$)
- (UL20) $(\forall x)(\chi \rightarrow \varphi) \rightarrow (\chi \rightarrow (\forall x)\varphi)$ (x is not free in χ)
- (UL21) $(\forall x)(\varphi \rightarrow \chi) \rightarrow ((\exists x)\varphi \rightarrow \chi)$ (x is not free in χ)
- (UL22) $(\forall x)(\varphi \vee \chi) \rightarrow ((\forall x)\varphi \vee \chi)$ (x is not free in χ)

Deduction rules of $\forall\text{UL}^*$ are three rules. They are:

Modus Ponens (MP): from $\varphi, \varphi \rightarrow \psi$ infer ψ ;

Necessitation: from φ infer $\Delta\varphi$;

Generalization: from φ infer $(\forall x)\varphi$.

The meaning of “ t substitutable for x in $\varphi(x)$ ” and “ x is not free in χ ” in the above definition have the same meaning in the classical first-order predicate logic, moreover, we can define the concepts such as proof, theorem, theory, deduction from a theory T , T -consequence in the system $\forall\text{UL}^*$. $T \vdash \varphi$ denotes that φ is provable in the theory T . $\vdash \varphi$ denotes that φ is a theorem of system $\forall\text{UL}^*$. Let $\text{Thm}(\forall\text{UL}^*) = \{\varphi \in J \mid \vdash \varphi\}$, $\text{Ded}(T) = \{\varphi \in J \mid T \vdash \varphi\}$. Being the axioms of propositional system UL^* are in predicate system $\forall\text{UL}^*$, then the theorems in UL^* are theorems in $\forall\text{UL}^*$. According the similar proof in [1, 15, 16, 22] we can get the following lemmas.

Lemma 1 *The hypothetical syllogism holds in $\forall\text{UL}^*$, i.e. let $\Gamma = \{\varphi \rightarrow \psi, \psi \rightarrow \chi\}$, then $\Gamma \vdash \varphi \rightarrow \chi$.*

Lemma 2 $\forall\text{UL}^*$ proves:

- (1) $\varphi \rightarrow \varphi$; (2) $\varphi \rightarrow (\psi \rightarrow \varphi)$; (3) $(\varphi \rightarrow \psi) \rightarrow ((\varphi \rightarrow \gamma) \rightarrow (\psi \rightarrow \gamma))$;
 (4) $(\varphi \& (\varphi \rightarrow \psi)) \rightarrow \psi$; (5) $\Delta\varphi \equiv \Delta\varphi \& \Delta\varphi$.

Lemma 3 $\forall\text{UL}^*$ proves: *If $T = \{\varphi \rightarrow \psi, \chi \rightarrow \gamma\}$, then $T \vdash (\varphi \& \chi) \rightarrow (\psi \& \gamma)$.*

In order to prove the soundness of predicate system $\forall\text{UL}^*$, we should introduce the following definitions.

Definition 2 [22] Let L be a $(-, \Delta, *, \Rightarrow, \wedge)$ - type algebra. L is called a UL^* -algebra, if the following conditions hold:

1. $(L, *, \Rightarrow, \wedge, \vee, 0, 1)$ is a residuated lattice with an order \leq , where 1 is the greatest element and 0 is the least element, \vee is defined by $a \vee b = ((a \Rightarrow b) \Rightarrow b) \wedge ((b \Rightarrow a) \Rightarrow a)$, $\forall a, b, c \in L$. And \vee satisfied $(a \Rightarrow b) \vee (b \Rightarrow a) = 1$, $\forall a, b, c \in L$.
2. $-$ is an order-reversing involution on L with respect to \leq , that is if $a \leq b$ then $(-b) \rightarrow (-a)$, and $(--a) = a$ for any $a \in L$.
3. Δ is a unary operation on L , and it satisfies $(\forall a, b, c \in L)$:

$$\begin{aligned} \Delta a \cup (\Delta a \Rightarrow 0) &= 1, \\ \Delta(a \cup b) &\leq (\Delta a) \cup (\Delta b), \\ \Delta a &\leq a, \\ \Delta a &\leq \Delta \Delta a, \\ (\Delta a \Rightarrow b) &\leq (\Delta a) \Rightarrow (\Delta b), \\ (\Delta a \Rightarrow b) &= \Delta((-b) \Rightarrow (-a)), \\ \Delta 1 &= 1 \end{aligned}$$

Let J is first-order predicate language, L is linearly ordered UL^* -algebra, $M = (M, (r_p)_p, (m_c)_c)$ is called a L -evaluation for first-order predicate language J , which M is non-empty domain, according to each n -ary predicate P and object constant c , r_p is L -fuzzy n -ary relation: $r_p : M^n \rightarrow L$, m_c is an element of M .

Definition 3 Let J be predicate language, M is L -evaluation of J , x is object variable, $P \in J$.

- (i) A mapping $V: \text{Term}(J) \rightarrow M$ is called M -evaluation, if for each $c \in \text{Const}(J)$, $v(c) = m_c$;
- (ii) Two M -evaluation v, v' are called equal denoted by $v \equiv_x v'$ if for each $y \in \text{Var}(J \setminus \{x\})$, there is $v(y) = v'(y)$.
- (iii) The value of a term given by M, v is defined by: $\|x\|_{M,v} = v(x)$; $\|c\|_{M,v} = m_c$. We define the truth value $\|\varphi\|_{M,v}^L$ of a formula φ as following. Clearly, $*, \Rightarrow, \Delta$ denote the operations of L .

$$\begin{aligned}
\|P(t_1, t_2, \dots, t_n)\|_{M,v}^L &= r_P(\|t_1\|_{M,v}, \dots, \|t_n\|_{M,v}) \\
\|\varphi \rightarrow \psi\|_{M,v}^L &= \|\varphi\|_{M,v}^L \Rightarrow \|\psi\|_{M,v}^L \\
\|\varphi \&\psi\|_{M,v}^L &= \|\varphi\|_{M,v}^L * \|\psi\|_{M,v}^L \\
\|\bar{0}\|_{M,v}^L &= 0; \quad \|\bar{1}\|_{M,v}^L = 1 \\
\|\Delta\varphi\|_{M,v}^L &= \Delta\|\varphi\|_{M,v}^L \\
\|-\varphi\|_{M,v}^L &= -\|\varphi\|_{M,v}^L \\
\|(\forall x)\varphi\|_{M,v}^L &= \inf\{\|\varphi\|_{M,v'}^L \mid v \equiv_x v'\} \\
\|(\exists x)\varphi\|_{M,v}^L &= \sup\{\|\varphi\|_{M,v'}^L \mid v \equiv_x v'\}
\end{aligned}$$

In order to the above definitions are reasonable, the infimum/supremum should exist in the sense of L. So the structure M is L-safe if all the needed infima and suprema exist, i.e. $\|\varphi\|_{M,v}^L$ is defined for all φ, v .

Definition 4 Let $\varphi \in J$, M be a safe L-structure for J.

- (i) The truth value of φ in M is $\|\varphi\|_M^L = \inf\{\|\varphi\|_{M,v}^L \mid v \text{ M-evaluation}\}$.
- (ii) A formula φ of a language J is an L-tautology if $\|\varphi\|_M^L = 1_L$ for each safe L-structure M. i.e. $\|\varphi\|_{M,v}^L = 1$ for each safe L-structure M and each M-valuation of object variables.

Definition 5 A logic system is soundness if for its each theorem φ , we can get φ is a tautology.

Theorem 1 (Soundness) [23] *Let L is linearly ordered UL*-algebra and φ is a formula in J, if $\vdash \varphi$, then φ is L-tautology, i.e. $\|\varphi\|_M^L = 1_L$.*

Similarly, we can get the following strong soundness theorem.

Definition 6 Let T be a theory, L be a linearly ordered UL* - algebra and M a safe L-structure for the language of T. M is an L-model of T if all axioms of T are 1_L -true in M, i.e. $\|\varphi\| = 1_L$ in each $\varphi \in T$.

Theorem 2 (Strong Soundness) [23] *Let T be a theory, L is linearly ordered UL*-algebra and φ is a formula in J, if $T \vdash \varphi$ (φ is provable in T), then $\|\varphi\|_M^L = 1_L$ for each linearly ordered UL*-algebra L and each L-model M of T.*

Theorem 3 (Deduction Theorem) [23] *Let T be a theory, φ, ψ are closed formulas. Then $(T \cup \{\varphi\}) \vdash \psi$ iff $T \vdash \Delta\varphi \rightarrow \psi$.*

22.3 Completeness of $\forall UL^*$

Definition 7 Let T be a theory on $\forall UL^*$.

- (1) T is consistent if there is a formula φ unprovable in T .
- (2) T is complete if for each pair φ, ψ of closed formula, $T \vdash (\varphi \rightarrow \psi)$ or $T \vdash (\psi \rightarrow \varphi)$.
- (3) T is Henkin if for each closed formula of the form $(\forall x)\varphi(x)$ unprovable in T there is a constant c in the language of T such that $\varphi(c)$ is unprovable.

Lemma 4 T is inconsistent iff $T \vdash \bar{0}$.

Lemma 5 T is complete iff for each pair φ, ψ of closed formulas if $T \vdash \varphi \vee \psi$, then T proves φ or T proves ψ .

Proof Sufficiency: For each pair φ, ψ of closed formulas, being $(\varphi \rightarrow \psi) \vee (\psi \rightarrow \varphi)$ is theorem in $\forall\text{UL}^*$, so $T \vdash (\varphi \rightarrow \psi) \vee (\psi \rightarrow \varphi)$, thus $T \vdash (\varphi \rightarrow \psi)$ or $T \vdash (\psi \rightarrow \varphi)$. Thus T is complete.

Necessity: assume T is complete and $T \vdash \varphi \vee \psi$, Either $T \vdash \varphi \rightarrow \psi$ and then $T \vdash (\varphi \vee \psi) \rightarrow \psi$, thus $T \vdash \psi$, or $T \vdash \psi \rightarrow \varphi$ and then similarly $T \vdash \varphi$.

Definition 8 Let T be a theory, the set of all closed formulas over $\forall\text{UL}^*$ is denoted by $F^c(\forall\text{UL}^*)$. The definition of relation \sim_T on $F^c(\forall\text{UL}^*)$ is: $\varphi \sim_T \psi$ iff $T \vdash \varphi \rightarrow \psi$, $T \vdash \psi \rightarrow \varphi$.

Obviously, \sim_T is equivalent relation on $F^c(\forall\text{UL}^*)$, and holds on $\&, \rightarrow, \Delta, -$. So the quotient algebra $[F]_T = F^c(\forall\text{UL}^*) / \sim_T = \{[\varphi]_T \mid \varphi \in F^c(\forall\text{UL}^*)\}$ of $F^c(\forall\text{UL}^*)$ about \sim_T is UL^* -algebra, in which, $[\varphi]_T = \{\psi \in F^c(\forall\text{UL}^*) \mid \psi \sim_T \varphi\}$, the partial order \leq on $[F]_T$ is $[\varphi]_T \leq [\psi]_T$ iff $T \vdash \varphi \rightarrow \psi$.

Lemma 6 (1) If T is complete then $[F]_T$ is linearly ordered.

(2) If T is Henkin then for each formula $\varphi(x)$ with just one free variable x , $[(\forall x)\varphi]_T = \inf_c [\varphi(c)]_T$, $[(\exists x)\varphi]_T = \sup_c [\varphi(c)]_T$, in which c running over all constants of T .

Proof (1) is obvious since $[\varphi]_T \leq [\psi]_T$ iff $T \vdash \varphi \rightarrow \psi$.

(2) Clearly, $[(\forall x)\varphi(x)]_T \leq \inf_c [\varphi(c)]_T$ for each c , thus $[(\forall x)\varphi(x)]_T \leq \inf_c [\varphi(c)]_T$. To prove that $[(\forall x)\varphi(x)]_T$ is the infimum of all $[\varphi(c)]_T$, assume $[\gamma]_T \leq [\varphi(c)]_T$ for each c , we have to prove $[\gamma]_T \leq [(\forall x)\varphi(x)]_T$ (which means that $[(\forall x)\varphi(x)]_T$ is the greatest lower bound of all $[\varphi(c)]_T$). But if $[\gamma]_T \not\leq [(\forall x)\varphi(x)]_T$ then $T \not\vdash \gamma \rightarrow (\forall x)\varphi(x)$, thus $T \not\vdash (\forall x)(\gamma \rightarrow \varphi(x))$. So by the henkin property, there is a constant c such that $T \not\vdash \gamma \rightarrow \varphi(c)$, thus $[\gamma]_T \not\leq [\varphi(c)]_T$, a contradiction.

Similarly, $[\varphi(c)]_T \leq [(\exists x)\varphi(x)]_T$ for each c . Assume $[\varphi(c)]_T \leq [\gamma]_T$ for each c , we prove $[(\exists x)\varphi(x)]_T \leq [\gamma]_T$. Indeed, if $[(\exists x)\varphi(x)]_T \not\leq [\gamma]_T$ then $T \not\vdash (\exists x)\varphi(x) \rightarrow \gamma$, thus $T \not\vdash (\forall x)(\varphi(x) \rightarrow \gamma)$ and for some c , $T \not\vdash \varphi(c) \rightarrow \gamma$, thus $[\varphi(c)]_T \not\leq [\gamma]_T$, a contradiction. This completes the proof.

Lemma 7 For each theory T and each closed formula α , if $T \not\vdash \alpha$ then there is a complete Henkin supertheory \hat{T} of T such that $\hat{T} \not\vdash \alpha$.

Proof First observe that if T' is an extension of T , $T' \not\vdash \alpha$, and (φ, ψ) is a pair of closed formulas then either $(T' \cup \{\varphi \rightarrow \psi\}) \not\vdash \alpha$ or $(T' \cup \{\psi \rightarrow \varphi\}) \vdash \alpha$. This is proved easily using the deduction theorem (Theorem 3). Indeed, if $T', \{\varphi \rightarrow \psi\} \vdash \alpha$ and $T', \{\psi \rightarrow \varphi\} \vdash \alpha$, then $T' \vdash \Delta(\varphi \rightarrow \psi) \rightarrow \alpha$, $T' \vdash \Delta(\psi \rightarrow \varphi) \rightarrow \alpha$, so $T' \vdash \Delta(\varphi \rightarrow \psi) \vee \Delta(\psi \rightarrow \varphi) \rightarrow \alpha$, thus $T' \vdash \alpha$, a contradiction.

Put $T'' = T' \cup \{\varphi \rightarrow \psi\}$ in the former case and $T'' = T' \cup \{\psi \rightarrow \varphi\}$ in the latter, T'' is the extension of T' deciding (φ, ψ) and keeping α unprovable.

We shall construct \widehat{T} in countably many stages. First extend the language J of T to J' adding new constants c_0, c_1, c_2, \dots . In the construction we have to decide each pair (φ, ψ) of closed J' -formulas and ensure the Henkin property for each closed J' -formula of the form $(\forall x)\chi(x)$. These are countably many tasks and may be enumerated by natural numbers (e.g. in even steps we shall decide all pair (φ, ψ) , in odd ones process all formulas $(\forall x)\chi(x)$ —or take any other enumeration).

Put $T_0 = T$ and $\alpha_0 = \alpha$, then $T_0 \not\vdash \alpha_0$. Assume T_n, α_n have been constructed such that T_n extends T_0 , $T_n \vdash \alpha \rightarrow \alpha_n$, $T_n \not\vdash \alpha_n$; we construct T_{n+1}, α_{n+1} in such a way that $T_n \vdash \alpha \rightarrow \alpha_{n+1}$, $T_{n+1} \not\vdash \alpha_{n+1}$ and T_{n+1} fulfils the n th task.

Case 1 n th task is deciding (φ, ψ) . Let T_{n+1} be extension of T_n deciding (φ, ψ) and keeping α_n unprovable; put $\alpha_{n+1} = \alpha_n$.

Case 2 n th task is processing $(\forall x)\chi(x)$. First let c be one of the new constant not occurring in T_n .

Subcase(a) $T_n \not\vdash \alpha_n \vee \chi(c)$, thus $T_n \not\vdash (\forall x)\chi(x)$. Put $T_{n+1} = T_n$, $\alpha_{n+1} = \alpha_n \vee \chi(c)$.

Subcase(b) $T_n \vdash \alpha_n \vee \chi(c)$, thus $T_n \vdash \alpha_n \vee \chi(x)$ by the standard argument (in the proof of $\alpha_n \vee \chi(c)$ replace c by a new variable x throughout). Hence $T_n \vdash (\forall x)(\alpha_n \vee \chi(x))$ and using axiom (UL20) for the first time, $T_n \vdash \alpha_n \vee (\forall x)\chi(x)$. Thus $T_n \cup \{(\forall x)\chi(x) \rightarrow \alpha_n\} \vdash \alpha_n$ so that $T_n \cup \{\alpha_n \rightarrow (\forall x)\chi(x)\} \not\vdash \alpha_n$, $T_n \cup \{\alpha_n \rightarrow (\forall x)\chi(x)\} \vdash (\forall x)\chi(x)$ does not prove α_n but it does prove $(\forall x)\chi(x)$. Thus put $T_{n+1} = T_n \cup \{\alpha_n \rightarrow (\forall x)\chi(x)\}$ and $\alpha_{n+1} = \alpha_n$.

Now let \widehat{T} be the union of all T_n . Then clearly \widehat{T} is complete and $\widehat{T} \vdash \alpha$ (since for all n , $\widehat{T} \vdash \alpha$). We show that \widehat{T} is Henkin. Let $\widehat{T} \not\vdash (\forall x)\chi(x)$ and let $(\forall x)\chi(x)$ be processed in step n . Then $T_{n+1} \not\vdash (\forall x)\chi(x)$, $T_{n+1} \not\vdash \alpha_{n+1}$, thus subcase (a) applies and $\widehat{T} \not\vdash \alpha_{n+1}$, α_{n+1} being $\alpha_n \vee \chi(c)$. Hence $\widehat{T} \not\vdash \chi(c)$. This completes the proof.

Lemma 8 *For each complete Henkin theory T and each closed formula α unprovable in T there is a linearly ordered UL*-algebra L and L -model M of T such that $\|\alpha\|_M^L < 1_L$.*

Proof Take M be the set of all constants of the language of T , $m_c = c$ for each such constant. Let L be the lattice of classes of T -equivalent closed formulas, i.e. put $[\varphi]_T = \{\psi \mid T \vdash \varphi \equiv \psi\}$, $[\varphi]_T * [\psi]_T = [\varphi \& \psi]_T$, $[\varphi]_T \Rightarrow [\psi]_T = [\varphi \rightarrow \psi]_T$. So L is a linearly ordered UL*-algebra (since $T \vdash \varphi \rightarrow \psi$ or $T \vdash \psi \rightarrow \varphi$ for each pair (φ, ψ)).

For each predicate P of arity n , let $r_P(c_1, \dots, c_n) = [P(c_1, \dots, c_n)]_T$, this completes the definition of M . It remains to prove $\|\alpha\|_M^L = [\alpha]_T$ for each closed formula φ . Then for each axiom φ of T we have $\|\varphi\|_M^L = [\varphi]_T = [1]_T = 1_L$, but $\|\alpha\|_M^L = [\alpha]_T \neq [1]_T = 1_L$. For atomic closed formula φ the claim follows by definition; the induction step for connectives is obvious. We handle the quantifiers.

Let

$(\forall x)\varphi(x), (\exists x)\varphi(x)$ be closed, then by the induction hypothesis,

$$\|(\forall x)\varphi(x)\|_M^L = \inf_c \|\varphi(c)\|_M^L = \inf_c [\varphi(c)]_T = [(\forall x)\varphi(x)]_T$$

$$\|(\exists x)\varphi(x)\|_M^L = \sup_c \|\varphi(c)\|_M^L = \sup_c [\varphi(c)]_T = [(\exists x)\varphi(x)]_T$$

Here we use lemma and the fact that in our M , each element c of M is the meaning of a constant (namely itself); this gives $\|(\forall x)\varphi(x)\|_M^L = \inf_c \|\varphi(c)\|_M^L$ and the dual for \exists .

Using the above lemmas, we can get the following completeness theorem.

Theorem 4 (Completeness) *For predicate calculus system $\forall UL^*$, T is a theory, φ is a formula, $T \vdash \varphi$ iff for each linearly ordered UL^* -algebra L and each safe L -model M of T , $\|\varphi\|_M^L = 1_L$.*

22.4 Conclusion

In this chapter a predicate calculus formal deductive system $\forall UL^*$ according to the propositional system UL^* for Schweizer–Sklar t-norm is built up. We prove the system $\forall UL^*$ is complete, which shows that the semantic and syntactic of system $\forall UL^*$ are harmony.

Acknowledgments This work is supported by Scientific Research Program Funded by Shaanxi Provincial Education Department (Program No. 2010JK567).

References

1. Hajek P (1998) Metamathematics of fuzzy logic. Kluwer Academic Publishers, Dordrecht
2. Esteva F, Godo L (2001) Monoidal t-normbased logic: towards a logic for left-continuous t-norms. Fuzzy Sets Syst 124:271–288
3. Cignoli R, Esteva F, Godo L, Torrens A (2000) Basic fuzzy logic is the logic of continuous t-norms and their residua. Soft comput 4:106–112
4. Hohle U (1995) Commutative, residuated l-monoids. In: Non-Classical Logics, Their Application to Fuzzy Subsets. Hohle U, Klement EP (eds) Kluwer Academic Publishers. Dordrecht, London, pp 53–106
5. Esteva F, Godo L et al (2000) Residuated fuzzy logics with an involutive negation. Arch Math Log 39:103–124
6. Klement EP, Mesiar R, Pap E (2000) Triangular Norms. Kluwer Academic Publishers, Dordrecht

7. Pei DW, Wang GJ (2002) The completeness and applications of the formal system L^* . *Sci China (Ser F)* 45:40–50
8. Wang SM, Wang BS, Pei DW (2005) A fuzzy logic for an ordinal sum t-norm. *Fuzzy Sets Syst* 149:297–307
9. Wang GJ (2000) *Non-classical mathematical logic and approximate reasoning*. Science Press, Beijing (in Chinese)
10. Pei DW (2002) First-order Formal System K^* and its completeness. *Chinese Annals Math Ser A* 23(6):675–684
11. Wu HB (2009) Competeness of BL_{Δ}^* system. *J Jishou Univ (Nat Sci Ed)* 30(6):1–5
12. He HC et al (2001) *Universal Logic Principle*. Science Press(in Chinese), Beijing
13. Ma YC, He HC (2005) The fuzzy reasoning rules based on universal logic. In: 2005 IEEE international conference on GrC, IEEE Press, pp 561–564
14. Ma YC, He HC (2005) A propositional calculus formal deductive system $\mathcal{UL}_{h \in (0,1]}$ of universal logic. In: Proceedings of 2005 ICMLC, IEEE Press, pp 2716–2721
15. Ma YC, Li QY (2006) A propositional deductive system of universal logic with projection operator. In: Proceedings of 2006 ISDA, IEEE Press, pp 993–998
16. Ma YC, He HC (2006) The axiomatization for 0-level universal logic. *Lect Notes Artif Intell*, 3930:367–376
17. Ma YC, He HC (2008) Axiomatization for 1-level universal AND operator. *J China Univ Posts Telecommun*, 15(2):125–129
18. Klement EP, Navara M (1999) Propositional fuzzy logics based on Frank t-norms: a comparison in fuzzy Sets, logics and reasoning about knowledge Dubois D et al. (eds.), Kluwer Academic Publishers
19. Wu WM (2000) Generalized tautologies in parametric Kleene’s systems. *Fuzzy Syst Math*, 14(1):1–7
20. Wang GJ, Lan R (2003) Generalized tautologies of the systems H_x , *J Shaanxi Normal Univ (Nat Sci Ed)*, 131(2):1–11
21. Whalen T (2003) Parameterized R-implications. *Fuzzy Sets Syst* 134(2):231–281
22. Zhang XH, He HC, Xu Y (2006) A fuzzy logic system based on Schweizer–Sklar t-norm. *Sci China (Ser F: Inf Sci)* 49(2):175–188
23. Li QY, Ma YC, The predicate system based on Schweizer–Sklar t-norm and its soundness (submitted)

Chapter 23

Providing Timely Response in Smart Car

Jie Sun, Yongping Zhang, Bobo Wang and Kejia He

Abstract Driving is a complex process requiring timely interactions between the driver, the vehicle and the environment. This paper aims at providing time-stressed intelligent services in the smart car. To guarantee the system's timeliness, we identify the situations which require timely responses and determine the worst-case, the deadlines and durations of these tasks. The tasks are preemptive so that a task with higher priority will interrupt the current running task with lower priority. We implement a real-time scheduler to provide a satisfactory result in fewer amount of time.

Keywords Smart car · Real time · Slake stealing · Pervasive computing

23.1 Introduction

The smart car is becoming a promising application domain of ubiquitous computing, which aims at assisting the driver with easier driving, less workload and less chance of getting injured [1]. For this purpose, a smart car must not only

J. Sun (✉) · Y. Zhang · B. Wang · K. He
Ningbo University of Technology, Ningbo, China
e-mail: sunjie@nbut.cn

Y. Zhang
e-mail: ypz@nbut.cn

B. Wang
e-mail: swtbb@yahoo.cn

K. He
e-mail: start_learn@163.com

provide the correct assistance but provide it within strict time constraints. The smart car requires timely responses to the events in the driving environment and provides real-time intelligent services that will be reliable enough for critical applications.

Lots of smart cars have been developed in the past decade. However, little effort is done in quantified timely performance of smart cars. This is because of the nature of intelligent techniques, most of which need the prior knowledge of what will happen. Thus it is impossible to pre-calculate all possible combinations of situations which may occur. Therefore, it is difficult to provide a correct answer within a certain time constraint. This paper focuses on providing time-stressed intelligent services in the smart car. To guarantee the system's timeliness, we identify the situations which require timely responses and determine the worst-case, the deadlines and durations of these tasks. A real-time scheduler will use approaches that allow programs to meet deadlines.

The remainder of the paper is organized as follows. [Section 23.2](#) introduces the related work of real-time technologies and smart car. [Section 23.3](#) proposes the architecture of the smart car environment, with the support of time constraints. We introduce the real-time scheduler of the smart car to illustrate how our approach works in [Sect. 23.4](#). The performance evaluation is shown in [Sect. 23.5](#). The conclusions are given in [Sect. 23.6](#).

23.2 Related Work

Cars are becoming smarter and smarter. Automotive manufactures implement many novel ideas in their newest series of concept cars. Manufactures like BMW, Mercedes-Benz, Volvo, Lexus have developed different driver assistant systems to provides advanced active safety guarantees that monitors the roadway, the states of the vehicle and the driver. Time guarantee is critical especially to such safety-critical smart car.

Real-time systems are used to control physical processes that range in complexity from automobile ignition systems to controllers for flight systems and nuclear power plants. Such systems are referred to as hard real-time systems because the timing requirements of system computations must always be met or the system will fail. The scheduler for these systems must coordinate resources to meet the timing constraints of the physical system. This implies that the scheduler must be able to predict the execution behavior of all tasks within the system.

Real-time scheduling can be classified in two categories, static and dynamic scheduling. A static real-time scheduling algorithm such as Rate Monotonic schedules all real-time tasks off-line using static parameters and requires complete knowledge about tasks and system parameters, while dynamic task scheduler calculates the feasible schedule on-line and allows tasks to be invoked dynamically. These algorithms use dynamic parameters such as deadline and laxity [2].

The common scheduling solution of real-time systems is to use cyclical executives [3]. Cyclical executives provide a deterministic schedule for all tasks and resources in a real-time system by creating a static timeline upon which tasks and resources are assigned specific time intervals. This technique has been able to meet the timing requirements of many real-time systems. However, the cyclical executive approach has several serious drawbacks. Although a cyclical executive provides predictability through the use of a deterministic schedule, it is the requirement of determinism that is source of these drawbacks. A real-time system typically has many different timing constraints that must be met and no formal algorithm exists for creating timelines to meet these constraints. Therefore, each real-time system requires a different, handcrafted cyclical executive that requires extensive testing to verify its correctness. Also, because of the many ad hoc decisions made to create the timelines, changes to the system have unpredictable effects upon the correctness of the timelines and thus, a complete and new set of testing is required. Thus, the cyclical executive approach is typically expensive to create, verify, and update. As real-time systems become more complex, the problems associated with cyclical executives only become more difficult. A better scheduling solution is needed that provides predictability without the requirement of determinism.

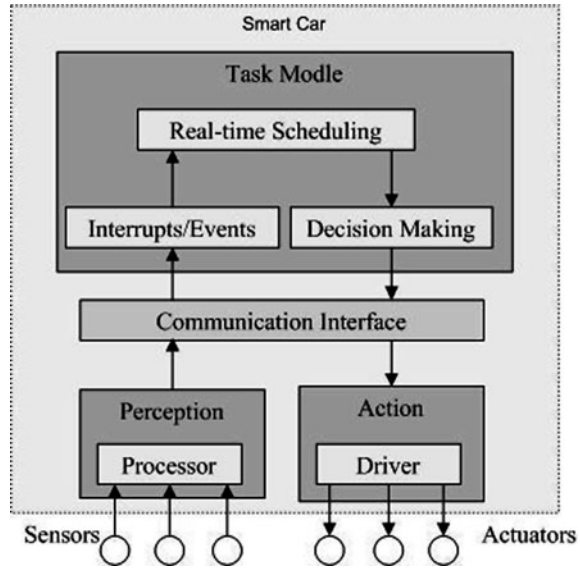
An alternative solution to the real-time scheduling problem is to use priority-driven, preemptive scheduling algorithms to schedule tasks and resources [4]. These algorithms assign priorities to each task in the system and always select the highest priority task to execute, preempting lower priority tasks. Such algorithms can describe the worst-case timing behavior of the system under all circumstances, and thus provide predictability by capturing the timing behavior of the system. Using well defined algorithms to schedule tasks and resources in a real-time system yields an understandable scheduling solution that is void of the ad hoc decisions typically found in cyclical executives. Thus, the resulting real-time system is easier to maintain and upgrade. Also, since the algorithmic approach accurately models the timing behavior of the system, the resulting schedule is easier to test. Because of these benefits, we will be investigating algorithmic based scheduling methods in this thesis as a means of creating efficient, predictable real-time systems.

Anytime algorithms [5] are also be used for satisfactory problem solving. They have such characteristics as: they can be preemptively scheduled; they can be terminated at any time and produce an answer; and the answers returned improve in some well-behaved manner with respect to time. In contrast, approximate processing algorithms can be terminated at any time after a given deadline and will return an answer, and the answer returned improves in a step-wise manner as the initial deadline moves farther away.

23.3 The Architecture of the Smart Car

A smart car is a comprehensive integration of many different sensors, control modules, actuators, and so on. A smart car can monitor the driving environment by

Fig. 23.1 The general architecture of a smart car



sensors, assess the possible risks, and take appropriate actions to avoid or reduce the risk. In this section, we briefly review the system model and task model of the smart car.

23.3.1 System Model

A general architecture of a smart car is shown in Fig. 23.1.

In order to achieve the perception of the driving environment, the smart car must be aware of the states of the traffic, the car and the driver through a variety of sensors. Active environments sensing in- and out-car will be a general capability in near future. Lidar-, radar- or vision-based approaches can be used to provide the positioning information. Multiple cameras are able to eliminate blind spots, recognize obstacles, and record the surroundings. The dynamics of a car can be read from the engine, the throttle and the brake. These data will be transferred by controller area networks (CAN) in order to analyze whether the car functions normally. Physiological sensors can detect whether the driver is in good condition. The sensor data is processed on a separate I/O processor. Most data just confirms regular patterns, but occasionally abnormal or unexpected data arrives, which must be given attention to. Filtering of regular data is necessary to avoid both overburdening the perception component with needless detail and not informing it of important new information.

The smart car is built by collections of heterogeneous devices and Electronic Controlling Units (ECUs) embedded in the car. These devices are specialized to

export their services unobtrusively. The availability of so many services is implemented by diverse individual tasks. Some services are routine and can be implemented as periodic task. While filtered sensor data are sent into the system and act as external events. Event list triggers sporadic and aperiodic tasks.

The task model supports timing constraints directly. Each task is associated with a deadline (latest response time). A correctly scheduled task must start executing and finish before its deadline. The task model is discussed in detail in Sects. 3.2.

Decision making module determines the reaction of the smart car to the external events. Different levels of risk will lead to different responses, including notifying the driver through Human Machine Interface (HMI) and taking emergency actions by car actuators. HMI warns the driver of the potential risks in non-emergent situations. For an example, a tired driver would be awakened by an acoustic alarm or vibrating seat. Visual indications should be applied in a cautious way, since a complex graph or long text sentence will seriously impair the driver's attention and possibly cause harm. The actuators will execute specified control on the car without the driver's commands. The smart car will adopt active measures such as stopping the car in case that the driver is unable to act properly, or applying passive protection to reduce possible harm in abrupt accidents, for example, popping up airbags.

23.3.2 Task Model

The task model of a smart car consists of a set of tasks, and the tasks are distinguished by its timing behavior. A periodic task is released in a steady pace determined by its predefined period. Each task has a predefined relative deadline. A reactive task, on the other hand, is released by an external event in the driving environment. The default state of the system is ready to react to whatever it is defined to react on.

The periodic task set T is defined by $T = \{T_i^p(C_i, D_i, P_i) : i = 1, \dots, n_p\}$ with $1 \leq C_i \leq D_i \leq P_i$, where C_i is the worst-case execution time (WCET), D_i is the relative deadline and P_i is the period of task T_i with the priority p .

The reactive task set R can be defined as $R = \{R_i^p(C_i, D_i, L_i) : i = 1, \dots, n_r\}$ with $1 \leq C_i \leq D_i$ where C_i is the worst-case execution time, D_i is the relative deadline and L_i is the release time of reactive task R_i with the priority p .

23.4 The Scheduler

Only one task can run on a processor at the same time. The selection of the task is done by the scheduler, which is one of the main tasks of a real-time operating

system (RTOS). There is some features of task scheduling in the smart car that should be stood to: first, most of the tasks that need to be taken account are event-driven reactive tasks; second, the scheduling should be priority-driven and preemptive, so that the emergent situation can be handled as soon as possible; third, the smart car is a personal space and the driver is the center of the personal space, so the intent conflict of different users in traditional smart space rarely exists and thus the resource competition between different tasks can be avoided.

For the reasons analyzed above, we design the principle of the task scheduling as: (1) Periodic tasks are be scheduled to complete before their deadline; there may be some slack time between completion of the periodic tasks and its deadline. (2) Execute reactive tasks in the slack time, ahead of periodic tasks. The main goal is to schedule in such a way that it maximizes the number of accepted reactive tasks without interference with the deadlines of periodic tasks. The main advantage is the reduction of response time for reactive tasks.

The algorithm guarantees a given response time for a reactive task based on extending the slack stealing algorithm for EDF schedulers. It calculates the slack time for each priority level and the amount of slack is the longest time the periodic task at a priority level might be delayed without missing its deadline.

Let $S_i(t)$ denote the slack at priority level i at some arbitrary time t , $l_i(t)$ denote the time at which task T_i was last invoked, $x_i(t)$ denote the earliest possible next release of task T_i , $d_i(t)$ denote the next deadline on an invocation of task T_i , $c_i(t)$ denote the amount of time that the current invocation of task T_i was executed, $e_i(t)$ denote the time at which last completed invocation was finished. These data can be achieved from the task control block by operating system. The algorithm is shown as follows:

```

begin
   $S_i(t) := 0;$ 
   $w_i^{m+1}(t) := 0;$ 
  while  $w_i^{m+1}(t) \leq d_i(t)$  do;
     $w_i^m(t) := w_i^{m+1}(t);$ 
     $w_i^{m+1}(t) := S_i(t) + \sum_{\forall j \in hp(i) \cup i} (c_j(t) + \lceil \frac{(w_i^m(t) - x_j(t))_0}{T_j} \rceil C_j);$ 
    if  $w_i^m(t) := w_i^{m+1}(t)$  then
       $S_i(t) := S_i(t) + v_i(t, w_i^m(t));$ 
       $w_i^{m+1}(t) := w_i^{m+1}(t) + v_i(t, w_i^m(t)) + \varepsilon;$ 
       $S_i^{\max}(t) := S_i(t);$ 
end.

```

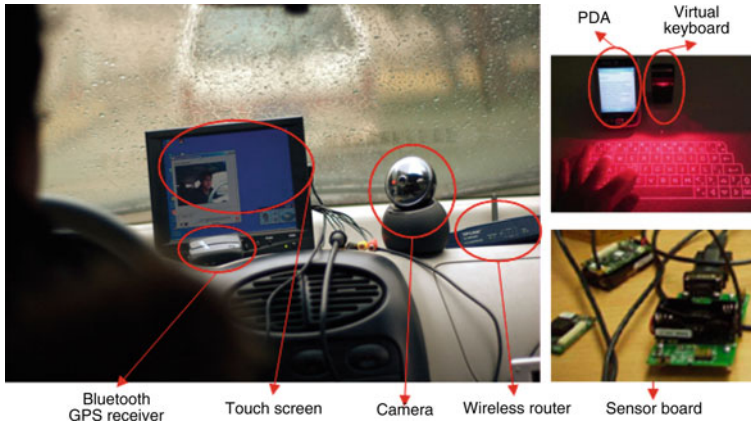


Fig. 23.2 The smart car prototype

23.5 Performance Evaluation

We build a smart car prototype to evaluate the performance. We integrate devices, sensors and network to build a smart car prototype, as shown in Fig. 23.2. The smart car includes intelligent devices, sensors, and ECUs. The devices used in the prototype are listed in Table 23.1, and the sensors are listed in Table 23.2. A notebook PC runs the applications such as driver face recognition. A PDA conducts the interaction between the driver and the smart car. The virtual keyboard and the touch screen are used to provide the passengers with easier interaction. Three ECUs are built using MPC555 processor. A wireless router and a CAN hub are adopted to provide different communication requirements. A RTOS SmartOSEK is developed [6].

23.5.1 Test Result

An ECU runs the main periodic task responsible for velocity monitoring. Besides, it detects the ambient light of the car. When the light is dim, the spotlight will be turned on. The experiment is shown in Fig. 23.3 and the task description is shown in Table 23.2.

Test 1: According to the algorithm, we can compute the slack:

$$\text{Iteration 0: } S_i(t) = 0, w_i^0 = 0, x_1(t) = 2, x_2(t) = 4, d_2(t) = 4, c_1(t) = 0, c_2(t) = 2, \varepsilon = 0;$$

$$\text{Iteration 1: } w_2^1(t) = S_2(t) + c_1(t) + c_2(t) = 0 + 0 + 2 = 2;$$

$$\text{Iteration 2: } w_2^2(t) = S_2(t) + c_1(t) + c_2(t) = 0 + 0 + 2 = 2.$$

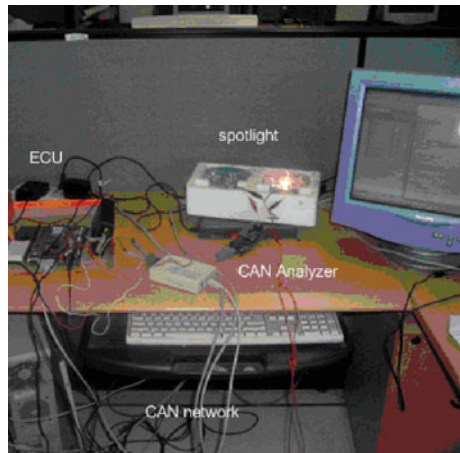
Table 23.1 Device list in the smart car prototype

Device	Type
Notebook PC	SONY VAIO FJ
ECU	Motorola MPC555, Atmel AT89C51
PDA	400 MHz Intel, XScal-PXA255
Wireless router	D-Link DIR-300
CAN analyzer	CANalsyt-II
CAN connection	CANHub-S5
Virtual keyboard	I-tech virtual laser keyboard
Touch screen	FA801-NP/C/T, 8"
Environment sensor	GAINSI-EVO
Velocity sensor	Hall-effect sensor

Table 23.2 The task example in the smart car prototype

Identity	Period	WCET	Deadline
T1 (velocity monitoring task)	4	4	4
T2 (light monitoring task)	6	3	6

Fig. 23.3 The spotlight experiment of the smart car prototype



$$\begin{aligned}
 \text{Since } w_2^1(t) &= w_2^2(t), S_2(t) = S_2(t) + v_i(t, 2) \\
 &= S_2(t) + \min(2, \min(0, 2)) = 0 + 0 = 0, w_2^2(t) = w_2^2(t) + \varepsilon = 3;
 \end{aligned}$$

Iteration 3: $w_3^3(t) = S_2(t) + c_1(t) + C_1 + c_2(t) = 3$ and since $w_2^2(t) = w_2^3(t)$, $S_2(t) = S_2(t) + v_i(t, 3) = S_2(t) + \min(2, \min(6, 1)) = 0 + 1$, $w_3^2(t) = w_3^2(t) + 1 + \varepsilon = 5$. $w_3^2 > d_2(t)$ so $S_2(t) = 1$.

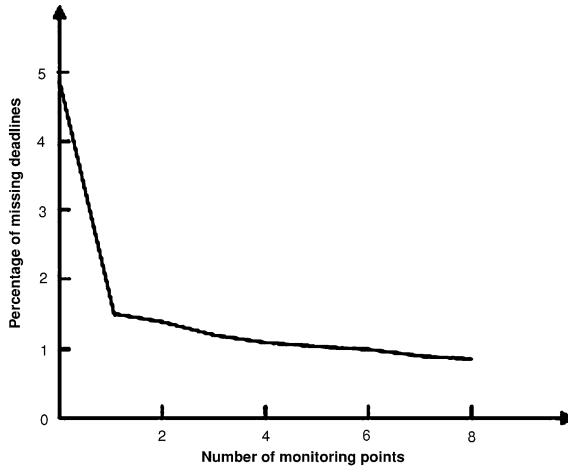


Fig. 23.4 Percentage of missing deadlines as the rate of monitoring in the smart car prototype

Test 2: We measure the quality of the smart car by the percent of tasks missing deadline. Figure 23.4 shows the average missing deadlines when the rate of monitoring is varied. These results suggest that the availability results leads to improved performance in terms of increased quality, but has little effect on the percentage of missed deadlines. The result relating to the number of possible quality values suggests that a scheduler could reduce its runtime by reducing the number of distinct quality values in the methods it is scheduling. This approximation will have the effect of reducing the precision of the final schedule. However, the scheduler will always find a schedule in those tasks.

23.6 Conclusion

As a promising application area of ubiquitous computing, smart car is becoming more and more important. The most concerned problem is the real-time performance. The contribution of this paper is to enhance the smart car with real-time guarantees. We develop a real-time scheduler to meet deadlines of periodic and reactive tasks simultaneously.

There are still many issues waiting for more consideration. So far, this paper just discusses tasks of soft deadline. Our future work will take tasks with hard deadline into account, and hence more improvement of timely performance can be reached.

Acknowledgments This research is supported by Ningbo Natural Science Foundation (2010A610108), NSF of Zhejiang Province, China (No.Y1080123), Foundation of MHRSS of China (No. 2009-416), and Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry (SRF for ROCS, SEM, No. 2009-1590).

References

1. Moite S (1992) How smart can a car be. In: Proceedings of the Intelligent Vehicles '92 Symposium, pp 277–279
2. Cheng AMK (2002) Real-time systems: scheduling analysis and verification. University of Huston, Huston
3. Locke CD (1992) Software architecture for hard real-time applications cyclic executives vs fixed priority executives. *J Real-Time Sys* 4(1):37–53 Kluwer
4. Cottet F, Mammeri Z (2002) Scheduling in real-time systems. John Wiley, England
5. Zilberstein S, Russell S (1995) Approximate reasoning using anytime algorithms. In: Natarajan S (ed) *Imprecise and approximate computation*. Kluwer Academic Publishers, The Netherlands
6. Zhao M, Wu Z, Yang G, Wang L, Chen W (2004) SmartOsek: a real time operating system for automotive electronics. The first IEEE international conferences on embedded software and systems, LNCS:3605, pp437–442

Chapter 24

A Study of Position Tracking Technology in Indoor Environments for Autonomous Wheelchair Driving

Sung-Min Kim, Jae-Hoon Jeong, Jung-Hwan Lee
and Sung-Yun Park

Abstract Various kinds of wheelchairs have been developed to secure mobility of the aged and disabled people for convenience of Activity of daily living (ADL). Accurate positioning technology is necessary to secure the mobility for realization of autonomous driving technology among the versatile wheelchair functions. In case of global positioning system (GPS), which is most widely used for the provision of positioning information, it is difficult to provide the accurate indoor positioning information. This paper has suggested active RFID system to overcome such shortcomings and to embody the indoor positioning technology. The error ratio has been checked out by the active RFID system implemented in the limited test space, after making compact and low-power active RFID module, measuring Time of Arrival (ToA), and estimating the positioning information using Received Signal Strength Indication (RSSI). It has been confirmed as of the estimation result of error ratio in the test space that accurate indoor positioning can be provided within about ± 2 cm deviation. Continuing research will be necessary in order to provide accurate indoor positioning information of autonomous driving wheelchair.

Keywords GPS · Active RFID · Wheelchair · Tracking · Environments

S.-M. Kim (✉) · J.-H. Jeong · S.-Y. Park
Department of Medical Biotechnology, College of Life Science and Biotechnology
Dongguk University, 26, Pil-dong, Jung-gu, Seoul 100-715, South Korea
e-mail: sungmin2009@gmail.com

J.-H. Lee
School of Biomedical Engineering, College of Biomedical & Health Science, Konkuk
University, 332 Danwol-dong, Chung-ju-si, Chungcheongbuk-do 380-701, South Korea

24.1 Introduction

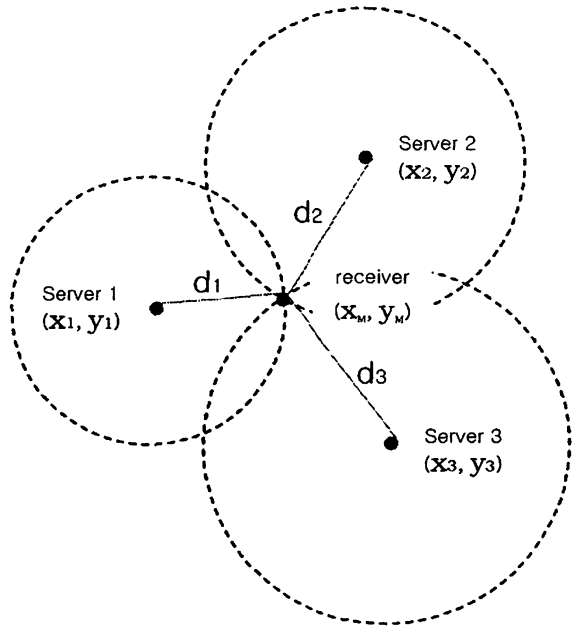
Global Positioning System (GPS) is for solving the position measuring problem in the global scale, which requires current time, location of satellite, and delay of signal to calculate the position of the GPS receiver [1]. GPS can measure the position accurately in open space without obstacles, as it is using the satellite signal [2], however, GPS signal may generate errors as the signal transmitted from satellite can be refracted and reflected by the terrain features such as buildings surrounding the receiver due to the impact of multi-path [1]. As of such errors, it is difficult to rely on the accuracy of indoor positioning information when using the GPS system in the indoor space [3].

7~10% of total world population are disabled, according to the WHO, and the aged ratio is expected to increase more and more as of the population's aging trend. In case of Korea, registration number of the disabled is 2 million 246 thousand as of Dec, 2008, which reaches 4.5% of total population, and the number has been increased by annual average 11.2% based on 2000. As the population aging trend and the number of disabled have been increasing also in the whole world, their social participation is being emerged as an important social issue. Variety of social accessibility is one of the core factors as a premise for the disabled to secure living quality through cultural life in the leisure and comfort activities. The most important element for the improvement of accessibility is how to secure the mobility. In order to secure the mobility, multifunctional wheelchairs have been developed to provide the convenience.

Most of the multifunctional wheelchair users are the severely disabled, so it is not easy for them to drive to the directions desired. The base technology for realization of autonomous driving for the wheelchair users experiencing the difficulty in autonomous driving is the accurate point positioning technology. As it is difficult to provide accurate indoor positioning information with GPS system [2], wireless indoor positioning systems have become very popular in recent years. Wireless indoor positioning systems are studying about indoor positioning based on infrared sensors [4], ultrasonic sensor [5, 6]. There are three technologies commonly used for indoor positioning systems—ultrasonics, infrared and rf. It will help smooth autonomous driving of wheelchair if indoor positioning information can be provided with active RFID system [7].

Among various researches on the indoor position measuring systems, this study has used the estimation method of indoor positioning information by calculation of Time of Arrival (ToA) values using the RF signal [8–10]. In order to embody the position measuring system that can be used connecting with wheelchair navigation or compact monitoring system, the positioning error ratio has been measured by measurement of the ToA between modules with Received Signal Strength Indication (RSSI) after implementing compact and low-power active RFID modules.

Fig. 24.1 ToA use to triangulation



24.2 Position Estimation Method

In the indoor environment, positioning technology of object such as wheelchair is an essential technology to construct the autonomous driving system.

The most basic principle of the technology is triangulation, which is difficult to trace accurate position due to the Non-Line-of Sight (NLOS) environment such as reflection, diffraction, and dispersion of RF signal by the obstacles of building exterior walls or metals as it is using the ToA signal. In order to solve such problem, it is necessary how to estimate the positioning data generated by triangulation [11]. This study has estimated the position by ToA using the RSSI.

24.2.1 Triangulation

Triangulation is a simple geometry, which is the most popular method to estimate the real-time position of moving object on the 2-dimensional plane. To estimate the real-time position of moving object, at least three or more datum points are required. The distance between the signal receiver and source can be obtained by measuring the arrival time of radio wave (ToA) between 3 signal sources as datum points and the receiver. As is shown in Fig. 24.1, ToA is the position tracing method by calculating absolute time from the terminal to the minimum 3 receivers [12, 13]

Lots of dispersed moving objects (receivers) are used for receiving the signal transmitted from signal source, in general, and the accurate position of receiver can be estimated by ToA or Time Difference of Arrival (TDoA). For the position calculation of ToA based receiver, the circles of uniform distance for the ToA are drawn, and the intersecting point of these circles is calculated by the following Formula (24.1). The receiver distance is calculated through the ToA received from at least three signal sources.

$$\begin{aligned} d_1 &= \sqrt{(x_1 - x_M)^2 + (y_1 - y_M)^2} \\ d_2 &= \sqrt{(x_2 - x_M)^2 + (y_2 - y_M)^2} \\ d_3 &= \sqrt{(x_3 - x_M)^2 + (y_3 - y_M)^2} \end{aligned} \quad (24.1)$$

24.2.2 Position Estimation Using the RSSI

Triangulation RSSI method is a position estimation method using the principle that RF signal strength is decreasing exponentially, which has a weak point that quite a lot of post-processing is required to cut down the errors caused by multi-path and fading phenomenon, on the contrary it has strong points that power consumption is very low and realization of low cost sensor network is available. [14]

As RSSI method has adopted the triangulation principle as it is, it needs the process the obtaining distance from moving object to three datum points. The distance from received signal strength to moving object can be calculated by Friis formula.

$$L = 20 \log_{10} \left(\frac{4\pi f d}{\lambda} \right) [dB] \quad (24.2)$$

The Friis formula in Formula (24.2) is for obtaining the path loss in free space, where λ indicates the wavelength of radio wave, which uses the same unit of distance d . When expressing Formula (24.2) for d , the distance between two points, it becomes Formula (24.3), where c is the radio speed and f is its frequency.

$$D = \frac{\lambda}{4\pi} \cdot 10^{\frac{L}{20}} = \frac{c}{4\pi f} \cdot 10^{\frac{L}{20}} \quad (24.3)$$

24.3 Realization of Active RFID Module

24.3.1 RF Transmission and Receiving

There are Zigbee method that satisfies IEEE 802.15.4 specification and open platform method that is using only baseband (2.4 GHz) for obtaining the RSSI signal of RF. This study has adopted the CC2500 IC, which is a closed protocol

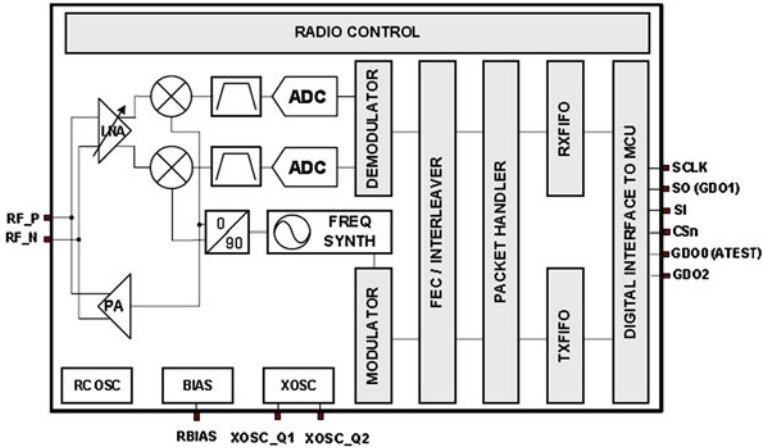


Fig. 24.2 Block diagram of CC2500

method to avoid conflict with other peripherals of computer and to apply various protocols.

CC2500 IC is using the frequency band width of 2400–2483.5 MHz Industrial, Scientific and Medical (ISM) and Short Range Device (SRD), and its transmission speed of 500 k Baud. Also it is very useful to control the distance change by variation of environment as the RF output can be controlled. Figure 24.2 shows the structure of RF Tx/Rx IC. The signal received is passing the Low Noise Amplifier (LNA) and Quadrature (I, Q) Down-Conversion after passing through the RF Front-End. This signal is processed at Limiter-Discriminator after conversion to 2 MHz Low-Intermediate Frequency (IF) and connected to the digital part, where the signal strength received is converted to a value for measuring the RSSI. Also filtering, demodulation, and frame synchronization of the channel are processed in this part. C2500 has been used for this purpose. It is the DSSS type, which can transmit 250 kbps and supports low power consumption (RX: 13.3 mA, TX: 17.4 mA). The operation power is available up to 1.6–3.6 V, which is thought to be suitable to embody the sensor node using battery.

24.3.2 Micro Controller for the Control

MSP430 F2274 of low-power and high performance has been used as micro controller for the control of CC2500 and for the verification of algorithm (Fig. 24.3).

The MSP430 series processor is constituted with 16 bit core, of which computing capability is very high. It shows low-power capability of 270 μ A at 2.2 V@1 MHz as it is driven at low-power of 1.8–3.6 V. Its inside Digital Clock

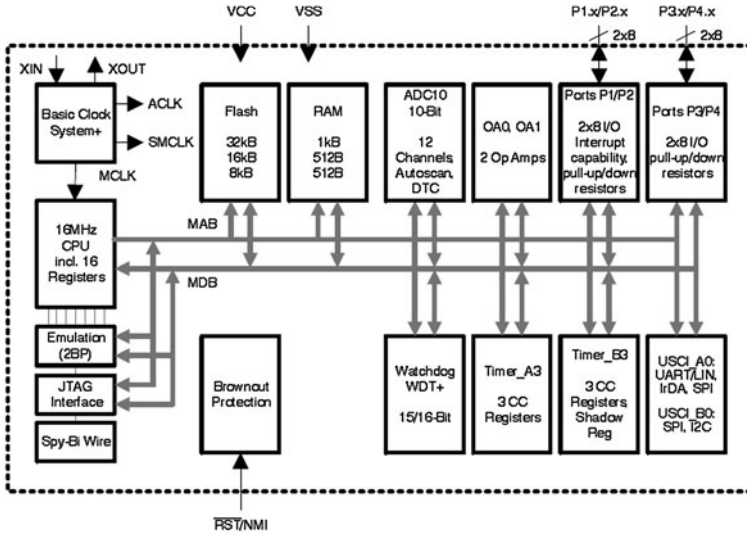


Fig. 24.3 Functional block diagram of MSP430F2274

Oscillator (DCO) logic can provide the operation frequency of 16 MHz \pm 1% without connecting any crystal externally. The flash memory for internal programs has the 32 kB, 1 kB RAM area, which is suitable for the various protocol applications of sensor network.

24.3.3 Embodiment of Active RFID Module

Experimental circuits for wireless Tx/Rx module had been designed using the CC2500 chip and MSP430F2274 as micro controller, and its PCB has been manufactured. Figure 24.4 shows the wireless Tx/Rx board made up of consolidated CC2500 RF chip and low-power MSP430F2274 explained in the above.

24.4 Experiment

24.4.1 Experiment to Estimate the RSSI Value

Experiment has been carried out using the active RFID module implemented in order to induce out the estimation formula for RSSI value complying with the variation of distance. The RSSI value of CC2500 can be obtained easily by reading the register value of RSSI status, and the reading value (RSSI dBm) is the

Fig. 24.4 RF transceiver board

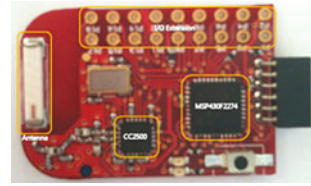
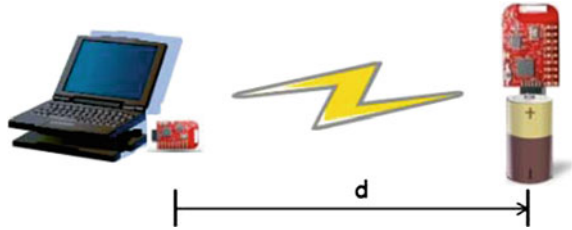


Fig. 24.5 The composition of RSSI test



complement on 2. By adding internal offset (-75 dBm; CC2500) on to the value, the RSSI of actually received signal is calculated with the Formula (24.4).

$$RSSI_{dBm} = RSSI_{Reg} + RSSI_{offset}[dBm] \tag{24.4}$$

The experimental configuration for measuring the variation of RSSI value complying with the distance between active RFID modules manufactured is illustrated in Fig. 24.5.

The graph as a result of signal vs. distance relation is shown in Fig. 24.6 which is obtained using the RSSI values measured by changing the distances of 3 modules when active RFID module output is 0 dBm. The three Tx/Rx modules have indicated almost same relation, which shows the characteristics of diminishing exponentially in the 0~50 cm section and slowly decreasing linearly in the section over 100 cm.

Through the RFID values obtained by each RF Tx/Rx modules and the distance data, the formula has been estimated in the form of 2nd fractional polynomials, which is shown in Formula (24.5).

$$Distance(RSSI) = \frac{P_1 \times RSSI^2 + P_2 \times RSSI + P_s}{RSSI^2 + q_1 \times RSSI + q_2}$$

$$p_1 = -13.63$$

$$p_2 = -1674$$

$$p_3 = -3.922 E + 04$$

$$q_1 = 135.6$$

$$q_2 = 4645 \tag{24.5}$$

The receiver position tracing experiment has been performed in the limited test space with the estimation formula loaded on the module.

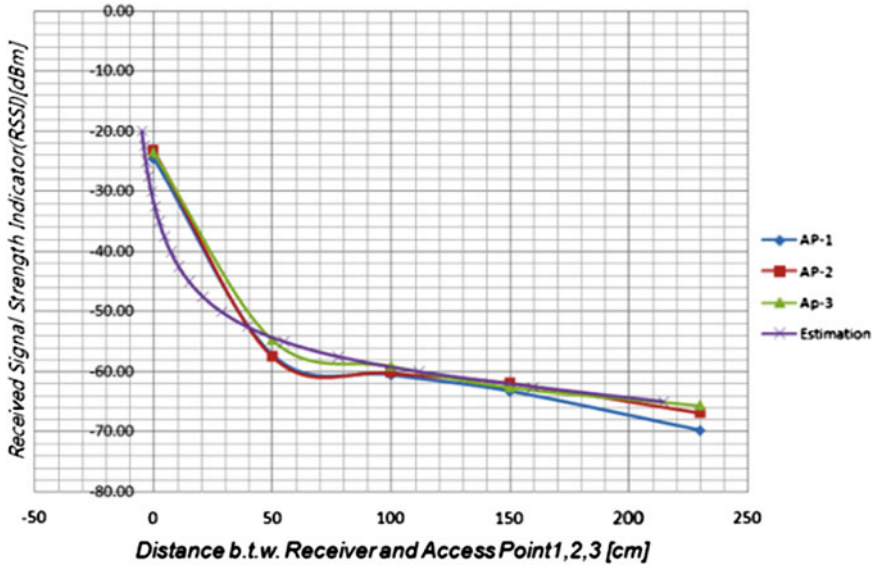
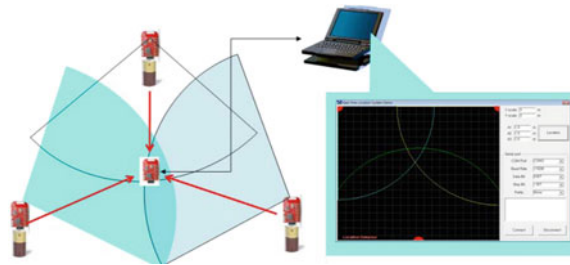


Fig. 24.6 Measurement data of three modules

Fig. 24.7 Test setup & display



24.4.2 Position Tracing Experiment in the Test Space

RF Tx/Rx modules have been arranged in triangular structure on arbitrary locations in the test space of $3\text{ m} \times 3\text{ m} \times 3\text{ m}$, and the locations are defined as A1, A2, and A3. RF modules had been fixed on each A1, A2, and A3 location, and additional modules were arranged on the locations for measuring. Then the locations were changed by 1 m from A1, A2, and A3 and the changed locations have been traced.

The principle of measurement in the test space has been the execution of position estimation by triangulation as is shown in Fig. 24.7, which was monitored on the screen in order to check the position tracing of moved receiver in the

Fig. 24.8 Indoor test space

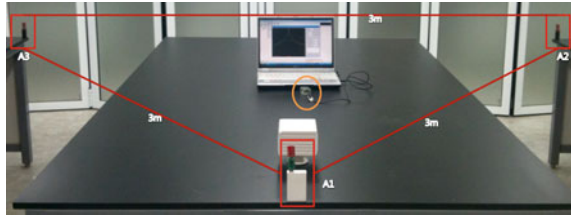


Table 24.1 Font sizes of headings. Table captions should always be positioned *above* the tables

Position	Distance 1 m from A1			Distance 1 m from A2			Distance 1 m from A2			Max error
	1	2	3	1	2	3	1	2	3	
A1 RSSI	-66	-65	-67	-81	-77	-81	-81	-82	-82	4
A2 RSSI	-76	-80	-80	-66	-64	-63	-78	-78	-77	4
A3 RSSI	-74	-73	-73	-81	-80	-82	-65	-66	-66	2



Fig. 24.9 Result of A1, A2, A3 position

triangular structure in arbitrary test space. Actual test space is shown in Fig. 24.8, and Table 24.1 shows the measurement result after moving receivers on the table to the locations 1 m away from each corner, and the positions of each module on the screen are shown in Fig. 24.9.

This study has checked the variation of RSSI values complying with the variation of distances using stable RSSI values after arranging RF Tx/Rx modules on the locations 1 m away respectively from the A1, A2, A3 positions.

24.5 Conclusion

Active RFID system has been implemented for tracing the position in the indoor space together with the GPS information used for outdoor position tracing, paying attention to the point that most of the wheelchair users are living in the limited space. RSSI value has been measured by using the modules implemented, and the variation of signal complying with the module distance has been checked out.

After analyzing the characteristic that signal is slowly declined linearly according to the distance, the estimation formula has been extracted in the form of 2nd fractional polynomials. The experiment for tracing the positions of receivers in limited test space has been carried out with the estimation formula loaded on the modules manufactured. As a estimation result of the position in the limited indoor test space (3 m × 3 m × 3 m), maximum error ratio has been ± 4 dBm. When converting the result value into distance, the distance error ratio is ± 2 cm. It has been proved, through this study, that accurate positioning information can be provided when active RFID system is used for the provision of indoor positioning information. However, additional researches will be required in order to provide accurate positioning information of moving wheelchair.

It is expected that algorithm research will be necessary for tracing mobile positions based on the result of this study, also it is understood that similar error ratio to this study must be maintained for moving positions, considering the space allocation of indoor living environment, in order to provide accurate indoor positioning information for autonomous driving of wheelchair in the end.

Acknowledgments This work was supported by the Technology Innovation Program (Industrial Strategic Technology Development Program, 10032055) funded by the Ministry of Knowledge Economy (MKE, Korea)

References

1. Martin EH (1980) GPS user equipment error models. *Glob Position Syst* 1:109–118
2. Van Dierendonck AJ et al (1978) The GPS navigation message. *J Navig* 25(2):147–165
3. Kaplan ED (1996) *Understanding GPS: principles and applications*, 1st edn. Artech House, Boston
4. Koren Y, Borenstein J (1989) Obstacle avoidance with ultrasonic sensors. *IEEE J Robot Autom* 4(2):213–218
5. Kim HT, Park M (1944) Tangential segment based half step path planning for sonar based navigation. In: *Proceedings of the 1st Asian control conference, Japan*, vol 2, pp 945–949
6. Argyros AA, Bergholm F (1999) Combining central and peripheral vision for reactive robot navigation In: *IEEE computer society conference on computer vision and pattern recognition*, vol 2, pp 23–25
7. Wang H, Lee R, Hsiao C, Hsieh G (2010) Active RFID system with cryptography and authentication mechanisms. *J Inf Sci Eng* 26:1323–1344
8. Pahlavan K et al (2006) Indoor geolocation in the absence of direct path. *IEEE Wirel Commun Mag* 13(6):50–58
9. Pahlavan K, Krishnamurthy P, Beneat J (1998) Wideband radio propagation modeling for indoor geolocation applications. *IEEE Commun Mag* 36(4):60–65
10. Heidari M, Pahlavan K (2008) Identification of the absence of direct path in ToA-based indoor localization systems. *Int J Wirel Inf Netw* 15:117–127
11. Bae J, Choi Y, Kim J (2010) Performance analysis of TWR positioning technique with ultra wideband in indoor channel. *Inst Electron Eng Kores* 47(7):18–22
12. Bechteler TF, Yenigun H (2003) 2-D localization and identification based on SAW ID-tags at 2.5 GHz. *IEEE Trans Microw Theory Tech* 51(5):1584–1590

13. Thomas F, Ros L (2005) Revisiting trilateration for robot localization. *IEEE Trans Robot* 21(1):93–102
14. Véronneau S, Roy J (2009) RFID benefits, costs, and possibilities: the economical analysis of RFID deployment in a cruise corporation global service supply chain. *Int J Prod Econ* 122(2):692–702

Chapter 25

Adaptive Regulation of a Class of Nonlinear Systems with Unknown Sinusoidal Disturbances

Pengnian Chen

Abstract This chapter addresses adaptive output regulation for a class of nonlinear systems with sinusoidal disturbances. The gain matrix of the sinusoidal disturbances depend the system output. The disturbances undertaken are unmatched. The chapter presents a new method for estimating sinusoidal disturbances. Based on the estimation method, a method of adaptive output regulation is presented, which can guarantee that all the signals in the closed loop are bounded and the output converges to zero.

Keywords Adaptive control • Nonlinear system • Disturbance rejection

25.1 Introduction

Consider a nonlinear control system of the form

$$\begin{aligned}\dot{x} &= Ax + \phi(y) + bu + G(y)w \\ y &= Cx \\ \dot{w} &= Sw\end{aligned}\tag{25.1}$$

where $x(t) \in R^n$ is the state vector, $u(t) \in R$ the control input, $y(t) \in R$ the output vector, and $w \in R^l$ is the disturbance vector; $\phi(y)$ is a known vector of smooth functions with $\phi(0) = 0$, the linear system $\sum (C, A, b)$ is in the canonical form

P. Chen (✉)
College of Mechatronics Engineering, China Jiliang University
Hangzhou 310018, People's Republic of China
e-mail: pnchen@cjlu.edu.cn

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}, \quad c = (1 \quad 0 \quad \cdots \quad 0),$$

and S is an unknown matrix.

The adaptive control of uncertain nonlinear systems such as (25.1) has been received much attention [1–5]. Adaptive tracking is considered in [2], in which the control direct is unknown and the gain of disturbances is constant. In [3], the adaptive output regulation is considered under the assumption that the exosystem $\dot{w} = Sw$ has $2m + 1$ simple distinct eigenvalues on the imaginary axis and all the modes of the exosystem are persistently excited. The method of output regulation in [3] can guarantee the system output converges to zero exponentially. Under the assumption that all the frequency components in the disturbance system are fully excited, adaptive stabilization is studied in [4]. Adaptive tracking for a class of nonlinear system of triangular form with sinusoidal disturbances by state feedback is considered in [6].

The gain matrix of the disturbances, that is, $G(y)$, is a constant matrix in the studies mentioned above studies [1–5]. However, the gain matrix of disturbances, in general, is not a constant matrix in many practical applications [6–8].

In this study, we consider the problem of adaptive output regulation for system (25.1), in which the gain matrix depends on the output. A new estimation method for disturbances is presented. Based on the estimation method, a method of adaptive output regulation is presented, which can guarantee all the signals in the closed loop system are bounded and the output converges to zero.

We make the following assumptions for system (25.1):

Assumption 1: The eigenvalues of S are all simple and on the imaginary axis.

Assumption 2: $\sum (C, A, b)$ has relative ρ and b is a Hurwitz vector with $b_\rho = 1$, that is, $b_1 = \cdots = b_{\rho-1} = 0$, and $p(s) = s^{n-\rho} + b_{\rho+1}s^{n-1-\rho} + b_{n-1}s + b_n$ is a Hurwitz polynomial.

The rest of the chapter is organized as follows. Section 2 is a system transformation. Section 3 presents an estimation method of the sinusoidal disturbances. Section 4 is the main result of the chapter, in which a law of adaptive output regulation for system (25.1) is presented. Some concluding remarks are drawn in Sect. 5.

25.2 System Transformations

For control design, we transform system (25.1) into a new form via a filtered transformation.

Let $\bar{G}(y) = G(y) - G(0)$. Then

$$G(y) = G(0) + yH(y) \tag{25.2}$$

where

$$H(y) = \int_0^1 \frac{d\bar{G}(\sigma y)}{d\sigma} d\sigma.$$

Let $\lambda_i > 0, i = 1, 2, \dots, n - 1$, and define

$$\begin{aligned} d_n &= (0, 0, \dots, 0, 1)^T \\ d_i &= Ad_{i+1} + \lambda_i d_{i+1} \quad i = 1, 2, \dots, n - 1. \end{aligned} \tag{25.3}$$

According to Assumption 1, there exist $\Pi \in R^{n \times l}$ and $L \in R^l$ such that the following equations hold [3]:

$$\Pi S - A\Pi + d_1 L^T - G(0) = 0 \quad C\Pi = 0. \tag{25.4}$$

It is easy to see that d_1 is a Hurwitz vector with $d_{11} = 1$, where d_{11} is the first element of d_1 , and d_1, d_2, \dots, d_n are linearly independent. Consequently, there exist smooth functions $h_i(y), i = 1, 2, \dots, n$, such that

$$H(y) = \sum_{i=1}^n h_i(y) d_i. \tag{25.5}$$

Let $\xi = (\xi_1, \xi_2, \dots, \xi_{n-1})^T$ and $H_0(y) = (h_2(y), h_3(y), \dots, h_n(y))^T$. Define a dynamic system

$$\dot{\xi} = A_\xi \xi + yH_0(y)w \tag{25.6}$$

where

$$A_\xi = \begin{pmatrix} -\lambda_1 & 1 & & & & \\ & -\lambda_2 & 1 & & & \\ & & \ddots & \ddots & & \\ & & & -\lambda_{n-2} & 1 & \\ & & & & -\lambda_{n-1} & \end{pmatrix} \tag{25.7}$$

Since A_ξ is a Hurwitz matrix, there exists an $(n - 1) \times (n - 1)$ positive definite symmetric matrix P_ξ such that

$$P_\xi A_\xi + A_\xi^T P_\xi = -(k_1 + 1)I_{n-1} \tag{25.8}$$

where $k_1 > 0$ is a constant and I_{n-1} is the unit matrix of order $n - 1$.

Let $V_\xi = \xi^T P_\xi \xi$. Evidently, it is easy to prove that the following inequality holds:

$$\dot{V}_\xi \leq -k_1 \|\xi\|^2 + y^2 \|H_0(y)\|^2 \|P_\xi\|^2 \|w\|^2. \quad (25.9)$$

Now we perform the filtered transformation:

$$x_f = x - \Pi w - \sum_{j=2}^n d_j \zeta_{j-1}. \quad (25.10)$$

By a straightforward calculation, we have

$$\begin{aligned} \dot{x}_f &= \dot{x} - \Pi \dot{w} - \sum_{j=2}^n d_j \dot{\zeta}_{j-1} \\ &= Ax + \phi(y) + bu + G(0)w + yH(y)w - \Pi Sw \\ &\quad - \sum_{j=2}^n d_j (-\lambda_{j-1} \zeta_{j-1} + \zeta_j + y h_j(y)w). \end{aligned} \quad (25.11)$$

By (25.3) and (25.4), we have

$$G(0)w - \Pi Sw = -A\Pi w + d_1 L^T w \quad (25.12)$$

and

$$\begin{aligned} yH(y)w - \sum_{j=2}^n d_j (-\lambda_{j-1} \zeta_{j-1} + \zeta_j + y h_j(y)w) \\ &= y d_1 h_1(y) + \sum_{j=2}^n \lambda_{j-1} d_j \zeta_{j-1} - \sum_{j=2}^{n-1} d_j \zeta_j \\ &= y d_1 h_1(y) + \sum_{j=2}^n (d_{j-1} - A d_j) \zeta_{j-1} - \sum_{j=2}^{n-1} d_j \zeta_j \\ &= d_1 (\zeta_1 + y h_1(y)w) - A \sum_{j=2}^n d_j \zeta_{j-1} \end{aligned} \quad (25.13)$$

where $\zeta_n = 0$. According to (25.11)–(25.13), system (25.1) is transformed into

$$\begin{aligned} \dot{x}_f &= Ax_f + \phi(y) + bu + d_1 (\zeta_1 + y h_1(y)w + L^T w) \\ y &= Cx_f. \end{aligned} \quad (25.14)$$

In order to simplify the notations, we rewrite (25.14) as

$$\begin{aligned} \dot{x} &= Ax + \phi(y) + bu + d_1 (\zeta_1 + y h_1(y)w + L^T w) \\ y &= Cx. \end{aligned} \quad (25.15)$$

We make the following assumption for system (25.15):

Assumption 3: The pair of matrices (L^T, A) is observable.

In what follows, we assume that Assumptions 1–3 hold.

25.3 Estimation of Disturbances

In this section, we use the method developed in [4] to establish an estimation of states of system (25.1) and to parameterize w .

Let

$$\dot{\hat{x}} = (A - L_1 C)\hat{x} + \phi(y) + bu + L_1 y \quad (25.16)$$

where $L_1 \in R^n$ is chosen such that $A - L_1 C$ is a Hurwitz matrix. Let $\tilde{x} = x - \hat{x}$. Then we have

$$\dot{\tilde{x}} = (A - L_1 C)\tilde{x} + d_1(\xi_1 y h_1(y)w + L^T w). \quad (25.17)$$

Let $Q \in R^{n \times l}$ be a solution of the matrix equation

$$QS = (A - L_1 C)Q + d_1 L^T \quad (25.18)$$

where d_1 is defined in (25.3) and L is defined in (25.4). Let Q_1 be the first row of Q . According Assumption 3, (L^T, S) is observable. Thus (Q_1, S) is also observable [4]. Let $q = Qw$. Then we have

$$\dot{q} = (A - L_1 C)q + d_1 L^T w. \quad (25.19)$$

Let $\varepsilon = x - \hat{x} - q$. Then

$$\dot{\varepsilon} = (A - L_1 C)\varepsilon + d_1(\xi_1 + y h_1(y)w). \quad (25.20)$$

Choose $F \in R^{l \times l}$ and $g \in R^l$ such that F is a Hurwitz matrix and (F, g) is a controllable pair. Let E be a solution of the matrix equation

$$ES - FE = gQ_1 \quad (25.21)$$

Then E is non-singular [4]. Now we parameterize w as $w = E^{-1}\eta$. By a straightforward calculation, we have

$$\dot{\eta} = F\eta + gq_1 \quad (25.22)$$

where q_1 is the first element of q . Let $\hat{\eta}$ be an estimate of η , defined by

$$\dot{\hat{\eta}} = F\hat{\eta} + g(y - \hat{x}_1) \quad (25.23)$$

where \hat{x}_1 is the first element of \hat{x} . Let $\tilde{\eta} = \eta - \hat{\eta}$. Then

$$\dot{\tilde{\eta}} = F\tilde{\eta} + g\varepsilon. \quad (25.24)$$

Since $A - L_1C$ and F are Hurwitz matrices, there are matrices $P_\varepsilon > 0$ and $P_{\tilde{\eta}} > 0$ such that

$$\begin{aligned} P_\varepsilon(A - L_1C) + (A - L_1C)^T P_\varepsilon &= -(k_2 + 2)I_n \\ P_{\tilde{\eta}}F + F^T P_{\tilde{\eta}} &= -(k_3 + 1)I_l \end{aligned} \quad (25.25)$$

where $k_2 > 0$ and $k_3 > 0$. Let

$$V_\varepsilon = \varepsilon^T P_\varepsilon \varepsilon V_{\tilde{\eta}} = \tilde{\eta}^T P_{\tilde{\eta}} \tilde{\eta} \quad (25.26)$$

Then by a straightforward calculation, we have

$$\dot{V}_\varepsilon \leq -k_2 \|\varepsilon\|^2 + \|\xi_1\|^2 \|P_\varepsilon d_1\|^2 + y^2 \|h_1(y)\|^2 \|P_\varepsilon d_1\|^2 \|w\|^2 \quad (25.27)$$

and

$$\dot{V}_{\tilde{\eta}} \leq -k_3 \|\tilde{\eta}\|^2 + \|\varepsilon\|^2 \|P_{\tilde{\eta}} g\|^2. \quad (25.28)$$

25.4 Adaptive Regulation Laws

In this section, let $\rho = 1$. We first perform a coordinate transformation for (25.15). Let $d_1 = (d_{11}, d_{12}, \dots, d_{1n})^T$ be defined in (25.3). Define

$$z_i = x_{i+1} - d_{i+1}x_1 \quad i = 1, 2, \dots, n-1. \quad (25.29)$$

Then system (25.15) is transformed into

$$\begin{aligned} \dot{y} &= z_1 + \bar{\phi}_0(y) + u + \xi_1 + y h_1(y) w + L^T w \\ \dot{z} &= Dz + \bar{\phi}(y) + \bar{b}u \end{aligned} \quad (25.30)$$

where $z = (z_1, z_2, \dots, z_{n-1})^T$, $\bar{\phi}(y) = (\bar{\phi}_1(y), \bar{\phi}_2(y), \dots, \bar{\phi}_{n-1}(y))^T$ and

$$D = \begin{pmatrix} -d_{12} & 1 & 0 & \cdots & 0 \\ -d_{13} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ -d_{1n-1} & 0 & 0 & \cdots & 1 \\ -d_{1n} & 0 & 0 & \cdots & 0 \end{pmatrix}, \quad \bar{b} = \begin{pmatrix} b_2 - d_{12} \\ b_3 - d_{13} \\ \vdots \\ b_n - d_{1n} \end{pmatrix}.$$

According to Assumption 2, D is a Hurwitz matrix. We now design the following observer to reconstruct z :

$$\dot{\hat{z}} = D\hat{z} + \bar{\phi}(y) + bu \quad (25.31)$$

where $\hat{z} = (\hat{z}_1, \hat{z}_2, \dots, \hat{z}_{n-1})^T$ is an estimate of z .

Let $\tilde{z} = (\tilde{z}_1, \tilde{z}_2, \dots, \tilde{z}_{n-1})^T$ with $\tilde{z}_i = z_i - \hat{z}_i$, $i = 1, 2, \dots, n-1$. Then we have

$$\dot{\tilde{z}} = D\tilde{z}. \quad (25.32)$$

This follows that $\tilde{z}(t)$ converges to zero exponentially as t approaches to positive infinity.

According to (25.32), the first equation of (25.30) can be written as

$$\dot{y} = \hat{z}_1 + \tilde{z}_1 + \bar{\phi}_0(y) + u + \xi_1 + yh_1(y)w + \eta^T \theta \quad (25.33)$$

where $\theta = (E^{-1})^T L$, which is an unknown constant vector. Let

$$\mu = \max \left\{ \|P_\xi\|^2 \|w\|^2, \|P_\varepsilon d_1\|^2 \|\varepsilon\|^2, \|w\| \right\}.$$

We propose the following adaptive control law for system (25.30):

$$u = -\alpha y - \hat{z}_1 - \bar{\phi}_0(y) - \hat{\eta}^T \hat{\theta} - y\phi(y)\hat{\mu} \quad (25.34)$$

where $\alpha > 0$; $\hat{\theta}$ and $\hat{\mu}$ are defined by

$$\begin{aligned} \dot{\hat{\theta}} &= \gamma_1 \hat{\eta} y \\ \dot{\hat{\mu}} &= \gamma_2 y^2 \phi(y) \end{aligned} \quad (25.35)$$

where $\gamma_1 > 0$, $\gamma_2 > 0$ and $\phi(y) = \|H_0(y)\|^2 + 1/2(3\|h_1(y)\|^2 + 1)$.

Theorem 1 Consider the closed loop system consisting of (25.30), (25.34) and (25.35). Let $\alpha > 3/2$. Then all signals in the closed loop system are bounded and

$$\lim_{t \rightarrow +\infty} y(t) = 0. \quad (25.36)$$

Proof. Under the control law (25.34), the closed loop system can be written as

$$\begin{aligned} \dot{y} &= -\alpha y + \hat{\eta}^T \tilde{\theta} + \tilde{\eta}^T \theta - y\phi(y)\hat{\mu} + \xi_1 + yh_1(y)w + \tilde{z}_1 \\ \dot{z} &= Dz + \bar{\phi}(y) + \bar{b}u. \end{aligned} \quad (25.37)$$

Let

$$V = \frac{1}{2}y^2 + \frac{1}{2\gamma_1}\tilde{\theta}^2 + \frac{1}{2\gamma_2}\tilde{\mu}_l^2 + l_1V_\xi + l_2V_\varepsilon + l_3V_\eta \quad (25.38)$$

where $\tilde{\theta} = \theta - \hat{\theta}$, $\tilde{\mu}_l = l\mu - \hat{\mu}$; $l > 0$ and $l_i > 0$, $i = 1, 2, 3$, are constants, which are chosen later. Differentiating V yields

$$\begin{aligned} \dot{V} = & -\alpha y^2 + y\dot{\eta}^T\tilde{\theta} + y\tilde{\eta}^T\theta - y^2\phi(y)\hat{\mu} + y\xi_1 + y^2h_1(y)w \\ & - \frac{1}{\gamma_1}\tilde{\theta}^T\dot{\tilde{\theta}} - \frac{1}{\gamma_2}\tilde{\mu}_l\dot{\hat{\mu}} + l_1\dot{V}_\xi + l_2\dot{V}_\varepsilon + l_3\dot{V}_{\tilde{\eta}} + y\tilde{z}_1. \end{aligned} \quad (25.39)$$

Since

$$\begin{aligned} y\tilde{z}_1 & \leq \frac{1}{2}y^2 + \frac{1}{2}\tilde{z}_1^2 \\ y\xi_1 & \leq \frac{1}{2}y^2 + \frac{1}{2}\xi_1^2 \\ y\tilde{\eta}^T\theta & \leq \frac{1}{2}y^2 + \frac{1}{2}\|\tilde{\eta}\|^2\|\theta\|^2 \end{aligned} \quad (25.40)$$

by (25.39) and (25.40), we have

$$\begin{aligned} \dot{V} \leq & -(\alpha - \frac{3}{2})y^2 + y\dot{\eta}^T\tilde{\theta} - y^2\phi(y)\hat{\mu} + \frac{1}{2}\|\xi_1\|^2 + \frac{1}{2}\|\tilde{\eta}\|^2\|\theta\|^2 \\ & + y^2|h_1(y)|\|w\| - \frac{1}{\gamma_1}\tilde{\theta}^T\dot{\tilde{\theta}} - \frac{1}{\gamma_2}\tilde{\mu}_l\dot{\hat{\mu}} + l_1\dot{V}_\xi + l_2\dot{V}_\varepsilon + l_3\dot{V}_{\tilde{\eta}} + \tilde{z}_1^2. \end{aligned} \quad (25.41)$$

We choose l_i , $i = 1, 2, 3$, such that $l_1 > 1 + l_2\|P_\varepsilon d_1\|^2$, $l_2k_2 > 1 + l_3\|P_{\tilde{\eta}}g\|^2$ and $l_3k_3 > 1 + \frac{1}{2}\|\theta\|^2$. Then we have

$$\begin{aligned} l_1\dot{V}_\xi + l_2\dot{V}_\varepsilon + l_3\dot{V}_{\tilde{\eta}} & \leq -l_1k_1\|\xi\|^2 + l_1y^2\|H_0(y)\|^2\|P_\xi\|^2\|w\|^2 \\ & \quad - l_2k_2\|\varepsilon\|^2 + l_2\|\xi\|^2\|P_\varepsilon d_1\|^2 + l_2y^2\|h_1(y)\|^2\|P_\varepsilon d_1\|^2\|w\|^2 \\ & \quad - l_3k_3\|\tilde{\eta}\|^2 + l_3\|\varepsilon\|^2\|P_{\tilde{\eta}}g\|^2 \\ & \leq -\frac{1}{2}\|\xi\|^2 - \frac{1}{2}\|\varepsilon\|^2 - l_3k_3\|\tilde{\eta}\|^2 \\ & \quad + ly^2\left(\|H_0(y)\|^2 + \|h_1(y)\|^2\right)\mu \end{aligned} \quad (25.42)$$

where $l = \max\{l_1, l_2, 1\}$. By (25.41) and (25.42), we get that

$$\begin{aligned} \dot{V} \leq & -(\alpha - \frac{3}{2})y^2 - \|\xi\|^2 - \|\varepsilon\|^2 - \|\tilde{\eta}\|^2 \\ & + y\dot{\eta}^T\tilde{\theta} + y^2\phi(y)\tilde{\mu}_l - \frac{1}{\lambda_1}\tilde{\theta}^T\dot{\tilde{\theta}} - \frac{1}{\gamma_2}\tilde{\mu}_l\dot{\hat{\mu}} + \frac{1}{2}\tilde{z}_1^2. \end{aligned} \quad (25.43)$$

It follows from (25.35) and (25.43) that

$$\dot{V} \leq -\left(\alpha - \frac{3}{2}\right)y^2 - \|\xi\| - \|\varepsilon\|^2 - \|\tilde{\eta}\|^2 + \frac{1}{2}\tilde{z}_1^2. \quad (25.44)$$

Since $\alpha > 3/2$ and $\int_0^{+\infty} \tilde{z}_1^2(\tau)d\tau < +\infty$, by Barbalat's Lemma [2], it can be easily proved that (25.44) implies that

$$\lim_{t \rightarrow +\infty} y(t) = 0. \quad (25.45)$$

On the other hand, we can prove that all signals in the closed loop system are bounded.

The proof is completed.

25.5 Conclusion

In the chapter, we study the problem of output regulation for a class of nonlinear systems with sinusoidal disturbances. The gain matrix of the disturbances depends on the system output. The problem of adaptive output regulation has been solved in the case of relative degree one. If the gain matrix depends on the system states, the problem is still open. In addition, to extend the result of the paper to the case of the relative degree more than one is interesting.

Acknowledgments This work was supported by National Nature Science Foundation of China under Grants 21006086.

References

1. Bodson M, Douglas SC (1997) Adaptive algorithms for the rejection of sinusoidal disturbances with unknown frequencies. *Automatica* 33:2213–2221
2. Marino R, Tomei P (2005) Adaptive tracking and disturbances rejection for uncertain nonlinear systems. *IEEE Trans Autom Control* 50:90–95
3. Marin R, Santosuosso GL (2005) Global compensation of unknown sinusoidal disturbances for a class of nonlinear nonminimum phase systems. *IEEE Trans Autom Control* 50:1816–1822
4. Ding Z (2006) Adaptive estimation and rejection of unknown sinusoidal disturbances in a class of non-minimum-phase nonlinear systems. *IEEE Proc Control Theory Appl* 153:379–386
5. Ding Z (2003) Universal disturbance rejection for nonlinear systems in output feedback form. *IEEE Trans Autom Control* 48:1222–1227
6. Qu X, Sun M, Liu X (2010) Asymptotic attenuation of a class of nonlinear systems with unknown sinusoidal disturbances. *J Control Theory Appl* 8:509–514
7. Xu J-X, Xu J (2004) Observer based learning control for a class of nonlinear systems with time-varying parametric uncertainties. *IEEE Trans Autom Control* 49:275–281
8. Xu J-X (2004) A new periodic adaptive control approach with known periodicity. *IEEE Trans Autom Control* 49:579–588

Chapter 26

Research on Improved Ant Colony Algorithm for TSP Problem

Xiu-juan Qiu and Ting-gui Chen

Abstract A bionic optimization algorithm called ant colony optimization was introduced in this chapter. Based on the basic ant colony algorithm, this paper improves ant colony algorithms as follows: (1) increase the local pheromone updating link and change the original algorithm in the state transition principle; (2) select the next city by pseudo-random proportional rule instead of selection directly by probability. The simulation experiments show that the improved algorithm is better than the traditional one.

Keywords TSP · Combinatorial optimization problems · Ant colony optimization algorithm

26.1 Introduction

Ant colony optimization was first proposed in 1991 by the Italian scholars Dorigo M, which initially derived from the imitation of natural foraging behavior of biological ant colony [1, 2]. It is a bionic algorithm applied to the combinatorial optimization problems heuristic search algorithm which enjoys strong robustness, fine distributed compute system. So it is apt to combine with other methods, and has advantages in

X. Qiu (✉) · T. Chen

College of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou, 310018 Zhejiang, China
e-mail: qiu_altitude@sina.com

T. Chen

Contemporary Business and Trade Research Center,
Zhejiang Gongshang University, Hangzhou, 310018 Zhejiang, China
e-mail: ctgsimon@gmail.com

partial convergence. Although the ant colony optimization was first successfully applied in solving the well-known TSP problem [3], and shows people that it has advantages in terms of local convergence. But it still remains to be improved, especially the ability to get the best solution. The improved ant colony algorithm which was run in the computer successfully got the expected results.

26.2 Basic Ant Colony Algorithm

26.2.1 Principles

Basic ant colony optimization derives from ant's foraging behavior. Ants can always find the optimal path from the nest to the food resource. The principal reason is that ants will leave a kind of pheromone on the path by which they can contact each others. The longer the ants advance, the little the pheromone they release. The subsequent ants would choose the path by judging the amount of the pheromone. By such a feedback mechanism, interaction and pheromone, the ants eventually find an optimal path.

26.2.2 Algorithm Model

In the algorithm model, the artificial ants come from imitation of the real ones, and are endowed with some special function as follows [4, 5]:

- ① Will secrete pheromones in the process of moving, the amount of information depends inversely on the length of the path, that is, the longer the path is, the less pheromones will be;
- ② Selection of the route is based on sensing the amount of information.
- ③ Every ant has the same task to finding the optimal path.
- ④ Artificial ants can be endowed with functions that the real ants do not possess, such as: memory, local optimization, it is based on the demand, and can solve the problem more efficiently.

In the ant colony algorithm, each artificial ant selects the direction by the state transition rules, and complies with the updating principle of pheromone as follows:

- (1) The state transition probability

$$p_{ij}^k(t) = \begin{cases} \frac{[\tau_{ij}(t)]^\alpha * [\eta_{ik}(t)]^\beta}{\sum_{s \in allowed_k} [\tau_{is}(t)]^\alpha * [\eta_{is}(t)]^\beta}, & j \in allowed_k \\ 0, & otherwise \end{cases} \quad (26.1)$$

$$\eta_{ik}(t) = \frac{1}{d(i, k)} \tag{26.2}$$

Let $p_{ij}^k(t)$ be the probability that ant k in node i select j as the next node; let $allowed_k$ be the set of cities that can be chosen by ant k for the next step; let $\tau_{ij}(t)$ be the amount of information in path (i, j) at t time; $\eta_{ik}(t)$ is the heuristic function; let $d(i, k)$ be the distance between node i and the node k ; α is the information heuristic factor; β is the desired heuristic factor which describes the relatively importance of visibility.

(2) Pheromone updating rules

Too much residual pheromone may result in premature convergence which can seriously prevent the algorithm from searching the optimal solution. To avoid such problems, pheromone updating rules were injected.

The next is the introduction of pheromone update rule.

$$\tau_{ij}(t + n) = (1 - \rho) * \tau_{ij}(t) + \Delta\tau_{ij}(t) \tag{26.3}$$

$$\Delta\tau_{ij}(t) = \sum_{k=1}^m \Delta\tau_{ij}^k(t) \tag{26.4}$$

Let $\tau_{ij}(t + n)$ be the amount of information in path (i, j) ; ρ is the pheromone evaporation coefficient, the greater the value means the faster evaporation rate of pheromone; $\Delta\tau_{ij}(t)$ is the incremental of information in the loop through path (i, j) in this cycle.

26.3 Ant Colony Optimization Algorithm

26.3.1 Algorithm Idea

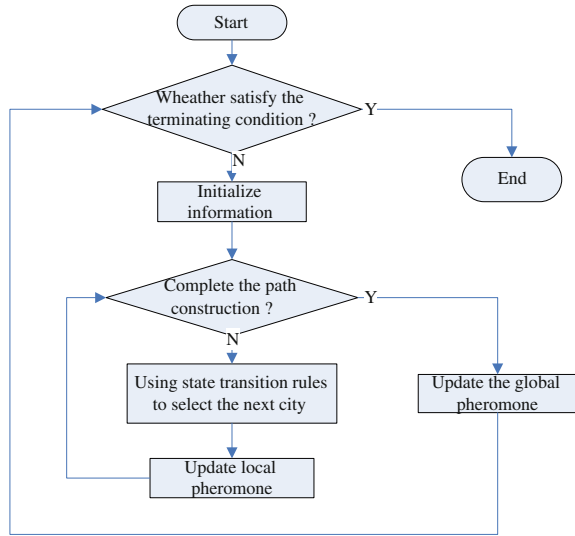
(1) The basic ant colony optimization updates the global pheromone after each completion of the global path. The improved algorithm retains the case of the global pheromone update, while the ants have a new ability to update the partial pheromone after selecting each city. This may enhance a single ant's influence on the pheromone. It is beneficial for the optimal solution, and effective to provide more information to find better path as well.

Mathematical model of the partial pheromone updating:

$$\tau(i, j) = (1 - \lambda) * \tau(i, j) + \lambda * \Delta\tau_{ij} \tag{26.5}$$

$$\Delta\tau_{ij} = \begin{cases} \frac{k}{d(i, j)}, & \forall (i, j) \in T \\ 0, & otherwise \end{cases} \tag{26.6}$$

Fig. 26.1 Flow chart of the improved algorithm



Let $\tau(i, j)$ be the amount of information between node i and node j ; λ is the pheromone evaporation coefficient, the greater the value means the faster evaporating of the volatile pheromone; K is a constant; $d(i, j)$ is the path length between the node i and node j ; $\forall(i, j) \in T$ means the truth that the ants in the Loop pass through the (i, j) path.

Mathematical model of the global pheromone updating:

$$\tau(i, j) = (1 - \lambda) * \tau(i, j) + \lambda * \Delta\tau_{ij} \tag{26.7}$$

$$\Delta\tau_{ij}^k = \Delta\tau_{ij}^{k-1} + Q / length_k \tag{26.8}$$

Where $\Delta\tau_{ij}$ is the incremental information in the path (i, j) in this loop; $d(i, j)$ is the path length between the node i and node j ; λ said the pheromone evaporation coefficient, the greater the value representative of the faster evaporation rate of pheromone; Q is a constant; $length_k$ is the total length of the path in the first k times cycle.

- (2) Use the pseudo-random rule to select the next city, rather than the direct use of transition probability.

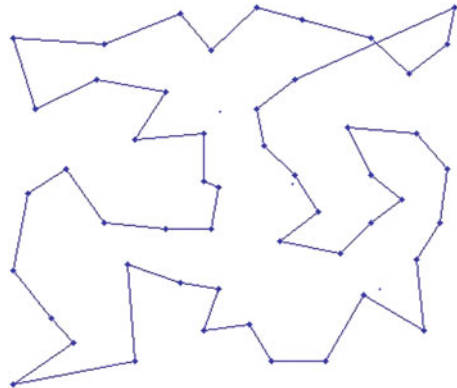
Pseudo-random rule: ① q_0 is a random number that uniformly distributed in $[0, 1]$; ② q_0 determines the size of the importance from the use of prior knowledge and explores new paths. The values can be obtained from experimental experience.

- (3) The improved algorithm process is shown in Fig. 26.1:

Table 26.1 Compare ACA with IACA

Algorithm	Cities	Ants	Iterations	The average length	Shortest path
ACA	30	16	1000	431.5	425
	41	29	500	377.9	374
	51	34	500	446.05	442
IACA	30	16	1000	427.9	422
	41	29	500	375.9	372
	51	34	500	435	429

Fig. 26.2 Best solution of ACA



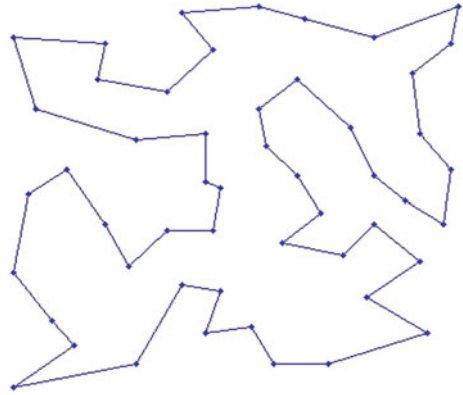
26.4 Simulation Analysis

The improved ant colony algorithm which is written by C++ has been proved to run successfully in VC6.0 under the 32-bit Win7 operating system with 2.0 GB memory. TSP test data was obtained from literature [6]. Experiment parameters are set as follows: traditional ant colony optimization: $\alpha = 1$, $\beta = 5$, $\rho = 0.5$, $Q = 100$; improved ant colony algorithm: $\alpha = 1$, $\beta = 4.7$, $\lambda = 0.4$, $q_0 = 0.8$.

Table 26.1 will show us the compare of the ability to find the path of TSP between the ant colony algorithm (ACA) and the improved ant colony algorithm (IACA).

We get the results of Table 26.1 through several simulation tests. The average length in the table was average of the value obtained by each algorithm’s test. The shortest path refers to the best results from the multiple tests in each algorithm. The tests show that the quality to get best solutions was improved, not only the single solution but also the overall one. The improved ant colony optimization has met the expected requires (Figs. 26.2 and 26.3).

Fig. 26.3 Best solution of IACA



26.5 Conclusion

Since the ant colony algorithm for the TSP was firstly proposed, many scholars have done a lot of researches on it in order to extend it to more optimization areas, including job scheduling problem, network routing, vehicle logistics and transportation problems, air combat decision-making, layout optimization applications. Currently, the fusion model remains to be improved, and the application needs further expansion. Applications of the ant colony, both in depth and breadth have a lot to be improved [4]. Improving the value of application integrated with superior performance of ant colony algorithm should be an important direction for future research.

Acknowledgments This research is supported by Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20103326120001), Zhejiang Provincial Natural Science Foundation of China (No. Z1091224, Y7100673 and Y1091164), Zhejiang Provincial Social Science Foundation of China (Grant No. 10JDSM03YB), the Scientific Research Fund of Zhejiang Province (No. 2010C11062), Research Project of Department of Education of Zhejiang Province (No. Y200907458 and Y201016434), the Contemporary Business and Trade Research Center of Zhejiang Gongshang University (No. 1130KUSM09013 and 1130KU110021). We also gratefully acknowledge the support of Science and Technology Innovative project (No. 1130XJ1710214).

References

1. Dorigo M, Maniezzo V et al (1991) Distributed optimization by ant colonies. In: Proceedings of the 1st European conference on artificial life, pp 134–142
2. Dcolorni A (1994) Ant system for job shop scheduling [J]. *Belgain J Oper Res Stat Comput Sci* 34(1):39–53
3. Dorigo M, Gam Bardella LM (1997) Ant colony system: A cooperative learning approaches to the traveling salesman problem [J]. *IEEE Trans Evol Comput* 1(1):53–66

4. Haibin Duan (2005) Ant colony algorithm and its applications. Science Press, Beijing
5. Dorigo M, Maniezzo V, Colomi A (1996) Ant System: optimization by a colony of cooperating agents. IEEE Trans Syst Man Cybern B 26(1):29–41
6. Shiyong Li (2004) Ant colony algorithms with applications. Harbin Institute of Technology Press, Harbin

Part III
Networks Systems and Web Technologies

Chapter 27

Research on Internet of Things Technology

Qinian Zhou, Lingling Chen, Ge Li and Zhenhao Zhang

Abstract At first, the concept and development of Internet of Things were introduced. And then the Internet of Things architecture was described. The key technologies of Internet of Things were analyzed, including RFID technology, sensor network technology, intelligent technology, nanotechnology and cloud computing technology. At the same time, some problems in the development of Internet of Things were presented. Finally, this paper gave the future prospects of Internet of Things Technology.

Keywords Internet of Things · RFID · Sensor · Cloud Computing

The construction of modern agriculture industry technology system special funds support (09030313-C).

27.1 Introduction

Because the concept of Internet of Things is emerging soon, its connotation is still in the development and perfection. At present, there is no agreement of the concept of it, but it is sure that Internet of Things is an another information industry wave

Q. Zhou · L. Chen (✉) · Z. Zhang
College of Informatics and Electronics, Zhejiang Sci-Tech University,
Hangzhou China
e-mail: chenling117@126.com

G. Li
College of Machinery and Automatic Control, Zhejiang Sci-Tech University,
Hangzhou 310018, China

after computer, internet and mobile communication. Different definitions were raised by the researchers whose starting and key points are different in different fields. The comparative accurate definition of it is that Internet of Things is a information network in which the things and things, people and things, people and people are linked together to realize intelligent identification, location, tracking, monitoring and management with the help of kinds of information sensing equipments and systems (sensor network, RFID systems, infrared sensors, laser scanners, etc.), barcode and two dimensional barcode, global positioning system [1, 2].

In 1988, Massachusetts Institute Technology raised the ideas of Internet of Things which was named as Electronic Product Code system in that time creatively. In 1999, America Auto-ID put the concept of Internet of Things forward firstly, however, the definition of it was very simple, it was that all the things were linked together via information sensing equipments such as RFID, etc. and internet to realize intelligent identification and management. Technology review of American regarded the sensor network technology as the first technology of ten that would change the life of people in the future in 2003 [3]. In 2005, the International Telecommunication Union (ITU) proposed the concept of Internet of Things officially in ITU Network Report 2005: Internet of Things, and at the same time, the report also pointed out that the current communication technologies were limited in the target of any time, any place and any one, and in the future it would be in the stage of anything, and then Internet of Things existed. In March 2006, European Union held a meeting named From RFID to the Internet of Things, and a deeper description of it was made in the meeting [4]. In March 2008, the first global and international Internet of Things meeting was held in Zurich, the new ideas and technologies of Internet of Things and how to boost the development of Internet of Things were discussed in the meeting. The CEO of IBM Sam Palmisano raised the concept of “Smart Planet” firstly in “Round-table conference” which was held by Obama and American business leaders. On August 2009, prime minister Wen Jiabao went to wuxi micro-nanometer sensor network engineering technology research center inspection and delivered an important speech, he raised the idea of “Perception China” firstly. On August 24, China mobile president Wang Jianzhou delivered his first public speech in Taiwan and Proposed the concept of Internet of Things. On March 2010, the Shanghai Internet of Things center was established formally. There were some explanations about Internet of Things in the annotation of The Government Work Report 2010, it was that Internet of Things was a network which put anything and internet together and they can change information to realize intelligent identification, location, tracking, monitoring and management via information sensing equipments according to the agreed protocol [5]. In 2005, one ITU report described the prospect in the era of Internet of Things, it is that the car will alarm automatically when the driver make a mistake; the briefcase could remind the master that what forgot to take; the clothes will tell the washing machine their requirements of color and temperature etc. It is no doubt that if the era of Internet of Things comes, people’s daily life will change strikingly [6].

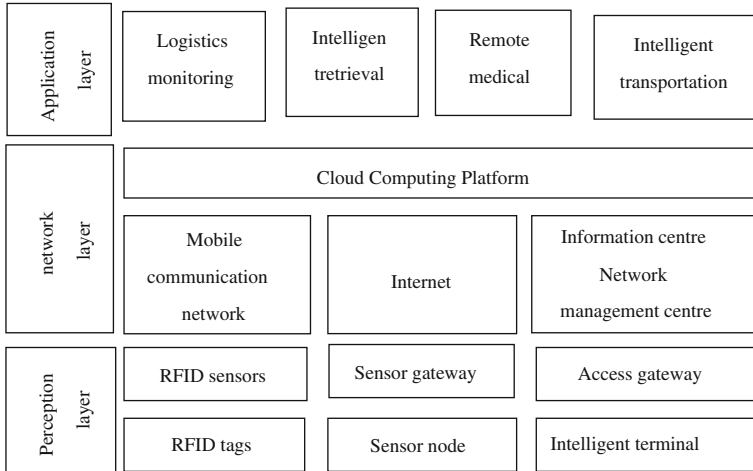


Fig. 27.1 Internet of things architecture

27.2 Internet of Things Architecture

The Internet of Things needs an open architecture to maximize interoperability among heterogeneous systems, providers and consumers of information and services. The architecture should have well-defined and granular layers because of the decentralized and heterogeneous nature of things. At present, there is no agreement of Internet of Things architecture because it is in the initial stage, but it can be divided into three levels according to the three characteristics that it should have. They are comprehensive perception, reliable transfer and intelligent processing, accordingly, the three levels of Internet of Things are shown in the Fig. 27.1 below.

As is shown in the Fig. 27.1 above, the main function of perception layer is perceiving and gathering information, it gathers information through RFID, sensor and two dimensional barcode technology etc. It also includes the sensor network of the data access to the gateway. The network layer is based on existing mobile communication network and internet. It assumes the data communication task between perception layer and network layer. Cloud Computing platform is mass data storage and analysis platform and is an important component of the network layer, it is also the foundation of many applications in application layer. The network layer is composed by many application servers, it provides rich and specific services for users by making use of analyzed and handled perception data, such as monitoring, query, control and scanning service etc.

27.3 The Key Technologies in the Internet of Things

There are many advance technologies included in the Internet of Things system. Some key technologies are introduced below. That are RFID technology, sensor network technology, embedded intelligent technology, nanotechnology and Cloud Computing technology.

27.3.1 *RFID Technology*

RFID is a non-contact automatic identification technology, it uses the frequency radio and transmission properties of space coupling to realize automatic identification of static or moving objects. Manual intervention is unnecessary in the process of identification. It can work in various harsh environment, identify high-speed objects and multiple tags at the same time, it is very convenient [7]. RFID is the main source of gathering information in the Internet of Things, so it plays a very important role in the Internet of Things system.

RFID can let objects talk. In the Internet of Things, standardized and interoperable information are stored in RFID tags, they are gathered in the center information system by wireless data communication network to realize the identification, then it can realize information exchange and sharing through open computer network, and it manages objects transparently finally [8].

The components of RFID system will be different because of the different applications, but it usually includes electronic tag, reader and antenna. Electronic tag is composed by coupling components and chips, there is the only Electronic Product Code (EPC) in each electronic tag. EPC is unique and retrievable identification code for each object and it is also unique ID. The reader is used to read label information. They are handed style and stationary style. Antenna can transmit signal between electronic tag and reader.

The working principle of RFID is that read transmit specific frequency radio through antenna, the induction current produces when the electronic tag enters into valid work area, and the electronic tag is activated, it transmit its coding information via internal antenna; the receiving antenna of reader receives the modulation signal from electronic tag, and the signal is sent to signal processing module by antenna modulator, effective information generated after harmonic and decoding and then be sent to back host system; the host system identifies the tag's identity according to the logic operation, makes corresponding treatment and control according to different set, and sends signal to control the reader's work finally [1].

The biggest advantage of RFID system is that it can work without manual intervention. It can be used in tracking and identifying objects, human and animals. It is still in the stage of development.

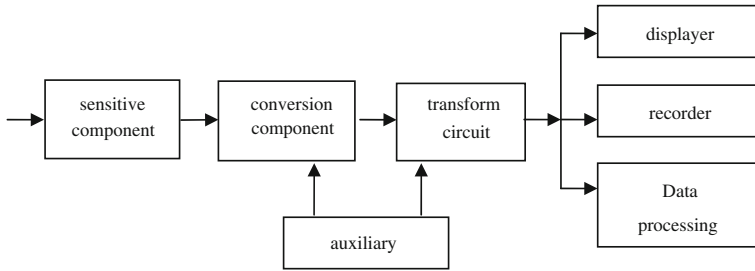


Fig. 27.2 The basic composition of the sensor

27.3.2 Sensor Network Technology

Sensor is a kind of equipment that can change specific measured information into some usable output signal by a certain rule, in order to meet the demand of information transmission, processing, recording, display and control [1]. With the development of science and technology, sensor network are developing toward miniaturization, integration, intelligence, and network. The development history is traditional sensor, intelligent sensor and embedded web sensor.

The basic components of sensor network node are sensing unit (composed of the sensor and A/D conversion function module), processing unit (includes CPU, memory and embedded operating system etc.), communication unit (composed of wireless communication modules) and power. At present, the main content of sensor network technology in the field of Internet of Things are followings.

- Advanced testing technologies and network measurement and control.
- Research on intelligent sensor network node.
- Research on sensor network organization structure and the underlying protocols.
- The test and control of sensor network.
- Security of sensor network.

A typical sensor usually consists of sensitive components, conversion components and transform circuit, and including auxiliary power sometimes. As is shown in the Fig. 27.2.

The sensitive component feels the measured signal directly, and output a physical signal that has definite relationship with the measured signal. Conversion component takes sensitive component output as input, it changes non-electric quantity into electrical signal and then outputs, it is the core component of sensor. Transform circuit inserts front circuit parameters into itself, and it will be changed into electric quantity and then output. Sensor can be used in many fields. For example, space satellites, rocket, military aircraft, buildings and underwater submarine etc. The development of sensor technology makes great contribution to the progress of science and technology [8].

27.3.3 Embedded Intelligent Technology

Intelligent technology is adopted various methods and means to reach an expected purpose. The object is intelligence when the intelligent system is inserted in, it can realize the communication between objects and users actively or passively [9]. In the present technical level, embedded technology is the key to realize intelligent technology, and intelligent system mainly consists of one or more embedded systems. Embedded system is a system in which application procedure, operating system and computer hardware are integrated together.

At present, there are also some technical difficulties that need further research, it mainly includes the following respects.

The theory research of artificial intelligence.

Advanced man-machine interactive technology and system.

Intelligent control technology and system.

Intelligent signal processing.

27.3.4 Nanotechnology

Nanotechnology is a science that study movement rules and interaction of 0.1–100 nm composition system and the technical problems in practical applications, it mainly includes nanophysics, nanometer chemical, nanometer material science, nanometer biology, nanometer electronics, nanometer processing learn, nanometer mechanics and so on [9, 10].

At present, the main applications of nanotechnology in the Internet of Things are miniaturization design of RFID equipment, sensing equipment and the processing technology of processing materials and micro-nanometer. The development of nanotechnology provides excellent sensitive material for sensor and many new methods for sensor production. Nanometer sensor is widely used in biology, chemistry, machinery, aviation and military. These applications signify that smaller objects in volume can interact and connect in the Internet of Things, what is more, the development of nanotechnology will promote electronic devices and systems to miniaturization, system integration, faster response speed and smaller consumption of a single device.

27.3.5 Cloud Computing Technology

As the internet technology develops, the generated data volume will be far exceed the internet data volume, therefore, cloud computing, pattern recognition and other intelligent computing technologies must be used to reach the intelligent processing, analyze and handle the mass data, and realize intelligent control finally.

Cloud Computing is composed of a series of dynamically upgraded and virtualization resources, these resources are shared by all users and they are also can be visited through the network. At the same time, users can access computing power, storage space and information services on-demand. In the Cloud Computing, the data of users is not stored in local, but stored in the clouds, the required application is not running in personal computers, but running in the clouds. Users can connect to the internet through any terminal to visit the data and use the application software in the clouds.

Cloud Computing has the following characteristics at the point of the current research status.

- Big scale.
- virtualization.
- High reliability and generality.
- High expansibility
- On-demand service.
- Extremely cheap.

There are four key technical points to realize the combination of Cloud Computing and Internet of Things [9]. First, the expansion of internet technology in the Internet of Things, including the realization of IP computing in kinds of objects, wireless access network, the fusion of internet and information. Second, IP virtualization technology, the resources sharing and IT service ability on-demand provided can be truly realized after the IP virtualization. Third, how to manage, control the Internet of Things and Cloud Computing platform and then make it more reliable and safe is also important. Finally, kinds of businesses and applications based on Cloud Computing and Internet of Things. An effective, benign value chain system and business ecosystem will be formed when Internet and Things and Cloud Computing provide service for human through the appropriate business model and practical actual service. So as to promote the entire information sector, IT and all walks of life to sustainable development.

27.4 Development Trends and Prospects

In the internet of Things, “things” are expected to become active participants in all kind of fields where they are enabled to interact and communicate among themselves and with the environment. In order to achieve the target, they exchange data and information and “sense” about the environment. At present, many specific applications of Internet of Things can be seen in the life, though it is in the primary stage, for example, intelligent library and remote anti-theft system etc. But many problems exist in the rapid development of Internet of Things. First, the product price, electronic tags and many read equipments and large information processing systems must be inserted into all the objects, it will cause higher costs. So how to reduce the price is a problem. Second, the product standardization, standards

should be designed to support a wide range of users and address common requirements from a wide range of industry sectors as well as the needs of environment, society and individual users. there is no a complete set of international standard in the field of RFID and WSN etc., the interoperability does not exist in the equipments of different manufactures, it is not convenient for users. Third, IP problems, IPv4 addresses will be exhausted at that moment, and IPv6 is needed. The transition from IPv4 to IPv6 is a long process, so the compatibility problem must be solved. Finally, security problem, in the Internet of Things, the relationship of things and things will be closer, things and people even be joined up, therefore, how to protect the large amount of data and privacy of users is needed to be resolved quickly.

Though there are many problems in the development of Internet of Things, it is sure that greater development of Internet of Things technology exists in the future. Application form is flexible, it will be used in every aspect of life. It is no doubt that we will benefit and it will accelerate the economic development in the age of Internet of Things.

References

1. Huajun L, Chuanqing L, Xiulin X (2010) Internet of things technology. Electronic Industry Press, Beijing
2. Amardeo C, Sarma JG (2009) Identities in the future Internet of Things. *Wireless Pers Commun*, Article 49 period(2009) pp 353–363
3. Xiangdong H (2010) Research on internet of things development. *Digit Commun* 17–21
4. Buckley J (2006) From RFID to the internet of things 22 pervasive networked systems. European Commission, DG Information Society and Media, Networks and Communication Technologies Directorate, Brussels
5. Jiabao W (2010) Government work report. http://www.gov.cn/2010lh/content_1555767.htm (2010)
6. BaiKe (2011) Internet of things development. <http://baike.baidu.com/view/1136308.htm>
7. Huajun L (2010) Research on key technology for internet of things. *Computer Age*, Article 7 period, pp 4–6
8. Runian L (2009) Study on the internet of things based on RFID technique. *J China Electron Sci Res Inst*, Article 6 period, Dec, pp 594–597
9. Yang G, Shen P, Zheng C (2010) Internet of things theory and technology. Science Press, Beijing
10. Feng Z (2010) The sensor technology based on internet of things. *Solutions*, Article 1 period, pp 28–31

Chapter 28

Hot Topic Detection Research of Internet Public Opinion Based on Affinity Propagation Clustering

Hong Liu and Bi Wei Li

Abstract Internet is becoming a spreading platform for public opinion. It is very important to grasp hotspot of internet public opinion (IPO) in time and understand the trends of them correctly. Aim at such drawbacks of some text clustering algorithm as information massive, curse of dimensionality, longer analysis time, lower analysis efficiency, this chapter introduces affinity propagation algorithm into hot topic detection of IPO, with the center of each clustering representing a topic. Through compared result of experiment affinity propagation clustering and K-Means algorithm, it shows that the efficiency and effectiveness of such the algorithm.

Keywords Hot topic detection · Internet public opinion · Affinity propagation · K-Means · Similarity matrix · Vector space model

28.1 Introduction

Nowadays, Internet transmission has become a new broadcasting format and played a more and more important role in public opinion, which refers to the society and politics attitude toward the social administrator in certain social space [1]. At present, a large number of current webs are full of the content of eroticism,

H. Liu (✉) · B. W. Li
College of Computer and Information Engineering, Zhejiang Gongshang University
HangZhou China
e-mail: LLH@mail.hzic.edu.cn

B. W. Li
e-mail: lbw@mail.hzic.edu.cn

reactionary/cult, and cybercrime. On the other hand, the occurrence of public emergency easily leads to grapevine popular on the network, which will cause public irrational judgments and disturbing behaviors. So it is especially important to analyze and detect topic information of internet public opinion (IPO for abbr.), make useful information quickly exposed to those seekers, and real-timely monitor the tendency of public opinion.

However, there are several challenges in analyzing and monitoring IPO. As we know, information on the internet is extremely disordered and constantly changing, so it technically is difficult to obtain a large number of relevant web pages rapidly, to judge the relevant degree of two text sections, and to process the public opinion change quickly. Other than static WebPages, unstructured or loosely formatted texts often appears at a variety of tangible or intangible dynamic interacting networks [2]. A web page is different from regular corpora of text documents.

In this chapter, aim at such drawbacks of some text clustering algorithm as k-centers clustering results are quite sensitive to the initial selection of exemplars, so it is usually rerun many times with different initializations in an attempt to find a good solution. This works work well only when the number of clusters is small and chances are good that at least one random initialization is close to a good solution. We introduce affinity propagation clustering in hot topic detection of IPO. The remainder of this chapter is organized as follows: [Section 28.2](#) introduces related works; [Sect. 28.3](#) introduces affinity propagation clustering. [Section 28.4](#) details our methodology application in hot topic detection of IPO; [Sect. 28.5](#) shows our experimental results and discussion. Finally, [Sect. 28.6](#) concludes the chapter and gives future work.

28.2 Related Work

Currently, people have also done some research about IPO, whose research methods are mainly clustering algorithm. Raymond W.M. Yuen and Terence Y.W. Chan [3] proposed a text opinion analysis approach based on morphemes. Kim Soo-Min and Hovy Eduard [4] presented an opinion analysis system, which could automatically extract opinions and opinion holders. Kim Soo-Min and Hovy Eduard [5] described a sentence-level opinion analysis system. Kim Soo-Min and Hovy Eduard [6] presented a novel method based on Semantic Role Labeling to extract opinions, opinion holders and opinion topics in text. Lun-Wei Ku, Yu-Ting Liang and Hsin-Hsi Chen [7] proposed an opinion identification algorithm at word, sentence and document level. Namrata Godbole, Manjunath Srinivasaiah and Steven Skiena [8] described a system that assigned scores to indicate positive or negative opinion for each distinct entity in the text corpus. Devitt and K. Ahmad [9] aimed to explore a computable metric of positive or negative polarity in financial news text for the financial markets. Dongjoo Lee, Ok-Ran Jeong and Sang-goo Lee [10] presented an opinion mining model especially for customer

feedback data about products and services. Selver Softic and Michael Hausenblas [11] gave a novel approach to opinion mining, which firstly RDFised discussion forums in SIOC, and then interlinked the created data with linked datasets such as DBpedia. Lan You, Yongping Du and et al. [12] utilized Back-Propagation Neural Network based classification algorithm to evaluate the hotness of the topic through its popularity, quality and message distribution. Tingting He, Guozhong Qu and et al. [13] presented a semi-automatic hot event detection approach. K. Chen and L.Luesukprasert [14] extracted hot terms by mapping their distribution over time, and identified key sentences through hot terms, then used multidimensional sentence vectors to group key sentences into clusters that represented hot topics. M. Platakis, D. Kotsakos and D. Gunopulos [15] proposed a hot topic detection method during a time interval which was based on bursty discovery. Qindong Sun, Qian Wang and Hongli Qiao [16], aiming at the mobile short message hot topic extraction, identified the hot topic through feature words and the association degree of these feature words. Yadong Zhou, Xiaohong Guan and et al. [17] utilized statistics and correlation of popular words in network traffic content to extract popular topics on the Internet.

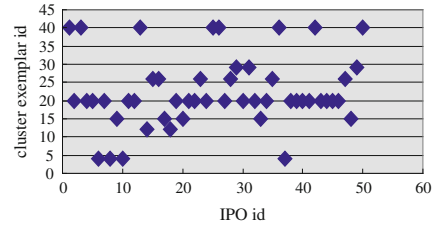
28.3 Affinity Propagation

Affinity propagation (AP) is a powerful clustering algorithm based on message passing technique was published in Science by Frey and Dueck [18–20]. Such method is a very simple but very efficient method for clustering data. It cannot only cluster data points into different classes, but also detect a representative sample for each class. It first simultaneously considers all data points as potential representative examples and, considers each data point as a node in a network, then recursively transmits real-valued messages along edges of the network until a set of representative examples and corresponding clusters emerges. Such AP clustering method has fast processing speed and can avoid many of the poor solutions which caused by unlucky initializations and hard decisions.

AP is based on similarities $s(i,k)$ (or $n \times n$ similarity matrix S for n data points) between pairs of data points, and simultaneously considers all the data points as potential cluster centers (called exemplars). To find appropriate exemplars, two kinds of messages (responsibility, availability) passed between data points, each of which each applies a different kind of competition, their illustrations are shown in Fig. 28.1. The responsibility $r(i,k)$ being sent from data point x_i to candidate representative example data point x_k , reflects how proper it would be for point x_k to serve as the representative example for point x_i , taking into account other potential exemplars for point x_i . It is updated using the rule:

$$r(i,k) = s(i,k) - \max_{j \neq k} [a(j,i) + s(i,j)] \quad (1)$$

Fig. 28.1 Clustering result of 50 data points



It is set to the similarity between point x_i and point x_k , minus the maximum of the similarities between point x_i and other candidate representative examples.

The availability $a(i,k)$, sent from candidate representative example point x_k to point x_i , reflects the accumulated evidence for how well-suited point x_i is to choose point x_k as its representative example, taking into account the support from other points that point x_k should be an exemplar. It is computed by the rule:

$$a(k, i) = \min \left\{ 0, r(k, k) + \sum_{j: j \notin \{k, i\}} \max[0, r(j, k)] \right\} \quad (2)$$

The availability $a(i,k)$ is set to the self-responsibility $r(i,k)$ plus the sum of the positive responsibilities candidate representative example x_k receives from other points. Unlike the responsibility update which lets all candidate representative examples fight for the ownership of a data point, the availability update gathers evidence from data points as to find whether each candidate representative example is suitable for acting as a good representative example.

Availabilities and responsibilities can be combined to recognize representative examples at any time during AP. For point x_i , the that maximizes $a(i,k) + r(i,k)$ indicates that point x_k serves as the representative example for point x_i . The message passing procedure is terminated after the local decisions stay constant for some number of iterations.

28.4 Hot Topic Detection of IPO Based on Ap Clustering

Application AP clustering in hot topic detection of IPO is mainly composed of following steps: data pretreatment, similarity matrix representation, AP clustering for IPO and results evaluation.

28.4.1 Data Pretreatment

IPO accomplished form search engine robot process has unstructured format, before analyzing, data pretreatment is applied for them, which includes data cleaning, text segmentation. For data cleaning, we remove noise data and

irrelevant data. Noise data include navigation bar, advertisement information, copyright information and investigate questionnaire, etc., which has no relation with topic content. Irrelevant data are from IPO where there are not enough postings or posting contents that are not related to the opinion topics at all. After removing noisy data and outliers, we apply Chinese text segmentation system (ICTCLAS system) of Chinese science research institute to do text segmentation for topic content of web pages, in the chapter. ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) is a Chinese lexical analysis system using an approach based on multi-layer HMM. ICTCLAS includes word segmentation, Part-Of-Speech tagging and unknown words recognition. After text segmentation, we remove stop-words, which are high frequent words that carry no information (i.e. pronouns, prepositions, conjunctions etc.), and do word stemming work, which means the process of suffix removal to generate word stems. This is done to group words that have the same conceptual meaning, in the analysis process; these group words define a feature. And then we need transform unstructured text format of web pages into structured text format for text representation, which typically are strings of characters into a suitable representation for the clustering task.

28.4.2 Similarity Matrix Representation

After data pretreatment, we need translate IPO into similarity matrix. Firstly, with TF-IDF (Term Frequency-Inverse Document Frequency) method, we transfer IPO into vectors space. Calculate formula is shown as follows:

$$w_{ij} = tf_{ij} \times idf_i \quad (3)$$

Where, w_{ij} stands for weight of feature word i in IPO j . tf_{ij} stands for frequency of feature word i in IPO j . idf_i stands for reciprocal of IPO frequency of feature word i , which is shown as follow:

$$IDF(t) = \log \left[\frac{D}{DF(t)} \right] \quad (4)$$

To strengthen the importance of one keyword in critical location, we enhance the weight of keywords that appear in head, title, and first-level title of that IPO. Then, every IPO d can be converted to vector space that is comprised of feature vectors, which is shown as $d_i = \{w_{i1}, w_{i2}, \dots, w_{in}\}$. Lastly, we calculate the similarity degree between every two vectors in the vector space to get similarity

matrix. We adopt cosine formula to calculate similarity degree between every two IPO based on vector space model, which is shown as follows:

$$Sim(d^i, d^j) = \cos(T^i, T^j) = \frac{\sum_{k=1}^n w_k^i * w_k^j}{\sqrt{\sum_{k=1}^n (w_k^i)^2} \sqrt{\sum_{k=1}^n (w_k^j)^2}} \quad (5)$$

Where, $Sim(d^i, d^j)$ stands for similarity degree between IPO d^i and d^j , w_k^i stands for weight of word k in IPO i , n stands for total amount of feature words in the vector space.

28.4.3 AP Clustering for IPO

The AP process is as follows:

- (1) Input: s, p, λ .
- (2) Initialization: $r(i, k) = 0, a(i, k) = 0$ for all i, k
- (3) Responsibility updates using Eq. (1) and following rules:

$$r^{old}(i, k) = r(i, k) \quad (6)$$

$$r^{new}(i, k) = (1 - \lambda)r(i, k) + \lambda r^{old}(i, k) \quad (7)$$

- (4) Availability updates using Eq. (2) and following rules:

$$a^{old}(k, i) = a(k, i) \quad (8)$$

$$a(k, k) = \sum_{j:j \neq k} \max[0, r(j, k)] \quad (9)$$

$$a^{new}(k, i) = (1 - \lambda)a(k, i) + \lambda a^{old}(k, i) \quad (10)$$

- (5) Iterative from step (3) and step (4) until a high-quality set of exemplars and corresponding clusters emerges. Then go to step (6).
- (6) Making assignments:

$$c_i^* = \arg \max_k [r(i, k) + a(k, i)] \quad (11)$$

28.4.4 Evaluation Criterion

In this chapter, we use such evaluation criterion as Macro_Rrecision, Macro_Recall, Macro_Fscore and average unmiss probability which are defined as follows:

$$\begin{aligned} Macro_F1 &= \frac{2 \times Macro_R \times Macro_P}{Macro_R + Macro_P}, \text{ where } Macro_R = \frac{1}{n} \sum_{i=1}^n R_i, Macro_P \\ &= \frac{1}{n} \sum_{i=1}^n P_i \end{aligned} \quad (12)$$

$$avgUnMiss = \sum_{i=1}^n \left(1 - \frac{\text{un detected opinion about topic } i}{\text{total opinion about topic } i}\right) \quad n \text{ is clustering number} \quad (13)$$

Where, Precision $P = TP/(TP + FP)$, Recall $R = TP/(TP + FN)$, and $F1 = 2 \cdot P \cdot R / (P + R)$. Where, TP(true positive) means number of correct positive predictions, FP(false positive) means number of document related to the category incorrectly, FN(false negative) means number of document not marked as related to a category but should be, and $TP + FP$ means number of positive predictions, $TP + FN$ means number of positive examples.

28.5 Experiment and Discuss

In the experiment phase, we collect data from some web sites as: news.sohu.com, www.china.com, www.zaobao.com, unn.people.com.cn and category of data include finance and economics, humanistic, life, entertainment, etc. We do experiment using AP algorithm, firstly we convert 50 pieces of IPO into similarity matrix. When applying AP clustering, max iterations sets 2000, stop criterion sets 200, damping factor set 0.5, value set 0.8 and preferences uses median similarity. The clustering result sketch map of IPO id is shown in Fig. 28.1. The sum of data point-to-exemplar similarities is 2.563664, and the sum of exemplar preferences is 7.4324. What's more, we compare AP clustering and K-Means in 50 data points with the same clustering value using Macro_P, Macro_R, and Macro_F1, whose compared result is shown in Fig. 28.2.

And then, we perform AP clustering and K-Means algorithm with data set number from 100 to 500. The avgUnMiss compared result is shown as Table 28.1.

From experiment, we can see that when at the same dimension of vector space, AP clustering has higher efficiency than K-Means algorithm.

Fig. 28.2 Comparison result of 50 data points

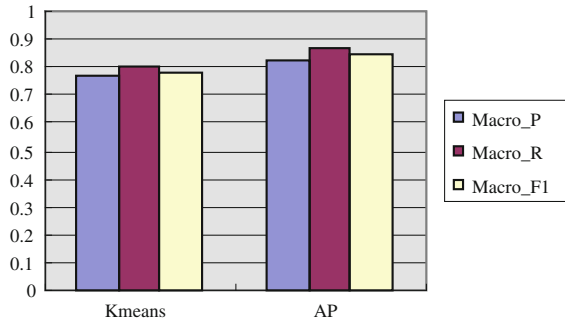


Table 28.1 Compare experiment result

Number of data points	Number of feature	avgUnMiss of AP	avgUnMiss of KMeans
100	1,069	0.875	0.782
200	1,854	0.825	0.799
300	2,462	0.889	0.804
400	2,859	0.901	0.854
500	3,249	0.946	0.898

28.6 Conclusions

In this chapter, according to the properties of IPO, we introduce vector space model to express word feature, adopt cosine formula to express similarity of IPO, and we perform AP clustering and K-Means algorithm in hot topic detection of IPO. The experimental result shows the efficiency and effectiveness of AP clustering algorithm. What’s more, our future works are shown as follows:

(1) Commerce application research is one of our future works. Using our IPO analysis approaches can help seeker get quickly information what they want to get. For market department, it can help them understand what their specific customers’ concerns regarding goods and services information. It is beneficial for us to provide best personalized services for people,

(2) Deeply approach research of IPO hot topic detection and forecast. Refinement for each step of the approach proposed above is needed. Dynamic monitoring technology is in demand which can monitor the web sites to detect change in time. Data cleaning is time-consuming and labor-intensive. Web content analysis can not stop at word frequency analysis because sometimes the result is poly semantic. How to improve process speed of massive data is also main research content of our future works.

Acknowledgment This chapter is supported by the Natural Science Foundation project of ZheJiang provincial (No. Y1110995). And this chapter is also supported the National Natural Science Foundation of China under Grant No. 60903053.

References

1. Wang LH, Liu Y (2005) An overview of China 2004 public sentiment research. *Xinhua Digest* 18:133–134
2. Khan JI, Shaikh S (2006) Relationship algebra for computing in social networks and social network based applications. In: *IEEE/WIC/ACM international conference on web intelligence*, pp 113–116
3. Yuen RWM, Chan TYW (2004) Morpheme-based derivation of bipolar semantic orientation of Chinese words. In: *Proceedings of the 20th international conference on computational linguistics (COLING-2004)[C]*, Geneva, Switzerland, pp 1008–1014
4. Kim S, Hovy E (2004) Determining the sentiment of opinions. In: *Proceedings of the 20th international conference on computational linguistics*, pp 1367–1373
5. Kim S, Hovy E (2005) Automatic detection of opinion bearing words and sentences. In: *The companion volume of the proceedings of IJCNLP-05*, Jeju Island, Republic of Korea
6. Kim S, Hovy E (2006) Extracting opinions, opinion holders, and topics expressed in online news media text. In: *Proceedings of the ACL workshop on sentiment and subjectivity in text*, Sydney, Australia, pp 1–8
7. Ku L-W, Liang Y-T, Chen H-H (2006) Opinion extraction, summarization and tracking in news and blog Corpora. In: *Proceedings of AAAI-2006 spring symposium on computational approaches to analyzing weblogs*, pp 100–107
8. Godbole N, Srinivasaiah M, Skiena S (2007) Large-scale sentiment analysis for news and blogs. In: *International conference on weblogs and social media (ICWSM'2007)*, Boulder, CO, pp 219–222
9. Devitt A, Ahmad K (2007) Sentiment polarity identification in financial news: a cohesion-based approach. In: *Proceedings of ACL2007*
10. Lee D, Jeong OR, Lee S (2008) Opinion mining of customer feedback data on the web. In: *Proceedings of the 2nd international conference on ubiquitous information management and communication*, Suwon, Korea, pp 230–235
11. Softic S, Hausenblas M (2008) Towards opinion mining through tracing discussions on the web. In: *Social data on the web (SDoW 2008) workshop at the 7th international semantic web conference*, Karlsruhe, Germany
12. You L, Du Y, Ge J, Huang X, Wu L (2005) BBS based hot topic retrieval using back-propagation neural network. In: *Proceedings of 1st international joint conference on natural language processing*, China, Springer, pp 139–148
13. He T, Qu G, Li S, et al (2006) Semi-automatic hot event detection. In: *Proceedings of the 2nd international conference on advanced data mining and applications*, LNAI 4093, pp 1008–1016
14. Chen K, Luesukprasert L (2007) Hot topic extraction based on timeline analysis and multidimensional sentence modeling. *IEEE Trans Knowl Data Eng* 19:1016–1025
15. Platakis M, Kotsakos D, Gunopulos D (2008) Discovering Hot Topics in the Blogosphere. In: *Proceedings of the 2nd Panhellenic scientific student conference on informatics, Related technologies and applications EUREKA*, pp 122–132
16. Sun Q, Wang Q, Qiao H (2009) The algorithm of short message hot topic detection based on feature association. *Inf Technol J* 8:236–240
17. Zhou Y, Guan X et al (2009) Approach to extracting hot topics based on network traffic content. *Frontiers Electr Electron Eng* 4:20–23
18. Frey BJ, Dueck D (2005) Mixture modeling by affinity propagation, *NIPS*
19. Leone M, Sumedha, Weigt M (2007) Clustering by soft-constraint affinity propagation: applications to gene-expression data. *Bioinformatics* 23(20):2708–2715
20. Frey J, Dueck D (2007) Clustering by passing messages between data points. *Science* 315(5814):972–976

Chapter 29

TTMP: A Trust-Based Topology Management Protocol for Unstructured P2P Systems

Zhiping Liao, Song Liu, Gelan Yang and Jiancun Zhou

Abstract Uncooperative behaviors of peers are great threats to current unstructured peer-to-peer (P2P) systems. To solve this problem, we present a trust-based topology management protocol (TTMP), which aims to promote the fairness and service quality of P2P system by integrating a trust model into its topology management. To achieve this goal, TTMP divides a single connection between every pair of neighboring peers in current networks into two links: an In-Link and an Out-Link. A peer can only get services by its In-Links and provide services through its Out-Links. And TTMP manages these two types of Links based on two different trust metrics defined in the proposed trust model. Simulation results indicate that TTMP improves the performance of P2P system.

Keywords p2p · Topology · Reputation · Trust · Uncooperative behavior

29.1 Introduction

Unstructured peer-to-peer (P2P) networks are highly decentralized systems where each peer represents a different self-interested entity. A peer can manipulate its local information to take advantage of other peers' resources. Therefore, the performance of a P2P system depends heavily on the cooperation degree of peers.

Z. Liao (✉) · G. Yang · J. Zhou
Department of Computer Science, Hunan City University
Yiyang 413000, People's Republic of China

S. Liu
Department of Economy Management, Hunan City University
Yiyang 413000, People's Republic of China

However, for the sake of inherent rationality which urges peers to struggle for their maximum benefit, there always exist many uncooperative behaviors which threat the utility of P2P system at any given time [1, 2]. For example, to save its communication bandwidth and energy a free-rider may consume services without contributing its own fair share. And in some popular P2P systems, there also exist peers who provide a great deal of fake and even polluted resources or services. Large amount of these uncooperative behaviors can degrade dramatically the quality of P2P services [3]. Hence, it is important to design a trust model to motivate peers to behave cooperatively. On the other hand, the utility and efficiency of a P2P system is bound up with its network topology which is also influenced by peers' creditability. However, current unstructured P2P systems usually lack fair topology structures where uncooperative behaviors of peer aren't taken into consideration reasonably.

To solve the problem of peers' uncooperative behaviors as well as construct fair network topologies for unstructured P2P systems, we present a trust-based topology management protocol (TTMP). The protocol aims to bring contributing peers closer to each other and push the uncooperative peers away from the contributors. TTMP divides a single connection between every pair of neighboring peers into two links: an In-Link and an Out-Link. And TTMP manages these two types of links separately according to two different trust metrics defined in the proposed trust model. The organization of the paper is as follows. In Sect. 29.2 we discuss some related works. And in Sect. 29.3 we present the TTMP protocol in detail. The simulation results of TTMP are described in Sect. 29.4 and we give our conclusions in Sect. 29.5.

29.2 Related Works

To improve the usability and efficiency of P2P networks many researches have been done. We present some of the works related to our proposal in the following.

The authors of [4] devise a model to study the phenomenon of free-riding in P2P systems and the authors of [5] propose for partially decentralized P2P systems a reputation management scheme which can detect and punish malicious peers. Being different from the both works, our solution is to transform the cooperation of peers into network topologies, push the uncooperative peers to the edge of the network, so as to make the P2P system be more efficient.

Article [6] presents how to self-organize and self-tune the P2P topology based on peers' differing capabilities. It divides the connections in overlay network into two types: Search Links and Index Links. Search Links are used to forward queries and Index Links are used to send copies of content indexes. Inspired by this idea, the TTMP protocol separates each connection into an In-Links and an Out-Links. A peer can only get services by In-Links while contribute services through Out-Links.

The adaptive peer topologies (APT) protocol [7] is based on the idea that a peer should directly connect to those peers from which it is most likely to download

satisfactory content. And the adaptive gnutella protocol (AGP) [8] transforms the overlay topology based on a reputation scheme that evaluates the provided services and offers a mechanism that organizes trusted nodes with similar content. Both APT and AGP protocols optimize the network topology based on trust mechanism. Unlike the former protocol our trust model considers many past history services instead of latest one. And differing from the later protocol, TTMP can not only discourage uncooperative peers but also promote the fairness and service quality of P2P system.

29.3 TTMP Protocol

29.3.1 Overview of TTMP

To encourage the cooperation of peers and provide fair topologies for unstructured P2P system, the TTMP protocol proposed in this paper integrates a trust model into its topology managing scheme. The adopted trust model has two important evaluation metrics associated with each peer's reputation. The first one termed R_{cons} is the balance of amount of services a peer has contributed and consumed. This metric is used by a peer to request services from others. When receiving service requests from other peers, a peer always gives more privilege to the ones with higher value of R_{cons} for contributing services to them. And the value of R_{cons} increases when a peer provides services and decreases after using services. By adopting this metric TTMP can provide incentives to contributing peers as well as discourage uncooperative behaviors. In this way, the fairness of P2P system is promoted. And the second metric is R_{serv} which indicates the amount of services a peer has provided. Opposite to the first one, this metric is used by peers to contribute services to others. The higher value of R_{serv} means that the peer is more powerful to provide services. After sending a service query a peer can usually get several responding peers, then it will chose the peer with the highest value of R_{serv} to become its service provider. Hence, TTMP can improve the service quality of the P2P system through the metric R_{serv} . The concrete computing of R_{cons} and R_{serv} are given in [Sect. 29.3.2](#).

The aim of TTMP is to embed the proposed incentive mechanism of R_{cons} and R_{serv} into its topology management, since P2P network topology affects the propagation of queries and the overhead imposed on the underlying physical network. To achieve this goal, TTMP divides the connection between every pair of neighboring peers into two links: In-Link and Out-Link. A peer can only get services through its In-Links and provide services by its Out-Links. In this way, a peer decides whether to maintain or terminate an In-Link according to the value of its counterpart's R_{serv} , and manages an Out-Links accordingly based on the corresponding peer's value of R_{cons} . Therefore, TTMP can adjust the network topology dynamically in reaction to peers' past behaviors, and the adapted

topology can promote fairness, service quality and efficiency for unstructured P2P systems. The adaptive management scheme of network topologies is represented specifically in Sect. 29.3.3.

29.3.2 Trust Model

The trust model embedded by TTMP includes two reputation evaluation metrics: R_{serv} and R_{cons} . We describe the function of them as follows. *Reputation of service ability* (R_{serv}): it evaluates a peer's service ability and the value of R_{serv} is related to how much services the peer has contributed. *Reputation of consumption capacity* (R_{cons}): it evaluates a peer's consumption capacity and its value is related to the privilege level a peer can consume services. In the proposed model, each peer evaluates and stores the value of all its neighbor peers' R_{serv} and R_{cons} . Since every service interaction between peers involves one server and one consumer, each peer can only evaluate one corresponding value of its opposite side after a transaction. For an example, if peer i has contributed a service to peer j , then peer i evaluates R_{cons} of peer j while peer j evaluates R_{serv} of peer i .

In the following, we present the computing process of R_{serv} and R_{cons} . And the calculation is based on a scenario that peer j has given a service to peer i , and we call one peer is the neighbor of another when they have an overlay network connection or have history transaction records between each other within a given time span.

29.3.2.1 Computing R_{serv} of peer j by peer i

After the transaction given above, peer i computes the R_{serv} of j in the following steps.

Step1. Peer i rates the service satisfaction of j by function $f(i,j)$ defined in (29.1). If peer i is totally satisfied with the service contributed by j , it will give a value of 1. And if peer i Fig. 29.1 that j has provided a fake or bad service then i gives 0 for the service. The higher the value is, the more satisfactory peer i feels.

$$f(i,j) = \begin{cases} 1, & \text{totallysatisfactory} \\ 0 & \text{totallyunsatisfactory} \\ e(e \varepsilon (0, 1)) & \text{partiallysatisfactory} \end{cases} \quad (29.1)$$

Step2. Peer i aggregates the quantity of services (Q_{serv}) by Eq. 29.2. Q_{serv} is the amount of services that peer j has provided to peer i within a *time span* which can be set and tuned by the P2P system based on the overhead of peer's storage.

$$Q_{serv} = \sum_{Navrpan} f(i,j) \quad (29.2)$$

Step3. Peer i computes the frequency of services (T_{serv}) using Eq. 29.3 where m_s indicates the times of services provided by peer j within the same *time span* as in (29.2), n_s is the average service times of all neighbors of i , and β is a regulating constant.. Our claim is that if peer j provides services to i more frequently than other neighbors of i , then peer j can achieve more trust from peer i . To incorporate this idea into the trust model we define the parameter T_{serv} as:

$$T_{serv} = \begin{cases} \frac{m_s}{n_s} * \beta^{\frac{1}{m_s}} & (\beta \in (0, 1), m_s \neq 0) \\ (m_s) = 0 & \end{cases} \quad (29.3)$$

Finally, following the above three steps, peer i computes the R_{serv} of j by (29.4).

$$R_{serv} = Q_{serv} * T_{serv} \quad (29.4)$$

29.3.2.2 Computing R_{cons} of peer i by peer j

Opposite to the computing of peer j 's R_{serv} conducted by i , peer j will compute the R_{cons} of peer i upon the same transaction. Firstly, peer j rates the consumption satisfaction of peer i by the function given in (29.1), and we express it as $f(j, i)$. If peer j considers that peer i has finished consumption successfully or the service has been accepted totally by peer i , then j will give a value of 1. Accordingly, 0 denotes totally failure, and e ($0 < e < 1$) denotes else situations. Secondly, peer j use (29.5) to calculate the quantity of i 's consumptions (Q_{cons}) by summing up the consumptions that peer i has got from j within the same *time span* as in (29.2).

$$Q_{cons} = \sum_{timespan} f(j, i) \quad (29.5)$$

Thirdly, with the similar motivation of adopting parameter T_{serv} while computing R_{serv} , we use the parameter T_{cons} to indicate the frequency of peer i 's consumption. And the computing of T_{cons} is given in (29.6) where m_c denotes the times of consumptions performed by peer i and n_c is the average times of peer j 's all neighbors.

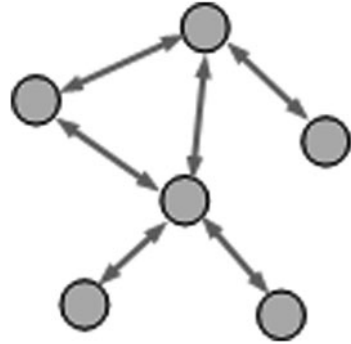
$$T_{cons} = \begin{cases} \frac{m_c}{n_c} * \beta^{\frac{1}{m_c}} & (\beta \in (0, 1), m_c \neq 0) \\ (m_c) = 0 & \end{cases} \quad (29.6)$$

To promote the fairness in P2P systems, we claim that a peer who has contributed more services to peer j should get more consumption privilege from j accordingly, while the consumption capacity must decrease as it consumes services. Therefore the R_{cons} of peer i is computed by peer j as:

$$R_{cons} = R_{serv} - Q_{cons} * T_{cons} \quad (29.7)$$

Although both (29.4) and (29.7) have the same " R_{serv} ", they are different in the scenario we presented here. Because the R_{serv} in (29.4) is about the service ability

Fig. 29.1 Network expressed in connections



of peer j while the one in (29.7) is that of peer i . And the R_{serv} in (29.7) is a history data which was computed and stored by peer j in another opposite transaction.

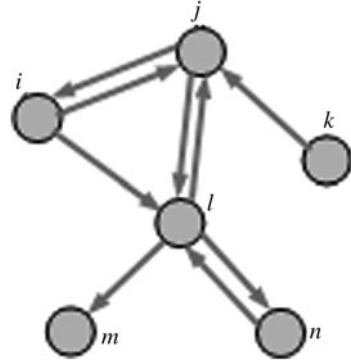
Therefore, from the computing process of R_{serv} and R_{cons} in the trust model, we can deduce that a malicious peer who provides bad services will have a low R_{serv} and a free-rider who contributes less service than its own fair share will arrive at a low R_{cons} , while cooperative peers can obtain high values on both reputation metrics.

29.3.3 Adaptive Topology Management

In general unstructured P2P overlay networks like Gnutella, every single topology connection can be used by its connected peer not only to provide services but also to get services. Differently, TTMP separates every single connection into two Links: an In-Link and an Out-Link. In-Links can only be used by the peer to obtain services, while Out-Links can only be used to contribute services. In this way, a peer uses In-Links to send queries and receive their replies, uses Out-Links to receive queries and reply them. The network topologies expressed in the form of connections and Links are shown separately in Figs. 29.1 and 29.2 where peers are represented as nodes and every line with arrow stands for a connection or a Link.

The main difference between Figs. 29.1 and 29.2 are the flowing directions of services which are represented as arrows. In Fig. 29.1, each line with two opposite arrow heads means that all topology connections can transport services in both directions. While in Fig. 29.2, the head of an arrow means an In-Link and the tail denotes an Out-Link. In an overlay network shown in Fig. 29.1, all neighboring peers have equal right to get and contribute services between each other, independent of their reputation values. However, the TTMP proposed overlay network shown in Fig. 29.2 can adjust its topology dynamically according to peers' reputation values of contribution and consumption, so as to provide incentives for cooperation of peers. Based on the considering of its bandwidth and capacity which constraint its maximum number of neighboring Links, each peer decides

Fig. 29.2 Network expressed in Links



whether to establish or release an Out-Link according to the value of neighbor's R_{cons} and performs similarly with an In-Link based on the neighbor's R_{serv} . As a result, a free-rider like peer k in Fig. 29.2 will eventually have no In-Link for receiving services from other peers, and a peer providing bad services like m can't dispatch its shares because of lacking Out-Link.

After finishing a service transaction, a peer of service contributor computes and updates the metric R_{cons} of its counterpart. And a peer of service consumer performs that both on R_{cons} and R_{serv} , because the changing of the counterpart's R_{serv} also influences the value of its R_{cons} . Then based on the computing results, the consumer manages its corresponding In-Link and Out-Link, and the contributor adjusts only its Out-Link. If the trust metric related to an existing Link is lower than the minimum threshold of the local peer, then the local peer will replace the Link with a new one. And if the trust metric of counterpart not having the corresponding Link is large enough or the Links of local peer is scarce, then the manager will try to establish a new Link to that peer. The following algorithm describes the Link managing of a peer.

Algorithm The pseudo-code that peer i manages its Link after finishing a service transaction with j

```

if (peer  $i$  is a service consumer) then
   $R_{\text{serv}} = \text{ComputingRservOf}(\text{peer } j)$ 
   $\text{StoreData}(R_{\text{serv}}, \text{ID}_j)$ 
  //managing In-Link of peer  $i$  in the following
   $\text{InLink} = \text{SearchInLinkTo}(\text{peer } j)$  //search for an In-Link with  $j$ 
  if (InLink is FOUND) then
    if ( $R_{\text{serv}} < \text{MinThresholdOfRserv}$ ) then //replace the In-Link
       $\text{ReleaseInLinkTo}(\text{peer } j)$ 
       $k = \text{SelectCandidate}()$  //select peer with highest  $R_{\text{serv}}$  from
candidates
      //to which  $i$  hasn't an In-Link, but  $i$  has their past transaction data locally.
      Send a ping message to  $k$  for creating an In-Link
      if (a pong arrives from  $k$ ) then //add new link by ping, pong

```

```

        EstablishInLinkTo(k)
    end if
end if
else
    if (NumberOfInLink < MaxNumOfInLink and Rserv > 0) then
        AddNewInLinkTo(j) //perform the same as add new link to k above
    else //decide if replace the In-Link of neighbor having minimum  $R_{serv}$ 
        k = SelectMinRserv() //select the neighbor with minimum  $R_{serv}$ 
        if (Rserv > RservOfK) then
            AddNewInLinkTo(j)
            If (pong message from j is received) then
                ReleaseInLinkTo(k)
            end if
        end if
    end if
end if
endif
//No mater peer i is a service consumer or contributor, it must manage Out-
Links
Rcons = ComputingRconsOf(peer j)
StoreData(Rcons, IDj)
ManageOutLinkTo(peer i) //The pseudo-code of this part is similar to that
of//managing In-Link. And it can be achieved by changing the code corresponding
//with  $R_{serv}$  into  $R_{cons}$  and changing the In-Link related pseudo-code into Out-
Link.

```

Upon receiving a ping message from another peer for establishing a new Link, a peer needs to manage its Link as well. And its action is similar to what we describe in the above algorithm except for the computing and storing of trust metrics. In a word, by the peers' adaptive managing of Links, contributing peers will move closer to each other and uncooperative peers will be pushed toward the edge of the network.

29.4 Simulation Results

PeerSim [9] is one of the most known simulators among P2P researchers. The PeerSim engine has two simulation models: event based model and cycle based model. We have chosen the cycle based model where nodes communicate with each other directly, and the nodes are given the control periodically, in some sequential order. The proposed TTMP protocol has been simulated under the PeerSim1.0.5 platform. And we have chosen the Gnutella v0.6 for experiment comparisons. The network size was 1,000 which included 600 cooperative peers, 300 free-riders and 100 malicious peers. The corresponding parameters' settings were: Random Graph network, max degree $k = 20$, $M = -5$, $\beta = 0.5$, number of cycles = 5,000.

Fig. 29.3 Number of links between cooperative peers

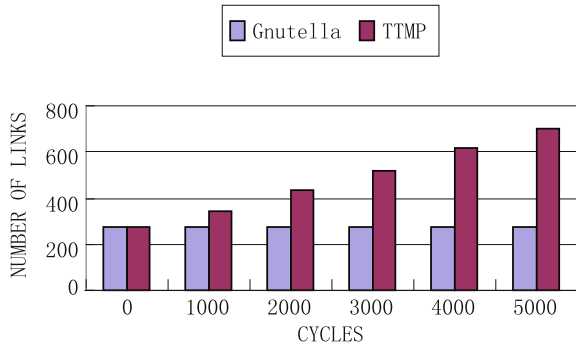


Fig. 29.4 Number of links to free-riders

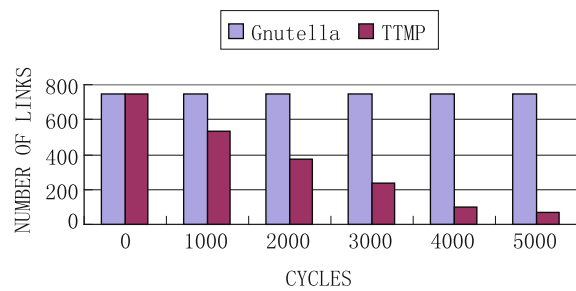


Fig. 29.5 Number of links to malicious peers

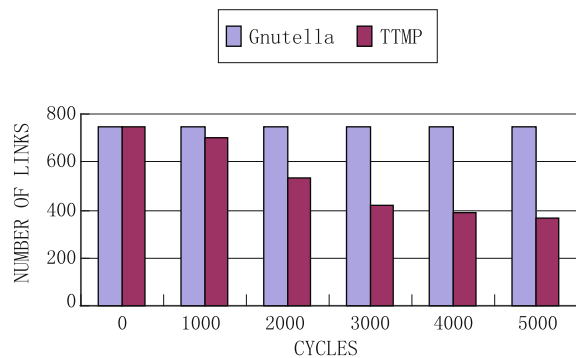
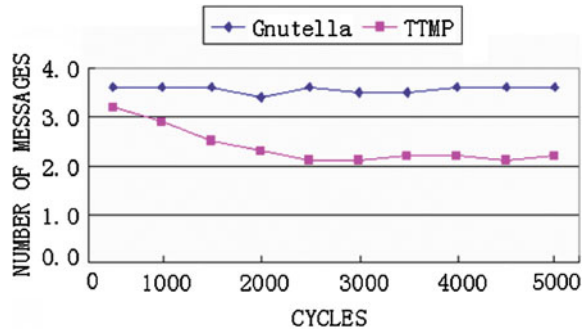


Figure 29.3 shows that the number of Links between cooperative peers increases over time. It indicates TTMP protocol brings contributing peers closer to each other by its topology management. And in Fig. 29.4, the number of Links to free-riders reduces slowly at the beginning. But the number decreases rapidly after 1,000 cycles when most free-riders’ values of R_{cons} become lower than their Minimum Thresholds.

As for malicious peers who provided bad services and requested fewer services than others, Fig. 29.5 shows that the number of Links to them decreases as well.

Fig. 29.6 Number of messages by uncooperative peers



But it is slower than that of free-riders because of their scarcer consumptions. And in Fig. 29.6, the average number of messages to and from an uncooperative peer per cycle is reduced greatly. We average it by every 500 cycles, and the metric drops from 3.6 in Gnutella to 2.2 in TTMP, a reduction of about 39%. This indicates that TTMP can alleviate the messaging overhead of P2P network, so as to improve its efficiency.

29.5 Conclusions

In this paper, we present TTMP, a TTMP which aims to provide incentives for peers to cooperate with each other in unstructured P2P network like Gnutella. TTMP protocol integrates a trust model into its topology managing scheme. And the solution is distributed and does not require a central entity to control and coordinate. The trust model includes two important evaluation metrics of R_{cons} and R_{serv} . TTMP can promote the fairness of P2P system by adopting R_{cons} and improve its service quality by embedding the metric R_{serv} . Each peer manages its In-Link and Out-Link according to the value of counterpart's R_{serv} and R_{cons} separately. As a result, uncooperative peers are pushed toward the edge of the network. And the simulation results indicate that the protocol increases the number of Links between cooperative peers while reduces the number of Links to uncooperative peers. Furthermore, the protocol can also alleviate the messaging overhead resulting from uncooperative peers, so as to improve the efficiency of unstructured P2P networks.

References

1. Shneidman J, Parkes D (2003) Rationality and self-interest in peer-to-peer networks. The second international workshop on peer-to-peer systems. Berkeley, CA
2. Keidar Idit, Melamed Roie, EquiCast Ariel Orda (2009) Scalable multicast with selfish users. *Comput Netw* 53:2373–2386

3. Hughes D, Coulson G, Walkerdine J (2005) Free-riding on Gnutella revisited: the bell tolls. In: IEEE distributed systems online
4. Feldman M, Christos P, Chuang J et al (2006) Free-riding and whitewashing in peer-to-peer systems[J]. IEEE Trans Syst Man and Cybernetics 24(5):228–236
5. Mekouar L, Iraqi Y, Boutaba R (2006) Peer-to-peer's most wanted: malicious peers. Comput Netw 50(4):545–562
6. Cooper BF, Garcia-Molina H (2003) Ad hoc, self-supervising peer-to-peer search networks[R]. Stanford University
7. Condie T, Kamvar SD, Garcia-Molina H (2004) Adaptive peer-to-peer topologies[C]. In: Lambrix P, Duma C (eds) Proceedings of the 4th international conference on peer-to-peer computing. IEEE Press, New York, pp 53–62
8. Pogkas I, Kriakov V, Chen ZQ, Delis A (2008) Adaptive management of communities based on peer content and reputation[C]. In: The third international multi-conference on computing in the global information technology, iccgi, pp 291–296
9. Jelasity M, Montresor A, Jesi GP, Voulgaris PeerSim S (2005) A peer-to-peer simulator, December

Chapter 30

Indoor Temperature and Humidity Monitoring System Based on WSN and Fuzzy Strategy

Bo Chang and Xinrong Zhang

Abstract According to the characteristics of large delay and large inertia of controlled object in the indoor temperature and humidity, an automatic monitoring system of indoor environment based on wireless sensor network is designed. By placing sensors network nodes in monitoring area, the monitoring information is aggregated to the central monitoring system in order to achieve a unified data management and network routing monitoring functions. The real-time data is collected, processed and sent to the receiver by wireless sensor network. At the receiving end, the data is received, saved and displayed to achieve temperature and humidity monitoring and fuzzy control by which the best expected value is get to achieve automatic control of indoor temperature and humidity. The monitoring results have shown that this system is stable, reliable, cost-effective, high regulation accuracy, adjust quickly and small overshoot.

Keywords Temperature and humidity · Wireless sensor network (WSN) · Fuzzy control strategy · Monitoring

B. Chang (✉)

Faculty of Computer Engineering, Huaiyin Institute of Technology, Huaian, Jiangsu
China

e-mail: mmm33534@sohu.com

X. Zhang

Faculty of Electrical Engineering, Huaiyin Institute of Technology, Huaian, Jiangsu
China

e-mail: nn33@163.com

30.1 Introduction

With the improvement of people's living standards, living environment quality is highly demanded more and more, especially in indoor environment such as temperature and humidity for human comfort. Most of the existing environmental monitoring system can not carry out remote access and environmental monitoring because their using wired data acquisition and transmission, on-site installation and wiring tedious workload, equipment, poor mobility, network complexity and high cost. Therefore, it is very necessary to research and develop an wireless intelligent environmental monitoring system [1]. WSN consists of information collection, processing and transmission, and it's widely used for monitoring environmental information so technical support is provided in many areas, such as national defense, environmental monitoring, engineering safety and agricultural and food processing, etc. [2]. Typical applications already have a lot of cases [3–6].

An integration solutions of remote monitoring system is proposed based ZigBee network [7, 8], which based on fast and flexible control features of Microcontroller and a powerful PC, monitoring and management capabilities and the development needs of access to information technology. Through ZigBee sensor network technology, wireless communication between sensor nodes and central node is achieved, to meet the requirements of multiple measurement points, multiple factors, mobility, convenience etc., during the course of environmental monitoring.

30.2 Monitoring System Design

Using wireless sensor nodes to achieve the indoor temperature, humidity and other physical information collection and transmission, and the fuzzy operation processing, through a multi-hop transmission to the host PC, System achieves adjustment and control of the temperature and humidity. The distributed structure of the overall system is adopted, taking into account the characteristics of the indoor environment monitoring parameters. The overall structure of this system is shown in Fig. 30.1.

Command response mode is adopted in the system. The data acquisition commands are issued by the master station and the address information sent from the master station is received and processed by the slave station. If the address information is consistent with the local address, the command is executed. Taking into account the conservation of energy, temperature and humidity sensing nodes would be periodically waked up and collect regularly indoor environmental data. By the signal conditioning and A/D conversion, these data will be sent to the central master microcontroller system by multi-hop routing. The data from the serial port is uploaded to a central monitoring PC, is saved, is analyzed and is

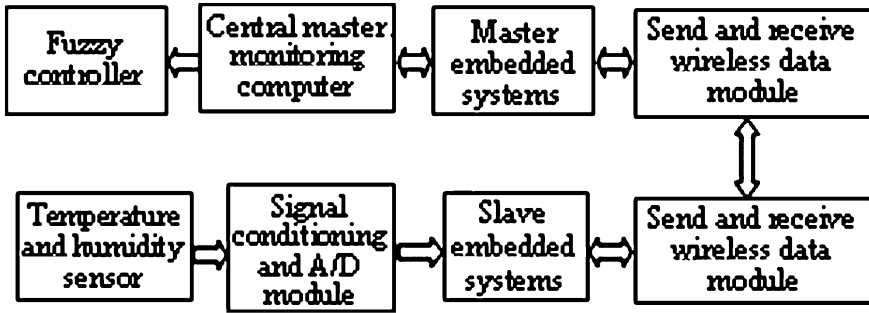


Fig. 30.1 System block diagram

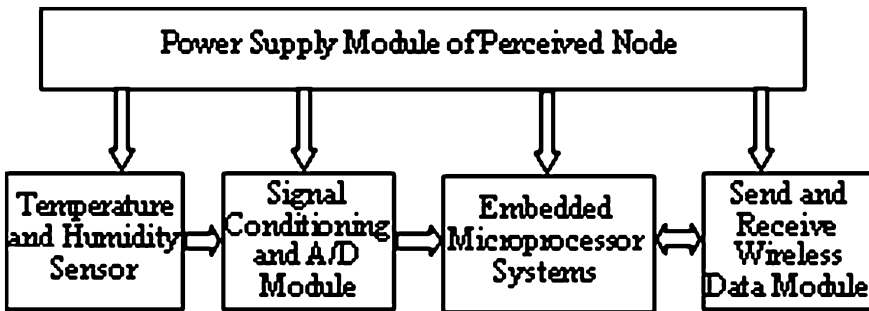


Fig. 30.2 Hardware structure of sensor node

managed. By the fuzzy operation, the output decision information is used for automatic adjustment of temperature and humidity parameters.

30.3 Sensor Node Design

According to the indoor environmental characteristics and the actual needs, temperature and humidity sensors are deployed by the artificial. The transfer rate of 250 Kb/s and the band of 2.4 GHz is preferred, which is the common band of current global sensor network. There are more than one ZigBee monitoring networks in the bottom and star structure is used. Node hardware structure is shown in Fig. 30.2.

30.3.1 Selection of the Embedded Microprocessor

A P89V51RB2 is used, which is an 80C51 microcontroller, including 16kB Flash and 1024 bytes of data RAM. The typical feature of P89V51RB2 is its X2 mode option. Using this feature the application can run in the clock frequency of 80C51, that is, 12 clocks per machine cycle, or the clock frequency of X2, that is, 6 clocks per machine cycle. So 2 times throughput can be obtained selecting the X2 mode at the same clock frequency.

30.3.2 Sensor Selection and Interface Design with Microcontroller

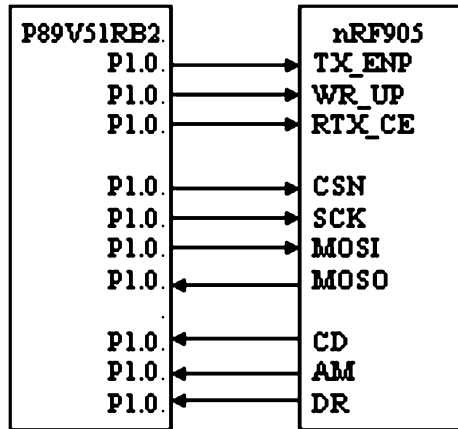
Digital sensor chip SHT11 [9], that produced by the Sensirion Company, chiefly consists of temperature and humidity sensor, amplifier, A/D converter circuit and memory. The salient features of this chip are: two-wire digital interface, fully digital output, without fine-tuning, can be directly connected with the MCU, the simple external circuit, small size, small current consumption etc. The temperature range of $-40.0 \sim 123.8^{\circ}\text{C}$, Humidity measuring range of $0 \sim 100\% \text{ RH}$, can meet the testing requirements of indoor temperature and humidity environment.

SHT11 connect with microcontroller P89V51RB2 through the two-wire serial interface circuit, that is, the port P2.0 connected to the SCK port of SHT11 and P2.1 connected to the port DATA. The SCK is used to synchronize the communication between a microcontroller and the SHT11. The DATA tristate pin is used to transfer data in and out of the device. DATA changes after the falling edge and is valid on the rising edge of the serial clock SCK. During transmission the DATA line must remain stable while SCK is high. To avoid signal contention the microcontroller should only drive DATA low.

30.3.3 Design of Send and Receive Wireless Data Module

The send and receive wireless data module nRF905 from Nordic company is a single-chip radio transceiver for the 433/868/915 MHz ISM band [10]. The transceiver consists of a fully integrated frequency synthesizer, receiver chain with demodulator, a power amplifier, a crystal oscillator and a modulator. The ShockBurst feature automatically handles preamble and CRC. Configuration is easily programmable by use of the SPI interface. Current consumption is very low, in transmit only 11 mA at an output power of -10dBm , and in receive 12.5 mA. Built in power down modes makes power saving easily realizable.

Fig. 30.3 Interface connection between nRF905 and P89V51RB2



30.3.4 Interface Design Between nRF905 and P89V51RB2

An nRF905 is connected to microcontroller P89V51RB2 by use of the SPI interface. P89V51RB2 works in the SPI host mode and nRF905 work in slave mode. The interface connection is shown in Fig. 30.3.

30.3.5 Data Transfer Module

Node ID occupies 2 bytes respectively and adopts uniform coding. Each node can monitor a variety of environmental parameters; distinguish between the different parameters with 4 bit. The analog–digital conversion accuracy of each node is 12 bit and for each sensor node, measurement data takes 2 Byte. Wireless data transfer process flow is shown in Fig. 30.4.

30.4 Fuzzy Control Strategy

A precise mathematical model can not establish because of the characteristics of the room temperature and humidity control, namely a multi-variable, large inertia, nonlinearity, parameter coupling, pure time delay and control regulation longer and producing significant overshoot. So it is difficult to obtain a good control accuracy using the classical control methods. As an important branch of the field of intelligent control, the fuzzy control strategy mimics the human thought to control, with simple design and robust advantages. Therefore, the fuzzy control

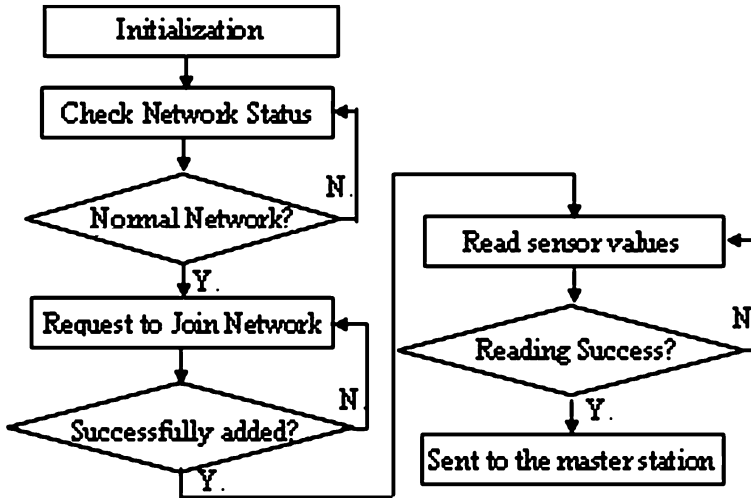


Fig. 30.4 Data transfer process flow chart

algorithm has been used to control the system environmental factors and software tools have been used for the aided design of fuzzy control system.

The task of fuzzy controller design is to select the input and output variables. The variables, that are observed by operator, based on operator experience, can be selected as input variables. The output variable is selected to control the actuator. The dual-input single-output control structure is used for this fuzzy control system. The input linguistic variables of the temperature fuzzy controller are the error and its rate of change between a given temperature and the actual temperature, while the output linguistic variable is switch state of heating and cooling equipment.

The design method of input linguistic variables and output linguistic variable membership function is that firstly the domain X of temperature error E and the domain Y of its change rate EC are determined, secondly the quantify factors K_E and K_{EC} are calculated separately, then language variables- NB , NS , Z , PS , PB are selected and lastly the membership function of describing fuzzy sets are determined by the experience of the operator.

By summarizing the technology and operator experience, control rule statements are accessed thus the temperature fuzzy control rule table is established. Owing to given the single chip can not do such a large amount of computation, the solution is to produce a fuzzy control table shown in Table 30.1 and then to search for the fuzzy operation through Software ways.

The fuzzy control table obtained by off-line operations is input into the computer. The controlled variable of non-fuzzy sets will be obtained by looking up fuzzy control table. The fuzzy control algorithm flow chart is shown in Fig. 30.5.

Table 30.1 Temperature fuzzy control query table

<i>EC</i>	<i>E</i>								
	-4	-3	-2	-1	0	+1	+2	+3	+4
-2	+1	+1	+1	+1	0	0	0	-1	-1
-1	+1	+1	0	0	0	0	0	-1	-1
0	+1	+1	0	0	0	0	0	-1	-1
+1	+1	+1	0	0	0	0	0	-1	-1
+2	+1	+1	0	0	0	-1	-1	-1	-1

Fig. 30.5 Fuzzy control algorithm flow chart

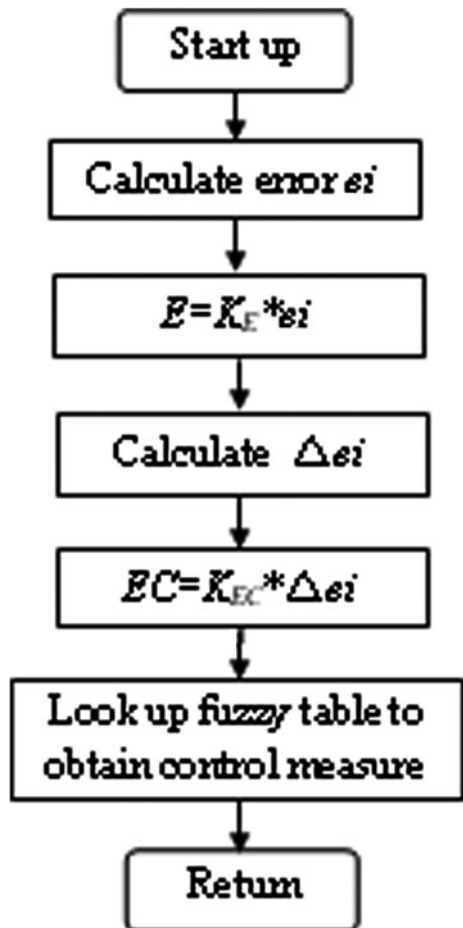
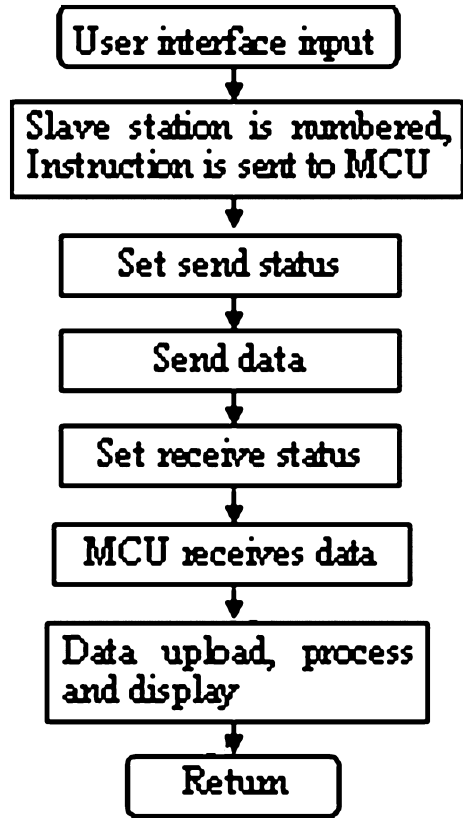


Fig. 30.6 Program flow chart of master station



30.5 System Software Design

Master computer provides human-machine interface and data processing, sends commands to the MCU through the RS232 interface and receives collected data returned from MCU. Based on modular design principles, the master control system functions are divided into different modules. In each functional module, a small amount of shared variables are shielded, which makes the modules to operate independently, so as to facilitate software modified, debugged and extended. By monitoring software, the data is acquired and displayed real-timely, and the reports are formed for the prediction and making decision. Master program flow chart is shown in Fig. 30.6.

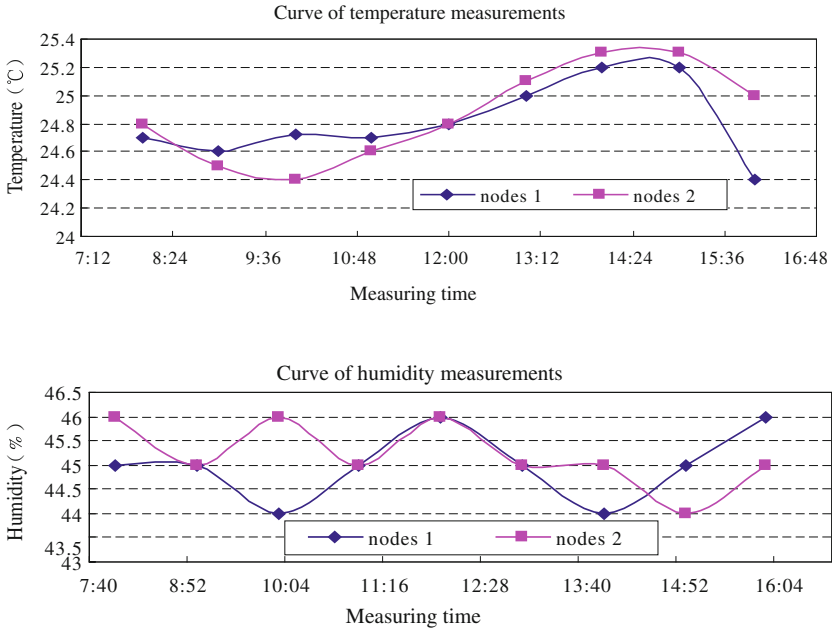


Fig. 30.7 Curve of temperature and humidity measurements

30.6 System Test Results and Analysis

This system was completed based on multiple detection and controls, combined with the latest ZigBee wireless networking technology. The system has been tested in the laboratory field (about 160 m²), where eight temperature and humidity sensing nodes and a central monitoring computer are arranged randomly.

By the testing, within the 15 m range temperature and humidity sensing nodes can pass through the sensor test bench equipment such as obstructions; can achieve data collection, transmission, and the list shows. Network is good in the self-correcting ability and the data accuracy. Temperature measurement error is less than $\pm 0.5^{\circ}\text{C}$, and relative humidity measurement error is within 5%. Once detection parameters is overrunning, the computer response time is usually less than 3 s, and then computer will display address and gauge parameter values and will raise the alarm. The monitoring data of part of the node at several time points is shown in Fig. 30.7.

30.7 Summary

The design of distributed control system of indoor wireless temperature and humidity improves the overall performance of the system in the aspects of convenience and networking flexibility. And this system can monitor node data real-time and refresh timely. In view of the low-power design, data transmission signal is relatively weak, anti-interference ability is poor and the phenomenon of loss of data is prone to come forth, which have been improved by using the software method to increase the frequency of data collection, extend data transmission cycle and reduce data loss rate. A feasible, practical, relatively low-cost solution has been provided for indoor temperature and humidity monitoring and control. The results have shown that the system can promptly detect changes in indoor temperature and humidity, can control the decisions.

Acknowledgments This work was supported by the 2010 Scientific and Technological Support Projects of Huai'an, China (Grant No. SN1045) and the 2010 Technology Research Fund of Huaiyin Institute of Technology, China (Grant No. HGB1010). We would like to thank the anonymous reviewers for their perspicacious comments.

References

1. Cook DJ, Das SK (2005) Smart environments:technologies, protocols, and applications[M]. John Wiley and Sons, Hoboken
2. Sun Z, Du KM, Han HF, et al. (2008) Design of a telemonitoring system for data acquisition of livestock environment[C]. Livestock environment VIII-proceedings of the 8th international symposium, Iguassu Falls, Brazil, ASABE, pp 995–1000
3. Ana Carolina de Sousa Silva, Aldo Ivan C'espedes Arce, S'ergio Souto, et al. (2005) A wireless floating base sensor network for physiological responses of livestock[A]. Computers and Electronics in Agriculture[C]. Elsevier, Amsterdam, The Netherlands, pp 246–255
4. Handcock RN, Swain DL, Greg J et al (2009) Monitoring animal behaviour and environmental interactions using wireless sensor networks, GPS collars and satellite remote sensing[J]. Sensors 9(5):3586–3603
5. Kim Y, Evans RG, Iversen WM, Pierce FJ (2006) Instrumentation and control for wireless sensor network for automated irrigation[C]. ASABE annual international meeting, Portland. ASABE Paper no 061105
6. Morais R, Valente A, Ser dio C (2005) A wireless sensor network for smart irrigation and environmental monitoring[C]. EFITA/WCCA Vila Real Portugal, pp 845–850
7. Barker N (2005) ZigBee and bluetooth strengths and weaknesses for industrial applications[J]. Computing & Control Engineering Journal 16(2):20–25
8. Hill JL, Culler DE (2002) Mica: a wireless platform for deeply embedded networks[J]. IEEE Micro 22(6):12–22
9. <http://www.sensirion.com/SHT1x/SHT7x> Relativek Humidity & Temperature Sensor System
10. Proakis JG, Salehi M (2001) Communication systems engineering[M]. PrenticeHal, UpperSaddle River

Chapter 31

Radioactive Target Detection Using Wireless Sensor Network

Tonglin Zhang

Abstract The detection of radioactive target is becoming more important recently in public safety and national security. By using the physical law for nuclear radiation isotopes, this chapter proposes a statistical method for wireless sensor network data to detect and locate a hidden nuclear target in a large study area. The method assumes multiple radiation detectors have been used as sensor nodes in a wireless sensor network. Radiation counts have been observed by sensors and each is composed of a radiation signal plus a radiation background. By considering the physical properties of radiation signal and background, the proposed method can simultaneously detect and locate the radioactive target in the area. Our simulation results have shown that the proposed method is effective and efficient in detection and location of the nuclear radioactive target. This research will have wide applications in the nuclear safety and security problems.

Keywords Decision and value fusion • Likelihood ratio test • Maximum likelihood estimates • Signal plus background model • Radiation and radioactive isotopes • Wireless sensor network

31.1 Introduction

Recent advances in wireless communications and electronics have enabled the development of low cost, lower-power, multifunctional sensor nodes that are small in size and efficiently communicate in short distance [1]. These tiny sensor nodes

T. Zhang (✉)

Department of Statistics, Purdue University, 150 N. University Street,

West Lafayette 47906, USA

e-mail: tlzhang@purdue.edu

which consist of sensing, data processing, and communicating components, enhanced the ideas of wireless sensor networks (WSN). The goal of any WSN is to provide measures of a given physical process, which may have many physical parameters, to detect specific events [2]. Therefore, there are substantial amount of work on WNS for target detection [3, 4, 5]. A general question for a WSN is how to efficiently integrate the available information from individual sensors to reach a global decision about the presence of a target over a monitoring area. The past work on signal detection with WSN can be categorized into two groups of methods: *decision fusion* and *value fusion* methods, respectively [6]. In decision fusion, each sensor makes its own binary decision and then the network will make a consensus by fusing all decisions [5, 7, 8]. In value fusion, sensors collect measurements and the network will make a decision by fusing the collected values.

In this chapter, we focus on the value fusion WSN methods because the original radiation data are counts, which are easily transmitted to the fusing center with low cost [9]. We assume a value fusion WSN has been used to monitor a large study area. In the monitored area, neither the location nor the strength of the radioactive target is known. A radiation detector, also known as a particle detector, is a device used to detect, track, and/or identify high-energy particles emitted from radioactive materials, including neutrons, alpha particles, beta particles, and gamma rays. In general, the observed radiation counts received by detectors can be modeled as a mixture of the signal from the radioactive source and the natural background radiation [10]. Let the area and efficiency of the radiation detector be A and e respectively. Then the total number of radiation count received by the detector (denoted by y) satisfies

$$y \sim \text{Poisson}\left(TAe\left(v_b + \frac{v_s}{4\pi r^2}\right)\right), \quad (31.1)$$

where T is the time duration, r is the distance between the detector and nuclear radioactive target, v_s is the surface radiation intensity rate, and v_b is the background radiation intensity rate. The research will have wide applications in the nuclear safety and security problem. We give the following two examples to highlight the importance.

Example 1 Since the world trade center tragedies on September 11, 2001, the United States government has acknowledged the high possibility of terrorist attacks using weapons of mass destruction (WMD) [11]. A good summary of physical models of the nuclear weapons can be found in Fetter et al. [12]. Because a hidden nuclear weapon contains radioactive materials, which emit radiation from its surface, the method developed by this chapter will likely be used to the defense of nuclear terrorism country-widely.

Example 2 The recent Fukushima nuclear power station disasters caused by the earthquake in Japan on March 11, 2001 have made serious nuclear pollution in the World. Currently, it is still too early to judge the final outcomes of the nuclear crisis that continues to Japan. This tragedy uncovers a huge safety and security problems that exist in all of the nuclear power plants in the world. In nuclear power

plants, nuclear fuel and waste contain uranium or other radioactive materials. In order to improve the safety and security level of a nuclear power plant, it is important to develop a quickly and accurately detection system for the leakage of the nuclear units. The method developed by this chapter will provide a theoretical base for the detection system.

This statistical model of the above examples is called the signal plus background model in the particle physics [13]. They can be written as special cases of Model (31.1). Based on the general framework of WSN, we may deploy many radiation detectors as sensor nodes at different locations. The statistical approach then can be proposed.

The rest of the paper is organized as follows. In Sect. 31.2, we introduce our statistical methods. In Sect. 31.3, we introduce our numerical algorithms. In Sect. 31.4, we display our simulation results. In Sect. 31.5, we present our conclusion.

31.2 Statistical Method

Assume a nuclear radiation target is hidden at an unknown location denoted by $\omega = (\omega_1, \omega_2, \omega_3)$. Suppose a WSN is used to detect and locate the target. Assume the WSN has m radiation detectors, which are deployed at $a_i = (a_{i1}, a_{i2}, a_{i3})$ for $i = 1, \dots, m$ respectively. Let

$$r_i(\omega) = \|a_i - \omega\| = \sqrt{(a_{i1} - \omega_1)^2 + (a_{i2} - \omega_2)^2 + (a_{i3} - \omega_3)^2}$$

be the Euclidean distance between the target and the i -detector. Let A_i be the area and e_i be the efficiency of the i -th detector respectively. Let T be the duration time of the detection (given by seconds). Let y_i be the total number of radiation count observed by the i -th detector during the detection period. Then, we have

$$y_i \sim \text{Poisson}\left(T\xi_i\left(v_b + \frac{v_s}{4\pi r_i^2(\omega)}\right)\right), i = 1, \dots, m, \quad (31.2)$$

where $\xi_i = A_i e_i$ is the capability of detection for i -th detector. The unknown parameters in model (31.2) are v_b , v_s and ω . Based on model (31.2), the detection problem is interpreted as the hypothesis test of

$$H_0 : v_s = 0 \text{ vs } H_1 : v_s > 0, \quad (31.3)$$

and the location problem is interpreted as the estimation and confidence interval for ω .

We propose a statistical method to test the significance of $H_0 : v_s = 0$. The method is based on the famous likelihood ratio test, which is accessed by the likelihood ratio statistic. The likelihood ratio statistic is defined by the ratio of likelihood functions derived under $H_0 \cup H_1 : v_s \geq 0$ and under $H_0 : v_s = 0$

respectively. When the test is significant, the location problem is accessed by the derivation of the estimate and confidence interval for ω , which can be derived by the maximum likelihood method.

In order to derive the likelihood ratio test statistic. We need to compute the maximum of the likelihood functions under $H_0 \cup H_1 : v_s \geq 0$ and $H_0 : v_s = 0$ respectively. In order to derive the estimate and confidence interval for ω , we need to derive the limiting distribution of the maximum likelihood of ω . The statistical approach is displayed in the remaining part of this section. The corresponding numerical algorithms will be displayed in the next section.

Straightforwardly, the log likelihood function under model (31.2) is

$$\ell(v_s, v_b, \omega) = - \sum_{i=1}^m \log(Y_i!) + \sum_{i=1}^m Y_i \log\left[T \xi_i \left(v_b + \frac{v_s}{4\pi r_i^2(\omega)}\right)\right] - T \sum_{i=1}^m \xi_i \left[v_b + \frac{v_s}{4\pi r_i^2(\omega)}\right]. \tag{31.4}$$

Under $H_0 : v_s = 0$, the loglikelihood function $\ell(v_s, v_b, \omega)$ in Equation (31.4) becomes

$$\ell_0(v_b) = - \sum_{i=1}^m \log(Y_i!) + \sum_{i=1}^m Y_i \log(T \xi_i v_b) - T v_b \sum_{i=1}^m \xi_i. \tag{31.5}$$

Note that $\ell_0(v_b)$ only contains parameter v_b . It is not necessary to compute the estimate of ω under the null hypothesis.

Let $\hat{v}_{b,0} = \sum_{i=1}^m Y_i / (T \sum_{i=1}^m \xi_i)$ be maximum likelihood estimate (MLE) of v_b under $H_0 : v_s = 0$. Then, \hat{v}_b can be analytically solved by $\hat{v}_{b,0} = \sum_{i=1}^m Y_i / (T \sum_{i=1}^m \xi_i)$. Let \hat{v}_s , \hat{v}_b and $\hat{\omega}$ be the MLE of v_s , v_b and ω under $H_0 \cup H_1 : v_s \geq 0$. Because the MLE \hat{v}_s , \hat{v}_b and $\hat{\omega}$ cannot be analytically solved, we have to develop a numerical algorithm of \hat{v}_s , \hat{v}_b and $\hat{\omega}$. This method will be introduced in the next Section. Suppose \hat{v}_s , \hat{v}_b and $\hat{\omega}$ are derived. Then, $(\hat{v}_s, \hat{v}_b, \hat{\omega})$ as $T \rightarrow \infty$ is asymptotically normal [16]. This property will be used to locate nuclear radioactive target when the test is significant.

We propose a loglikelihood ratio test to access the null hypothesis. By comparing Eqs. (31.4) and (31.5), we find that the location ω is not present in (31.5). Therefore, this problem is nonstandard because the classical loglikelihood ratio test does not possess it usually asymptotic null distribution [1, 8]. To define the loglikelihood ratio test, we propose a conditional test statistic and use it to formulate the test by maximizing the conditional test statistic.

Suppose ω is pre-selected. Then, Eq. (31.4) only contains parameters v_s and v_b . In this case, the testing problem becomes standard. Therefore, the loglikelihood ratio test can be formulated by the conditional test statistic given by

$$\Lambda(\omega) = 2[\ell(\hat{v}_{s,\omega}, \hat{v}_{b,\omega}, \omega) - \ell_0(\hat{v}_{b,0})] \tag{31.6}$$

where $\hat{v}_{s,\omega}$ and $\hat{v}_{b,\omega}$ are the conditional MLE of v_s and v_b under $H_0 \cup H_1 : v_s \geq 0$ respectively. It is easily to see that Λ approximately follows χ^2_1 distribution if T is large [16]. Because ω is unknown, we assume it belongs a set $D \subseteq R^3$. Then, the likelihood ratio test statistic is defined by

$$\Lambda = \sup_{\omega \in D} \Lambda(\omega). \quad (31.7)$$

The null hypothesis $H_0 : v_s = 0$ is rejected if Λ is large. Because neither the approximate nor the exact distribution of Λ is known, we propose a bootstrap method to access its p -value. If the p -value of Λ is less than a pre-selected significance level α , the test is significant; otherwise, the test is not significant.

When the test is significant, we need to develop a method to locate the radioactive target in the study area. Since the problem becomes standard under $v_s > 0$, the location ω of the radioactive target then can be estimated by the maximum likelihood method and its asymptotical distribution can be easily derived. Based on this property, we can compute the $100(1 - \alpha)\%$ elliptical confidence region for ω by

$$C_\alpha(\omega) = \left\{ \omega : \frac{1}{T} (\omega - \hat{\omega})' I(\hat{v}_s, \hat{v}_b, \hat{\omega}) (\omega - \hat{\omega}) \leq \chi^2_{\alpha,3} \right\},$$

where $\chi^2_{\alpha,3}$ is the upper α quantile of the χ^2_3 distribution, and $I(\hat{v}_s, \hat{v}_b, \hat{\omega})$ is the 3×3 Fisher Information matrix given by

$$I_{j_1, j_2} = \sum_{i=1}^m \frac{4\zeta_i \hat{v}_s^2 (a_{ij_1} - \hat{\omega}_{j_1})(a_{ij_2} - \hat{\omega}_{j_2})}{\hat{v}_b r_i^8(\hat{\omega}) + \hat{v}_s r_i^6(\hat{\omega})}, j_1, j_2 = 1, 2, 3. \quad (31.8)$$

31.3 Numerical Algorithms

In this section, we introduce our numerical methods. The methods includes the algorithm for $\hat{v}_{s,\omega}, \hat{v}_{b,\omega}$ for a given ω , the algorithm for $(\hat{v}_s, \hat{v}_b, \hat{\omega})$, and the algorithm for the bootstrap p -value of Λ . The algorithm for $(\hat{v}_s, \hat{v}_b, \hat{\omega})$ is developed under the algorithm for $\hat{v}_{s,\omega}, \hat{v}_{b,\omega}$. It is also used to develop the algorithm for the bootstrap p -value of Λ .

When ω is pre-selected, $\hat{v}_{s,\omega}, \hat{v}_{b,\omega}$ can be derived by solving the likelihood equation

$$\nabla_\omega(v_s, v_b) = \begin{pmatrix} \partial \ell / \partial v_s \\ \partial \ell / \partial v_b \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Let $H_\omega(v_s, v_b)$ be the Hessian matrix of $\ell(v_s, v_b, \omega)$ conditional on ω . Then, $H_\omega(v_s, v_b)$ is always negative definite because its eigenvalues are always negative

[17]. Therefore, $\ell(v_s, v_b, \omega)$ as a function of v_s and v_b is concave, which induces the following Newton–Raphson algorithm for $\hat{v}_{s,\omega}$ and $\hat{v}_{b,\omega}$ always converges.

Algorithm for $\hat{v}_{s,\omega}$ and $\hat{v}_{b,\omega}$.

- Derive an initial guess of $v_{s,\omega}^{(0)}$ and $v_{b,\omega}^{(0)}$.
- Let $v_{s,\omega}^{(u)}$ and $v_{b,\omega}^{(u)}$ be the solution in step u . Then, the solution of the next step is updated by

$$\begin{pmatrix} v_{s,\omega}^{(u+1)} \\ v_{b,\omega}^{(u+1)} \end{pmatrix} = \begin{pmatrix} v_{s,\omega}^{(u)} \\ v_{b,\omega}^{(u)} \end{pmatrix} - H_{\omega}^{-1}(v_{s,\omega}^{(u)}, v_{b,\omega}^{(u)}) \nabla(v_{s,\omega}^{(u)}, v_{b,\omega}^{(u)}).$$

- Iterate the algorithm until convergence.

The final solution from the above algorithm is the global maximizer of $\ell(v_s, v_b, \omega)$ for a given ω . It is the conditional MLE of v_s and v_b given ω . This is useful in the computation of the unconditional MLE $(\hat{v}_s, \hat{v}_b, \hat{\omega})$ of (v_s, v_b, ω) .

Because the $\ell(v_s, v_b, \omega)$ is not concave as functions of v_s, v_b and ω , the big concern is convergence of the Newton–Raphson method. Therefore, we discard this method and propose a bisection algorithm for \hat{v}_s, \hat{v}_b and $\hat{\omega}$ below. In this algorithm, we assume $\omega \in D$ with D covered by a three dimension rectangle W , as

$$D \subseteq W = \{\omega : C_j - h_j \leq \omega_j \leq C_j + h_j, j = 1, 2, 3\}.$$

The following algorithm is able to solve the global maximum of v_s, v_b and ω for any $\omega \in D$.

Algorithm for \hat{v}_s, \hat{v}_b and $\hat{\omega}$.

- Let $h_j^{(0)} = h_j$ for $j = 1, 2, 3$ and $\omega^{(0)} = (\omega_1^{(0)}, \omega_2^{(0)}, \omega_3^{(0)}) = (C_1, C_2, C_3)$.
- Define $\omega_{k_1 k_2 k_3}^{(u)} = (\omega_{1,k_1 k_2 k_3}^{(u)}, \omega_{2,k_1 k_2 k_3}^{(u)}, \omega_{3,k_1 k_2 k_3}^{(u)})$ with

$$\omega_{j,k_1 k_2, k_3}^{(u)} = \omega_j^{(u)} + \frac{(k_j - 3)h_j^{(u)}}{2}, k_1, k_2, k_3 = 1, 2, 3, 4, 5.$$

Then, $\{\omega_{k_1 k_2 k_3}^{(u)} : k_1, k_2, k_3 = 1, 2, 3, 4, 5\}$ is composed of a $5 \times 5 \times 5$ lattice centered at $\omega_{3,3,3}^{(u)}$ with side unit increment $h_j^{(u)}/2$ for $j = 1, 2, 3$.

- Compute $\hat{v}_{s,\omega}$ and $\hat{v}_{b,\omega}$ by letting $\omega = \omega_{k_1 k_2 k_3}^{(u)}$ for all $k_1, k_2, k_3 = 1, 2, 3, 4, 5$, and calculate all $\ell(v_{s,\omega_{k_1 k_2 k_3}^{(u)}}, v_{b,\omega_{k_1 k_2 k_3}^{(u)}})$ values. Let $\omega_{k_{m_1} k_{m_2} k_{m_3}}^{(u)}$ be the global maximizer for all of those $\omega = \omega_{k_1 k_2 k_3}^{(u)}$.
- Let $\omega^{(u+1)} = \omega_{k_{m_1} k_{m_2} k_{m_3}}^{(u)}$. If $1 < k_{m_1}, k_{m_2}, k_{m_3} < 5$, then let $h_j^{(u+1)} = h_j^{(u)}/2$; otherwise let $h_j^{(u+1)} = h_j^{(u)}$. Iterate this algorithm until convergence.

The above algorithm does not need the concavity of $\ell(v_s, v_b, \omega)$. It starts from an initial rectangle and then reduces the volume of the rectangle to $1/8$ at each step. Therefore, it definitely converges. If the final result claims $v_s < 0$, then $\hat{v}_s = 0$ which induces $\hat{v}_b = \hat{v}_{b,0}$. In this case $\Lambda = 0$. Otherwise, Λ is positive. This algorithm is used in the computation of the the bootstrap p -value of Λ .

Let $y_+ = \sum_{i=1}^m y_i$ and $\lambda_+ = \sum_{i=1}^m \lambda_i$. Then under the null hypothesis, we have

$$(y_1, \dots, y_m) | y_+ \sim \text{Multinomial}(y_+, (\frac{\xi_1}{\sum_{i=1}^m \xi_i}, \dots, \frac{\xi_m}{\sum_{i=1}^m \xi_i})). \quad (31.9)$$

Therefore, the algorithm for the bootstrap p -value is given as follows.

Bootstrap method for the p - value of Λ

- (i) Compute the observed value of Λ based on the observed counts y_1, \dots, y_m . Let it be Λ_0 .
- (ii) Generate K independent random samples from Model (31.9). Denote the k th sample as $(y_{1,k}, \dots, y_{m,k})$ with $k = 1, \dots, K$.
- (iii) Compute the value of Λ based on each generated $(y_{1,k}, \dots, y_{m,k})$. Let it be Λ_k .
- (iv) The bootstrap p -value of Λ is derived by $\#\{\Lambda_k \geq \Lambda_0 : k = 0, \dots, K\} / (K + 1)$, where $\#(S)$ is the number of elements contained in set S .

We choose $K = 999$ so that the bootstrap p -value is given by 0.001 increment. This p -value of Λ can be used to test the significance of the radioactive target. The null hypothesis will be rejected if the p -value is less than the significance level (e.g. 0.05).

31.4 Simulation Results

We evaluated the detection method by the behavior of its power function. We also evaluated the location method by the behavior of the mean square error. The evaluation was based on Monte Carlo simulations. In order to save the computational time, we assumed the third dimension of the deployed sensors were 0 and the hidden nuclear radioactive target was also installed on the plane of the deployed sensors. That is, we assumed $\omega_3 = a_{i3} = 0$ for all i .

In our simulation, we assumed the WSN had 100 radiation sensors. They were identical and deployed at the 10×10 lattice. Assume the radioactive target was installed at $(5.5, 5.5)$. Consequently, the distance between the radioactive target and the i detector is $r_i = [(a_{i1} - 0.5)^2 + (a_{i2} - 0.5)^2]^{1/2}$ with (a_{i1}, a_{i2}) be the i th lattice points of the 10×10 lattice. The radiation count y_i received by the i th detector then has the distribution

$$y_i \sim \text{Poisson}(T_{\xi_i}^{\xi}(v_b + \frac{v_s}{4\pi r_i^2})), i = 1, \dots, 100.$$

Table 31.1 Power function of Λ and MSE of $\hat{\omega}$

	v_s								
	0	0.1	0.2	0.4	0.5	0.6	0.7	0.8	0.9
Power function of Λ	0.049	0.073	0.170	0.443	0.746	0.992	0.986	1.000	1.000
MSE of $\hat{\omega}$	1.475	0.682	0.395	0.124	0.044	0.015	0.012	0.009	0.007

We assumed the sensors were identical so that ξ_i were all the same. To make our simulation simple, we assumed $\xi_i = 1$ for all i .

We fixed $v_b = 1$ and chose $T = 1000$. We let v_s vary from 0 to 0.8 with step increment 0.1. We computed the power function of Λ and derived the MSE of $\hat{\omega}$ based on 1000 simulation repetitions. The power function was evaluated by the percentage of the significance based on significance level $\alpha = 0.05$. The MSE of $\hat{\omega}$ was derived by the average of $\|\hat{\omega} - \omega\|^2$. The simulation result is given in Table 31.1.

Table 31.1 showed that the power function of Λ increased to 1 as v_s became stronger. The type I error probability was displayed by the case when $v_s = 0$. As we expected, the power function was close to 0.05 when $v_s = 0$. In addition, the table also showed that the MSE of $\hat{\omega}$ approached to 0 as v_s became stronger, which indicated the radiation signal can be successfully located as T became large. In our simulation, we have found that the numerical results were almost identically if $\sqrt{T}v_s/v_b$ kept constant. Therefore, Table 31.1 represented a group of numerical simulation scenarios.

31.5 Conclusion

In this chapter, we have proposed a statistical method as well as the algorithm for nuclear radiation target detection based on value fusion data of WSN. Comparing with methods using single detectors, our method can simultaneously detect and locate nuclear radioactive targets. Numerical results based on Monte Carlo simulations showed that our algorithm could successfully fulfill our tasks. Even though our simulation only considered a particular set out of WSN, the idea presented by this chapter will be extensively used to real world nuclear safety and security problems.

Acknowledgments This research was supported by the United States National Science Foundation Grant SES-07-52657.

References

1. Lewis FL (2004) Wireless sensor networks. smart environments: technologies, protocols, and applications. (eds) Cook DJ, Das SK Wiley, New York
2. Elson J, Estrin D Sensor networks: a bridge to the physical world. In: Raghavendra CS, Sivalingam M, Znati T (eds) by wireless sensor network. Kluwer Academic, Norwell (2004)
3. Chakrabarty K, Iyengar S, Qi H, Cho E (2002) Grid coverage for surveillance and target location in distributed sensor networks. *IEEE Trans Comput* 51:1448–1453
4. Clouqueur T, Phipatanasuphorn V, Saluja K, Ramanathan R (2003) Sensor deployment strategy for detection of targets traversing a region. *Mob Netw Appl* 28:453–461
5. Katenka N, Levina E, Michailidis G (2008) Local vote decision fusion for target detection in wireless sensor networks. *IEEE Trans Sign Proc* 56:329–338
6. Clouqueur T, Ramanathan P, Saluja K, and Wang K (2001) value-fusion versus decision fusion for fault-tolerance in collaborative target detection in sensor networks. In: Proceeding of 4th annual conference on information fusion, pp 25–30
7. Brennan SM, Maccabe AB, Mielke A.M. and Torney DC (2004) D.C. radiation detection with distributed sensor networks. *IEEE Computer* 37:(8)
8. Chong C, Kumar S (2003) Sensor networks: evolution, opportunities, and challenges. In: Proceedings of IEEE, August
9. Brooks RR, Iyengar SS (1998) Multi-sensor fusion: fundamentals and applications with software. Prentice Hall, New Jersey
10. Fetter S, Cochran TB (1990) Gamma ray measurements of a Soviet cruise missile warhead. *Sci* 248:828–834
11. Richelson J (2002) Defending nuclear terror. *Bull At Sci* 58:38–43
12. Fetter S, Frolov VA, Miller M, Mozley R, Prilutsky QF, Rodionov SN, Sagdeev RZ (1990) Detection nuclear warheads. *Sci Glob Secur* 1:225–302
13. Feldman GJ, Cousins R (1998) Unified approach to the classical statistical analysis of small signals. *Phys Rev* 57:3873–3889
14. Andrews DW, Ploberger W (1995) Admissibility of the likelihood ratio test when a nuisance parameter is present only under the alternative. *Annals of Statistics* 23:1609–1629
15. van der Vaart AW (1998) Asymptotic statistics. Cambridge University Press, UK
16. Davies R (1977) Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64:247–254
17. Wan H, Zhang T, and Zhu Y (2010) Detecting and locating hidden radioactive sources with spatial statistical methods. *Ann Oper Res* (in press)

Chapter 32

Optimization of Air Network Applied in the Express Based on Hub-and-Spoke Network: A Case Study of SF Express

Peng Jianliang, Si Jiandong and Bao Fuguang

Abstract The air network applied in express based on Hub-and-spoke network contributes to fully giving scope to the advantage of competitiveness and enlarging the market. In this paper, we choose SF express as the example and get the data of portfolio and the distance of two different cities through the strict survey. Based on the data, we build the simplified model of Hub-and-spoke network, determine the district and cities which are covered by the air hubs then plan the air lines and scheduled flights. The conclusion of study proves that Hub-and-spoke network can make full use of air resource, improve the load rate of the plane and the coverage.

Keywords Hub-and-spoke network · Express · Air network · Optimization

32.1 Introduction

Chinese freight air transportation has a rapid development in recent years. Each of the domestic airline company and famous expresses establish freight airline company. As one of the most important steps of the express transportation, the development of the air network is closely related to the development of the whole

P. Jianliang (✉) · S. Jiandong · B. Fuguang
College of Computer Science & Information Engineering, Zhejiang Gongshang
University, Hangzhou, 310018, CHINA
e-mail: pengjl@mail.zjgsu.edu.cn

S. Jiandong
e-mail: sijiandongwu@163.com

B. Fuguang
e-mail: baofuguang@126.com

express industry. When the foreign famous expresses start to occupy the Chinese express market, how to reasonably plan the air network, improve the timeliness of express delivery, reduce the cost of operation, enhance the customer service level become the most important problems which are eagerly need to be resolved. Traditional freight air network is mostly based on the conditions that the portfolio of between the two cities up to standard or the long-term accumulated experience of the related staffs. Obviously, this method is filled with many disadvantages and inefficiencies [1]. The built of Hub-and-spoke network contributes to fully giving scope to the advantage of competitiveness and enlarges the market [2].

In the Hub-and-spoke network, the airport of the panel point can be divided into hub and non-hub airport. It is good to save cost and to make full use of resources that configure all the resources on the basis of the function and the scope [3]. To the SF express, there are advantages as follows: (1) Building fully automatic sorting center on the hub, which leads to concentrated sorting, is good to improve sorting efficiency and reduce the error rate; (2) It will lead to the greatly increase on the portfolio of the hub to built the air hub, which urges the resource of the hub-airport to be scale; (3) It is good to the “density economy” [4]; With the improvement of the number of flights or the load rate of the plane, costs are coming down.

On the basis of reading lots of related documents, this paper takes the example of Chinese SF express limited company (calling SF express as follows). With the rapid development of SF express, we consider to use the Hub-and-spoke network to optimize the express air network.

32.2 The Model of Hub-and Spoke Network

32.2.1 Put Forward the Hub-and-Spoke Network

Hub-and-spoke network is also called concentrate-star structure [5] or axes-spoke network [6]. “Hub” is the special point used to strengthen the connection with other non-center point, there are some points usually. “Spoke” is the connection of the non-center point with the hub. The connection between non-center points is relied on the hub. The kind of air network connect the hub-airport and non hub-airport through the hub-city, and make a strict flight plan in order to improve the coverage, which finally lead to save the cost and improve the transportation rate.

32.2.2 Model Assume

Depend on the conditions raised by the model and the fact of the express’s air transportation, we come up with the model assume [7] as follows: (1) Hubs connect with each other directly. (2) There exists scale effect and all the hubs have

the same discount ratio. (3) Non-hubs cannot be connected with each other directly. (4) The distance of two cities meets the requirements of triangle non-equality. (5) The transport costs are same on the unit distance no matter the goods are different. (6) The distance of two cities is replaced by the direct distance of the two points. (7) The portfolio and transport costs are symmetrical.

32.2.3 Built Model

Ernst and Krishnamurthy [8] put forward the simplified model of Hub-and-spoke network [8]. Depending on the definition of Hub-and-spoke network, the goods need to be change trains through the hubs. When building the model, this paper conducts the transport cost as the objective function. When the objective function is minimum, the best scheme is coming out. The model is as follows:

$$MIN \sum_i \sum_k C_{ik} X_{ik} (O_i + D_i) + \sum_i \sum_k \sum_l \alpha C_{kl} Y_{kl}^i, \quad (32.1)$$

$$\sum_{k \in n} X_{kl} = 1, \forall i \in n, k \in H. \quad (32.2)$$

$$X_{ik} \in \{0, 1\}, \forall i \in n, k \in H. \quad (32.3)$$

$$\sum_l Y_{kl}^i - \sum_l Y_{lk}^i = O_i X_{ik} - \sum_j W_{ij} X_{jk}, \forall i \in n, k, l \in H, l \neq k. \quad (32.4)$$

$$Y_{kl}^i \geq 0, \forall i \in n, k, l \in H \quad (32.5)$$

In the model:

$$X_{ik} = \begin{cases} 1, & \text{non-hub } (i) \text{ connect with hub } (k) \\ 0, & \text{else} \end{cases} \quad (32.6)$$

In the model, C^{ik} means the transport costs from the point i to the point k . We express the path from the point i to the point k with $\bar{Y} = \{Y_{kl}^i, i \in n, k, l \in H, k \neq l\}$, n means the collection of non-hubs, H means the collection of hubs, Y_{kl}^i means the path is from i to another hub through one hub. O_i and D_i means the total quantity of flowing in and out from i , which was defined as $O_i = \sum_{j \in n} W_{ij}$, $D_i = \sum_{j \in n} W_{ji}$, the W_{ij} means the quantity of flow is from i to j . The objective function (32.1) searches for the minimum cost, including gathering cost, transport cost and delivery cost. Constraint condition (32.2) makes any point connects with a certain hub. Constraint condition (32.3) means all the variables are integers. Constraint condition (32.4) means, to any i , $Y_{kl}^i > 0$ or $Y_{lk}^i > 0$, but not to be zero at the same time. [As the Constraint condition (32.2) requires that i can only be connected with one hub $k, k \neq l$].

Table 32.1 The portfolio and the distance of any two cities

City	BJ	SH	CQ	SY	WH	CD	XA
BJ		715	1179	782	491	1402	890
SH	20425		1668	652	789	1919	1426
CQ	750	589		1958	879	254	294
SY	6190	3073	519		1164	2184	1671
WH	2822	1827	520	464		1131	649
CD	3036	1059		604	662		513
XA	1760	815	140	382	266	454	
ZZ	1726	1106	290	465	536	445	310
WX	16331		1697	5880	3129	4807	2364
HZ	23857		4524	9332	4366	9145	4773
FZ	7525		693	1894	1480	1922	855
CS	1470	799	144	265		628	171
SZ	53138	44075	6925	14984	11754	13746	5754
City	ZZ	WX	HZ	FZ	CS	SZ	
BJ	396	610	650	767	642	863	
SH	892	125	144	368	935	980	
CQ	813	1544	1533	1437	744	837	
SY	1152	684	749	995	1332	1477	
WH	204	665	654	593	168	373	
CD	1052	1794	1785	1692	999	1084	
XA	543	1300	1298	1236	550	722	
ZZ		767	773	762	283	539	
WX	2711		68	312	814	873	
HZ	4088			246	795	835	
FZ	1367				695	663	
CS	221	686	1105	478		261	
SZ	9673	48691	63753		9992		

The bottom left of the table expresses the portfolio (unit: kg) and the upper right expresses the distance (unit: km)

From the model, we would know that X_{ik} is the most important assigned vector, so the answer of the model is assured by X_{ik} . Obviously, this model is a NP question and can not be solved by simple linear programming. Therefore we may use genetic algorithm to get the relatively best answer.

32.3 The Application of Hub-and-Spoke Network on the Chinese Air Network

32.3.1 The Optimization of Hub-and-spoke Network in Express Air Network

SF express is mainly doing the business of express at home and abroad. Recent years, it built the network including southern, northern, eastern, central part of

Table 32.2 The result of planning the air network

Hub	Relative city
Beijing	Shenyang, Weifang, Wulumuqi, The district of Bohai Sea
Hangzhou	Shanghai, Wuxi, Yangtze River Delta
Shenzhen	Fuzhou, Pearl River Delta
Wuhan	Chongqin, Chengdu, Xi'an, Zhengzhou, Changsha, Zunyi, Nuijiang

Fig. 32.1 The result of planning the air network shows on map

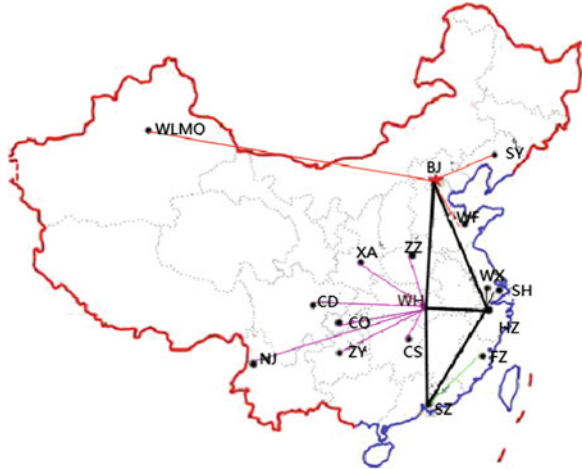


Table 32.3 The planning of scheduled flight in northern China, Beijing

Cargo origin	Destination	Plane type	Specified capacity	Average daily portfolio	Load rate (%)
Beijing	Wuhan	B737	14,000	14,004	100
	Wuhan	B757	28,000	28,400	104
	Hangzhou	A300	40,000	34,233	85.58
	Shenzhen	A300	40,000	35,670	89.175
	Shenzhen	B737	14,000	10,000	71.42
	Shenyang	B757	28,000	28,000	100
	Shenyang	B757	28,000	26,613	95.04
	Weifang	A300	40,000	40,000	100
	Weifang	A300	40,000	39,271	98.117
Shenyang	Beijing	B757	28,000	26,112	93.25
Weifang	Beijing	B737	14,000	13,783	98.45
	Beijing	A300	40,000	40,000	100
Average Load Rate					95

China. However, with the development of the portfolio, the now available air network is not suitable, it needs to break through. So we apply the Hub-and-spoke network to plan the SF express air network again. Based on the data from the strict survey, we get the portfolio and the distance of any two cities in Table 32.1.

Table 32.4 The planning of scheduled flight in southern China, Shenzhen

Cargo origin	Destination	Plane type	Specified capacity	Average daily portfolio	Load rate (%)
Shenzhen	Beijing	A300	40,000	40,000	100
	Beijing	A300	40,000	40,000	100
	Beijing	B757	28,000	24,906	88.95
	Wuhan	A300	40,000	40,150	100.3
	Wuhan	A300	40,000	40,150	100.3
	Wuhan	B757	28,000	28,201	100.7
	Hangzhou	A300	40,000	36,000	90
	Hangzhou	A300	40,000	36,519	91.30
Fuzhou	Fuzhou	B757	28,000	16,359	58.42
Fuzhou	Shenzhen	A300	40,000	30,062	75.15
Average load rate					91

Table 32.5 The planning of scheduled flight in central part of China—Wuhan

Cargo origin	Destination	Plane type	Specified capacity	Average daily portfolio	Cargo origin (%)
Wuhan	Beijing	B757	28,000	24,088	86.03
	Hangzhou	A300	40,000	36,000	90
	Hangzhou	B737	14,000	9,978	71.27
	Shenzhen	A300	40,000	38,633	96.58
	Chongqi	B757	28,000	20,830	74.39
	Chengdu	A300	40,000	34,000	85
	Chengdu	B737	14,000	9,825	70.18
	Xi'an	B757	28,000	22,496	80.34
	Zhengzhou	B757	28,000	27,159	96.7
	Changsha	B757	28,000	22,526	80.45
	Nujiang	A300	40,000	36,000	90
	Nujiang	B737	14,000	9,242	66.01
	Zunyi	A300	40,000	36,000	90
	Zunyi	B737	14,000	13,724	98.02
	Chongqin	Wuhan	B737	14,000	6,416
Chengdu	Wuhan	B737	14,000	14,000	100
Chengdu	Wuhan	Scattered Navigation		1,825	
Xi'an	Wuhan	B737	14,000	9,175	65.53
Zhengzhou	Wuhan	B737	14,000	12,114	86.53
Changsha	Wuhan	B737	14,000	11,483	82.02
Zunyi	Wuhan	A300	40,000	34,350	85.87
Nujiang	Wuhan	B757	28,000	27,008	96.45
Average Load Rate					82.72

Table 32.6 The planning of scheduled flight in eastern China—Hangzhou

Cargo origin	Destination	Plane type	Specified capacity	Average daily portfolio	Cargo origin (%)	
Hangzhou	Beijing	B757	28,000	27,000	96.43	
	Beijing	B757	28,000	27,734	99.05	
	Wuhan	B757	28,000	28,000	100	
	Wuhan	B757	28,000	28,000	100	
	Wuhan	B757	28,000	27,886	99.59	
	Shenzhen	A300	40,000	32,000	80	
	Shenzhen	A300	40,000	39,419	98.54	
	Shanghai	B757	28,000	28,000	100	
	Shanghai	B757	28,000	28,000	100	
	Shanghai	B737	14,000	11,412	81.51	
	Wuxi	B757	28,000	27,000	96.42	
	Wuxi	B757	28,000	26,690	95.32	
	Shanghai	Hangzhou	B757	28,000	28,000	100
		Hangzhou	B757	28,000	28,000	100
Hangzhou		B737	14,000	13,160	94	
Hangzhou	Hangzhou	B757	28,000	28,000	100	
	Hangzhou	B757	28,000	28,000	100	
	Hangzhou	B737	14,000	9,072	64.8	
Average Load Rate					95.03	

According to the data of Table 32.1, replace the C^{ik} with the direct distance between two cities and calculate. Then the answer is

$$X^T = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{32.7}$$

After analyzing, the final result shows in Table 32.2 and Fig. 32.1.

With the portfolio up to the standard of direct flight, the non-stop flight can save the most costs. In the fact, Shanghai to Beijing, Shanghai to Shenzhen, Wuxi to Beijing, Wuxi to Shenzhen, and these four lines is better to choose non-stop flight. Therefore, we can calculate again except the portfolio on the four lines.

32.3.2 Measure the Real Effectiveness of the Hub-and-Spoke Network

According to the Hub-and-spoke express air network after optimization, we can get the final scheduled flight, shows in Tables 32.3, 32.4 and 32.5.

Planning the flights on the base of Hub-and-spoke network can make full use of air resource and improve the load rate of the plane and the coverage. From the result we can know that the average load rate is up to 90 percent, higher than 88% in SF express.

SF now has air line among 9 cities, whereas other cities are scattered navigation. After optimization, the coverage is up to 94.3%, higher than 52.9% in SF express.

32.4 Conclusion

This paper applies the Hub-and-spoke network for planning the air network on the basis of SF express. Obtain the final result after modeling and calculating. It is obvious that the coverage and load rate is higher than the current situation.

Acknowledgment This research is supported by Ministry of Education's Social Sciences and Humanities Research Project (10YJC630018) and Zhejiang Provincial University Students Scientific Innovative Project.

References

1. O'Kelly ME (1998) A geographer's analysis of Hub-and-spoke network. *J Transp Geogr* 6:171–186
2. Giovanni N (1999) A note on competitive advantage of large Hub-and-spoke network. *Transp Res* 35:225–239
3. Ding L-L (2008) Deregulation and the improvement of Hub-and-spoke network of civil aviation transportation industry. *J Harbin Univ Comm* 5:118–120
4. Jin F-J, Wang C-J (2009) Hub-and-spoke system and China aviation network organization. *Geogr Res* 24:774–784
5. Zhong S (2003) The supply of concentrate-star structure in the air transportation. *Transp Syst Eng Inf* 3:9–14
6. Zhang H.-M (2000) The operating condition of Axes-spoke network. *Int Aviat* 21:37–39
7. Yang H-Y (2010) The application of Hub-and-spoke network in Chinese aeronautic transportation. *J Beijing Inst Technol* 12:27–30
8. Ernst AT, Krishnamurthy M (1996) Efficient algorithms for the incapacitated single allocation-hub median problem. *Locat Sci* 4:139–154

Chapter 33

Verification and Analysis for Ethernet Protocols with NuSMV

Yujia Ge, Xiaofei Feng and Fangcheng Tang

Abstract Verifying the correctness of computer systems is becoming necessary for most of systems, especially safety–critical systems. Symbolic model checking is one of the effective ways to do automatic model checking. In this chapter, we verified some properties of Ethernet protocols, especially CSMA/CD protocol, using NuSMV tool. NuSMV is proved to be a suitable tool for its input a transition based model of processes by communicating to each other through shared and global variables. The methods we proposed to analyze Ethernet protocols, such as synchronization modeling and time delay modeling, can also be extended to analyze similar related networking protocols and their properties.

Keywords Model checkin · Protocol analysis · NuSMV · CTL

33.1 Introduction

Verifying the correctness of computer systems is becoming necessary for most of systems, especially safety–critical systems. Model checking is an automatic approach which is a model-based and property-verification method [1]. There are

Y. Ge (✉) · X. Feng

College of Computer and Information Engineering, Zhejiang Gongshang University,
No.18 Xuezheng Str., Hangzhou, China
e-mail: yge@mail.zjgsu.edu.cn

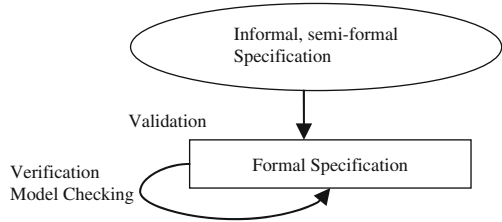
X. Feng

e-mail: xffeng@mail.zjgsu.edu.cn

F. Tang

Department of Computer Science, Iowa State University, Ames, IA, USA
e-mail: tfc@cs.iastate.edu

Fig. 33.1 V&V using formal specification



several different kinds of model checking. Symbolic model checking is one effective way to do it. In this chapter, we verify some internet protocols, such as IEEE Ethernet CSMA/CD protocol using NuSMV tool.

NuSMV [2] is a symbolic model checker originated from the re-engineering, re-implementation and extension of “symbolic model checker” (SMV) which is the original BDD-based model checker developed at Carnegie Mellon University. SMV provides a language for describing the models and it directly checks the validity of CTL formulas. NuSMV allows for the representation of synchronous and asynchronous finite state systems, and for the analysis of specifications expressed in computation tree logic (CTL) and linear temporal logic (LTL), using BDD-based and SAT-based model checking techniques [3]. NuSMV is an effective method to analysis network protocol by model checking. Some research works have been done as in [4], a sliding window protocol is validated, and as in [5], security protocols are validated and analyzed.

Usually using formal methods to validate and verify a system follows the process in Fig. 33.1. We use NuSMV to both check the properties on the formal specification and to do the validation part partially. The chapter describes our method to validate CSMA/CD protocol by using NuSMV. Section 33.2 first gives a formal description of CSMA/CD protocol and specifies our way to model the protocol using NuSMV language. Section 33.3 shows the property verification results. Section 33.4 discusses the traces and some simulation issues in validation. Section 33.5 is our conclusion.

33.2 Modeling CSMA/CD Protocol with NuSMV

We choose CSMA/CD protocol for its wide applications. The basic idea of pure carrier sense multiple access (CSMA) [6] is when a station has data to transmit, the station first listens to the cable to see if a carrier (signal) is being transmitted by another node. This may be achieved by monitoring whether a current is flowing in the cable, but it cannot prevent data collision because two stations can start transmitting data at the same time. Carrier sense multiple access with collision detection (CSMA/CD) is a modification of pure CSMA. If a transmitting data station detects a collision while transmitting a frame, it will stop transmitting that

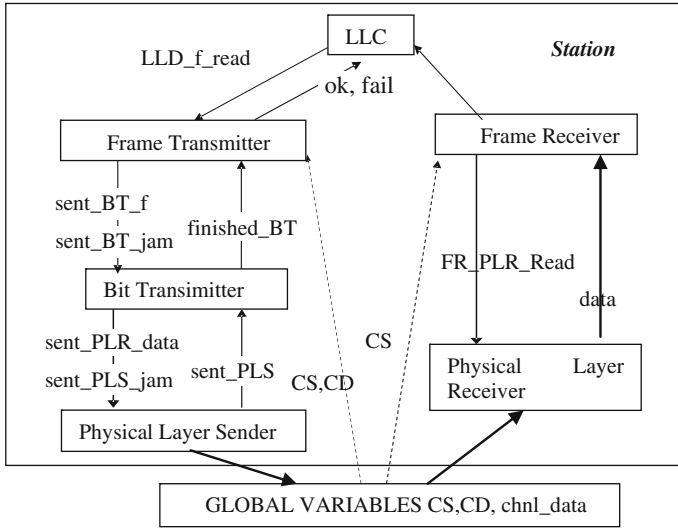


Fig. 33.2 Formal model of CSMA/CD protocol

frame and transmit a jam signal. Then it waits for a random delay before trying to send that frame again. Collision detection is used to improve CSMA performance by terminating transmission as soon as a collision is detected, and reducing the probability of a second collision on retry. The algorithm of choosing the delay is called “Binary Exponential Backoff”. This mechanism ensures that the delay time adapts to the number of stations involved on collision and collision is resolved in a reasonable interval if many stations collide. The number of slots for which a station is going to wait depends on a random number selected by the station. After the first collision each station waits for 0 or 1 slot times before trying again. After the second collision the stations wait for a random period of between 0, 1, 2 or 3 slot times and try again. In general, after i th collision, a random number between 0 to $2i-1$ is chosen. When $i \geq 16$, the value is always 0–1023.

33.2.1 Description of Formal Specification

According to a formal specification of the CSMA/CD protocol [7], each station consists of a set of processes communicating through shared variables. Each process is modeled as a timed transition system with upper and lower bounds on each transition. Fig. 33.2 shows the model of CSMA/CD protocol.

The following code shows the module for each station:

```
MODULE station(CS, CD, chnl_data, ticks)
VAR
```

```

LLC_f_ready: boolean;
ft: FT (LLC_f_ready, bt.finished_BT, CS, CD, ticks);
bt: BT (ft.sig_FT_BT, pls.sent_PLS);
pls: PLS(bt.sig_BT_PLS, bt.next_bit, CD, ft.delay);
fr: FR(CS);
plr: PLR(fr.FR_PLR_read, chnl_data);

```

ASSIGN

```

LLC_f_ready : = 1; --LLC_f_ready always is true

```

The five processes in each station are described in detail as below:

1. Frame Transmitter (FT) The process takes input from the variable `LLC_f_ready` (always is 1 in our model) to indicate that the frame is ready for transmission. It checks variable `CS` to see if the carrier is busy. If it's not busy and the delay is 0, it sends a `send_BT_f` signal to the bit transmitter (BT). If BT is not done with sending data and a collision is detected (`!finished_BT & CD = true`), it sends a `send_BT_jam` signal to the BT. If it received the `finished_BT` signal, it sends an `ok` signal to the LLC indicating that the data transmission has been successful (`sig_FT_LLC = ok`). If there is a collision when send data to BT (`sig_FT_BT = send_BT_f & CD`), set the delay to a random number according to the binary exponential backoff algorithm. Here we use approximate algorithm to generate the delay number. After waiting for the amount of delay time, the process continues to try to send data to BT (`send_BT_f`) when the channel is not busy. When it reaches maximum number of attempt = 4, we consider it as failure.

2. Bit Transmitter (BT) This process is the link between the MAC layer and the processes in the physical layer. If BT receives the signal `send_BT_f` from FT (`sig_FT_BT = send_BT_f`), it will send the `send_PLS_data` signal to the physical layer sender (PLS) to transmit the next bit from the data frame. If it receives a `send_BT_jam` signal, it indicates to the PLS to send the jam data by sending a `send_PLS_jam` signal to PLS. It sends `finished_BT` to FT when jam data is sent or all bits in one frame is sent. The process keeps track of the number of bits sent and the index of the next bit to be sent. We suppose the bit number of one frame = 3.

3. Physical Layer Sender (PLS) When receiving the signal from the BT to transmit data bit, PLS checks if the data to be transmitted is a jam bit or a data bit (`send_BT_PLS = send_PLS_data` or `send_BT_PLS = send_PLS_jam`). If it's data bit, `DATA = 0,1,2`. After the last data bit, an additional bit containing the value ND (no data) is sent. If it is jam bit, `DATA = J`. After sending a bit successfully, send a signal `sent_PLS` to BT.

4. Frame Receiver (FR) This is the data receiver process of the MAC layer. It waits for the variable `CD = 1` and then communicates with the physical layer receiver (PLR), set `FR_PLR_read = 1` and receives the data bits.

5. Physical Layer Receiver (PLR) This is the receiver section of the physical layer. It receives data from the data channel (`chnl_data`). When it receives a signal from the FR to read the data (`FR_PLR_read = true`), it reads data and store it.

To model the communication between stations, a set of global variables *CD*, *CS*, *chnl_data* are defined in main module. The details of several aspects for modeling are described as follows:

Data channel modeling. The data channel was one of the complex aspects to model in the transition system. Some assumptions are made in modeling data channel. The most significant one is that the channel is assumed to be just one bit length, as opposed to that in real life as a stream of bits whose bit length depends upon the physical length of the cable. The data channel is assumed as a set of data bit variables with one data bit for each station on the network. Each *DATA* variable is defined as:

```
DATA: {0, 1, 2, J, ND}
```

A value of {0, 1, 2} indicates good data bits. If the value of data variable is J then it indicates that the station is sending jam sequence. The data value of ND indicates that there is no data. The number of data variable depends on the number of stations in this system. In our project, we simulate two-station system, which has two data variable. The actual data on the channel is given by the variable *chnl_data* which is the composition of all the data variables. The definition of *chnl_data* is:

```
chnl_data :=
```

```
  case
DATA1 = ND : DATA2 ;
DATA2 = ND : DATA1 ;
1 : J ;
  esac ;
```

Carrier Sense (CS). The carrier sense (CS) variable which represents the status of the channel is defined as follows:

```
CS := !(DATA1 = ND & DATA2 = ND) ;
```

CS is set true if either station1 or station2 is sending data to channel.

Collision Detection (CD). The collision detection (CD) variable which is set to true if there is collision on the channel. This is defined as:

```
CD := !(DATA1 = ND | DATA2 = ND) | (DATA1 = J | DATA2 = J) ;
```

Random number generation. Since there is no *rand()* function provided by NuSMV, we simulate the process of generating random number for delay by the “random” function $f(x)$:

```
f(x) = 3*x mod 7 ;
```


Initially, let $x = 1$; we have $f(x) = 3$, $f(f(x)) = 2$, $f(f(f(x))) = 6$, and so on. By this way an approximately random sequence: 1, 3, 2, 6, 4, 5, 1, 3, 2 ... is produced.

33.2.2 Synchronous Model and Asynchronous Model

For the synchronous model, each value written by a transmitter in a clock cycle is read by all the receivers in the next clock cycle. No centralized delay monitoring system was needed. Each transmitter decrements its delay in successive clock cycles, and when the delay becomes zero, it tries to transmit again. It simulates the reality, but each step of all the transmitters is synchronized which is not the same as reality.

For the asynchronous model, due to the synchronization requirement for this protocol, we need to implement synchronization mechanism by ourselves. There are two main mechanisms that we did:

1. Synchronization of transmitters and receivers

The data variables are set by the transmitter and the receivers read it from the `chnl_data` variable. In the real life system there is time bound on each of the transitions which occur in the transmitting section and the receiving section. The rate at which the receiver will never read the same data bit twice from the channel and the transmitter will not overwrite the last written data. For modeling all processes as asynchronous, we achieve the synchronization between transmitters and receivers by using an array variable `read` which is defined as:

`read: array 0..1 of boolean`

A transmitter wait for `read[0]` to be true when it wants to write data to channel. If `read[0]` is true, transmitter1 writes data to channel and set `read[0]` to false, and waits for it to be true again to write data again. After a receiver reads data which is written by the transmitter, it sets `read[0]` to true (indicating it has read this data). The transmitter can later write data again.

2. Modeling of delay after collision

The ethernet protocol uses a Binary Exponential Backoff algorithm for handling the post collision arbitration. When two stations get into collision then each of them picks up a random number which depends on the current attempts number for sending data. Each station then waits for an amount of time given by the delay. Clearly, in an asynchronous model, since there is no global clock between stations, wait time at each station cannot achieve to count down by one after each cycle time. Since the station with least delay time will attempt to transmit first in the next try, another variable `go` is introduced to module FT indicating whether this station is qualified to transmit or not:

```

stn1_go: = !(stn1.ft.delay = 0) & !(stn2.ft.delay
= 0) & (stn1.ft.delay <= stn2.ft.delay);

```

The variable *go* for the station one is set to true when both stations' delay is greater than zero and station ones' delay is less than station twos'. The station which gets the *go* signal to be true will go ahead to transmitting data.

33.3 Property Verification

We checked several properties which are critical to the protocol for both synchronous model and asynchronous model. All the properties above are verified as true in NuSMV. In synchronous model, we checked the following 3 properties.

Property 1 *Whenever LLC layer requests the transmission of a frame, the MAC layer (i.e. Frame Transmitter) responds with a success or a failure message.*

```

SPEC AG(stn1.LLC_f_ready = 1 -> AF(stn1.ft.sig_FT_LLC =
ok | stn1.ft.sig_FT_LLC = fail)

```

Property 2 *Whenever there is a collision, both stations reach an ok state or both states reach a fail state.*

```

SPEC AG(CD -> AF((stn1.ft.sig_FT_LLC = ok & stn2.ft.sig_
FT_LLC = ok) | (stn1.ft.sig_FT_LLC = fail & stn2.ft.sig_FT_
LLC = fail)))

```

Property 3 *The frame received by a receiver is the correct frame. That is, when LLC of transmitter receives OK signal, BR of receiver receives a frame's data. To verify both the property, we define a transmitter and a receiver to make clear the role of the station as a receiver (includes FT,BT, PLS) or transmitter (includes FR, PLR).*

```

SPEC AG(recv.plr.PLR_read = 1 -> AF(trans.ft.sig_FT_LLC = ok)

```

In asynchronous model, we checked the following properties.

Property 1 *Each station will eventually reaches a ok state or fail state*

```

SPEC AF(stn1.ft.sig_FT_LLC = ok | stn1.ft.sig_FT_LLC = fail)

```

Property 2 *In case of a collision, the station picking up the lower delay will successfully transmit*

```

SPEC AG(stn1_go -> AF(stn1.ft.sig_FT_LLC = ok)

```

33.4 Traces in NuSMV

Traces provided by NuSMV offer the user the possibility of exploring the possible executions. The functionality could show how our model works. The state trace result produced by NuSMV is illustrated in a more understandable way in Table 33.1.

In the first step, station one and station two try to begin to send data from LLC. In the fourth step, because both stations try to send data to data channel at the same time, data collision happens. In the next step, station one sets the delay to four and station two sets the delay to five. After four steps, the delay of station one becomes one and tries to send data again and sends it to the data channel successfully. In the next step, station two starts to send data too. Data collision happens again. Above analysis shows our model is similar to the real protocol although with some differences.

The difference between a real life system and the model is because the difficulty to model some characteristic using NuSMV. For example, the representation of channel through which the stations communicate consists of a stream of bits. These bits move at the speed of transmission that is 10 mb/s per second. The propagation delay on the channel results in the stream of data bits to be unavailable at the same instant to all the stations on the network. To avoid modeling the propagation delay the data channel is assumed to be one bit long. When modeling a Binary Exponential Backoff algorithm for handling the post collision arbitration, a countdown of the delay variable is used in synchronous model. In asynchronous model, we need to use centralized delay monitoring mechanism for achieving the effect. Because of the synchronized mode in NuSMV, the synchronized model is a little different from the real life that each step of all the transmitters is synchronized.

33.5 Conclusion

In this chapter, we verified some properties of ethernet protocols, especially CSMA/CD protocol. NuSMV is proved to be a suitable tool for its input a transition based model of processes by communicating to each other through shared and global variables. The methods we proposed to analyze ethernet protocols, such as synchronization modeling and time delay modeling, can also be extended to analyze similar related networking protocols and their properties. We proved that it is suited to verify networking protocols, but there are some works left:

1. We show the correctness of model partially and informally by showing the execution of our model. More formal validation should be further done.
2. When modeling the protocol, we need to deal with some difficulties, such as how to describe the random number, how to synchronize the system. Those

Table 33.1 Results of traces with 15 steps

Time	frame transmitter	Bit transmitter	Physical layer sender	Data channel	CS	CD
1	(Station1&2) Received frame ready signal from LLC and send signal to BT					
2		(Station1&2) Received signal from FT and send signal to PLS				
3			(Station1&2) Received signal from FT and send data			
4					1	1
5	(Station1&2) Collision detection and send jam signal to BT				1	1
	Sm1.delay = 4					
	Sm2.delay = 5					
6	Sm1.delay = 3	(Station1&2) Send jam signal to PLS		Jam	1	1
	Sm2.delay = 4					
7	Sm1.delay = 2					
	Sm2.delay = 3					
8	Sm1.delay = 1					
	Sm2.delay = 2					
9	Sm1.delay = 0					
	Sm2.delay = 1					
10	(Station1) send signal to BT					
11	(Station2) send signal to BT	(Station 1) Received signal from FT and send signal to PLS				
		(Station 2) Received signal from FT and send signal to PLS				
12			(Station1) Received signal from BT and send data			
13			(Station2) Received signal from BT and send data	Valid data	1	0
14				Jam	1	1
15	(Station1&2) send jam signal to BT				1	0

difficulties on modeling time-related properties are due to the absence of real time mechanism in the model checker.

Acknowledgments The work was partially supported by the Natural Science Foundation of Zhejiang under Grant No. Y1100824 and Zhejiang Public Interest Project of Technology Bureau under Grand No. 2010C31088.

References

1. Huth M, Ryan M (2000) Logic in computer science, Cambridge University Press, London
2. NuSMV project (2011) <http://nusmv.fbk.eu/index.html>
3. Biere A, Cimatti A, Clarke E, Zhu Y (1999) Symbolic model checking without bdds. In: Proceedings of 5th international conference on tools and algorithms for construction and analysis of systems (TACAS'99), pp. 193–207
4. Zhao Y, Yang Z, Xie J, Liu Q (2009) Formal model and analysis of sliding window protocol based on NuSMV. *J Comput* 4(6):519–526
5. Marrero W, Clarke E, Jha S (1997) Model checking for security protocols. Carnegie Mellon University technical report CM U-SCS 97–139
6. Tanenbaum A (1996) Computer networks 3rd edn. Prentice Hall, New Jersey
7. Weinberg H, Zuck L (1992) Timed ethernet: real-time formal specification of ethernet. In: Proceeding of the. 3rd CONCUR, Lect Notes Comput Sci vol 630

Chapter 34

The Analysis and Modeling on Internet AS-Level Topology Based on K-Core Decomposition

Jun Zhang, Hai Zhao and Bo Yang

Abstract The k -core analysis allows to characterize networks beyond the degree distribution and uncover structural properties and hierarchies due to the specific architecture of the system. The properties related with node coreness were studied firstly. The distribution on coreness of Internet AS-level topology was analyzed with the actual topology measuring data. The linking trend to higher shells of the nodes was analyzed and the linking probability function was induced. Then, a new framework for modeling Internet AS-level topology was proposed with the conclusion drawn and the model was implemented. The proposed model was based on k -core decomposition in order to reveal the hierarchy of the network. Experiment results show that the proposed model produces scale-free random graphs, in the meanwhile, it also exhibits the hierarchy of the network better.

Keywords Complex network · Model · K -core decomposition · Power-law

34.1 Introduction

Internet research requires realistic models to correctly generate Internet-like networks. Internet topology modeling is a complex task, involving network measurement, graph theory, algorithm design, statistics, data mining, visualization and mathematical modeling and other research areas [1]. It is due to the complexity and difficult, a large number of experts were attracted in this field and a

J. Zhang (✉) · H. Zhao · B. Yang
College of Information Science and Engineering, Northeastern University
Shenyang 110819, China
e-mail: zhangjun1@ise.neu.edu.cn

variety of models that are capable of constructing random networks with power-law degree distributions were proposed [2–9]. Among them, preferential attachment [3] and optimized-based construction [6] have become the two major paradigms for explaining the Internet topology. The former theory relies on the principle that each joining node attaches its links to existing nodes with a probability proportional to their current degrees. The main explanation behind this behavior is a premise that new users perceive large-degree nodes as being more “attractive” compared to low-degree nodes. The latter theory models node join as an optimization problem and argues that each joining ISP aims to solve a certain trade-off between the benefit of improved connectivity and the cost of adding new links. The existing evolution theories exhibit certain limitations in the context of the Internet AS-graph. While acceptable in certain cases (such as social networks), preferential attachment is usually too restrictive to realistically model the Internet graph as it bases link formation solely on the degrees of existing nodes and places too much weight.

Recent studies show that Internet topology does not only exhibit power-law degree distributions, but also exhibits evidently hierarchical properties. While the main focus of this paper is to understand the properties related with node coreness, then provides a specific algorithm that can be used to create new graphs and tests them the advantages in exhibiting the hierarchy of Internet.

34.2 *K*-core Decomposition: Main Definitions

Let us consider a graph $G = (V, E)$ of $|V| = n$ vertices and $|E| = e$ edges; a k -core is defined as follows [10]:

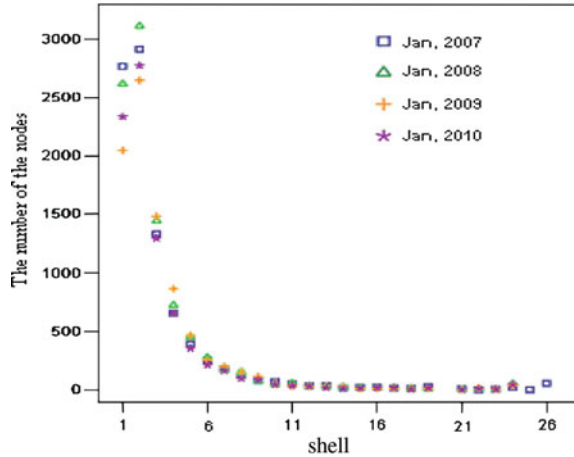
Definition 1 A subgraph $H = (C, E|_C)$ induced by the set $C \subseteq V$ is a k -core or a core of order k iff $\forall v \in C: \text{degree}_H(v) \geq k$, and H is the maximum subgraph with this property [10].

Definition 2 A vertex i has coreness c if it belongs to the c -core but not to $(c + 1)$ -core. We denote by c_i the coreness of vertex i [10]. The highest coreness of the vertex in the graph is called the coreness of the graph.

Definition 3 A shell C_c is composed by all the vertices whose coreness is c . The maximum value c such that C_c is not empty is denoted C_{\max} . The k -core is thus the union of all shells C_c with $c \geq k$ [10].

The k -core decomposition therefore identifies progressively internal cores and decomposes the networks layer by layer, revealing the structure of the different k -shells from the outmost one to the most internal one.

Fig. 34.1 The distribution on coreness of the nodes



34.3 The Distribution on Coreness of the Nodes

In order to describe the distribution on coreness of the nodes, we choose Internet AS-level topology measuring data of CAIDA in January of 2007–2010. Figure 34.1 shows the number of the nodes in every shell. We can see from it that although the period time of the four data sets is long, but their distribution trend is accordant. So we can put them together to analyze. Figure 34.2 shows the percentage of the nodes in each shell. The fitness result is as followed:

$$Y = 0.86 \times X^{-2.319}. \quad (34.1)$$

Because the coreness of Internet AS-level topology in Jan, 2007 is different with the others, and is inconvenient in analyzing with the others together, so the data sets used following are not include the data set of Jan, 2007. In the meanwhile, the nodes in the highest shell and the lowest shell are not included in the data sets. This is because the number of the nodes in the lowest shell is not satisfied the whole tendency which is descendent. In the other hand, because the nodes in the highest shell are very important, a little deviation will affect the whole topology, so we cannot deal them with fitness method.

34.4 The Linking Trend to Higher Shells

The nodes in each shell do not only have links with each other, but also have links with the nodes in different shells. According to the definition of coreness, we can find that the links between the nodes in lower shells and higher shells will be

Fig. 34.2 Distribution of the number of the nodes in different shells

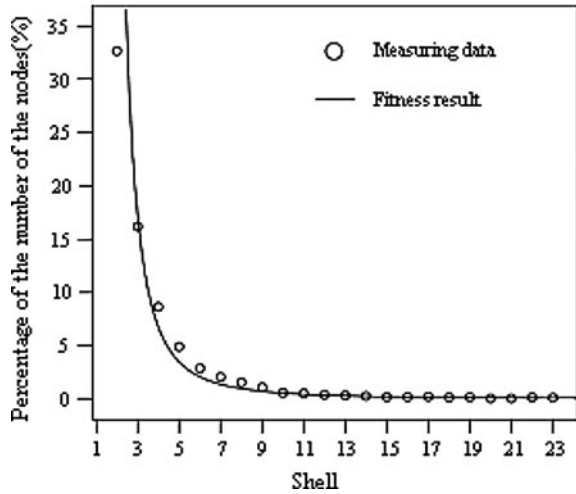
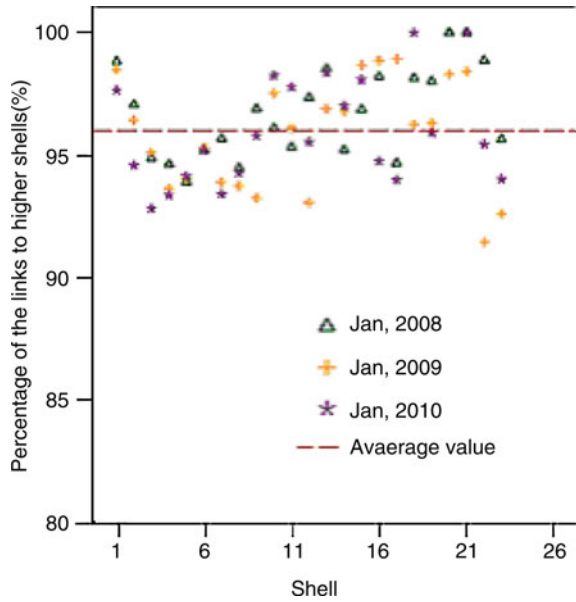


Fig. 34.3 Percentage of the links to higher shells



eliminated in the higher shells. We can consider these links as the links from the nodes in lower shells to the nodes in higher shells in a sense, and are treated in lower shells. Figure 34.3 shows the results of the nodes linking to higher shells. The highest shell is not included in the figure in respect that there is no shell higher than it.

We can see from Fig. 34.3 that more than 90% links in each shell are the links to higher shells, and only a few links are in the same shell. The dashed line in the figure

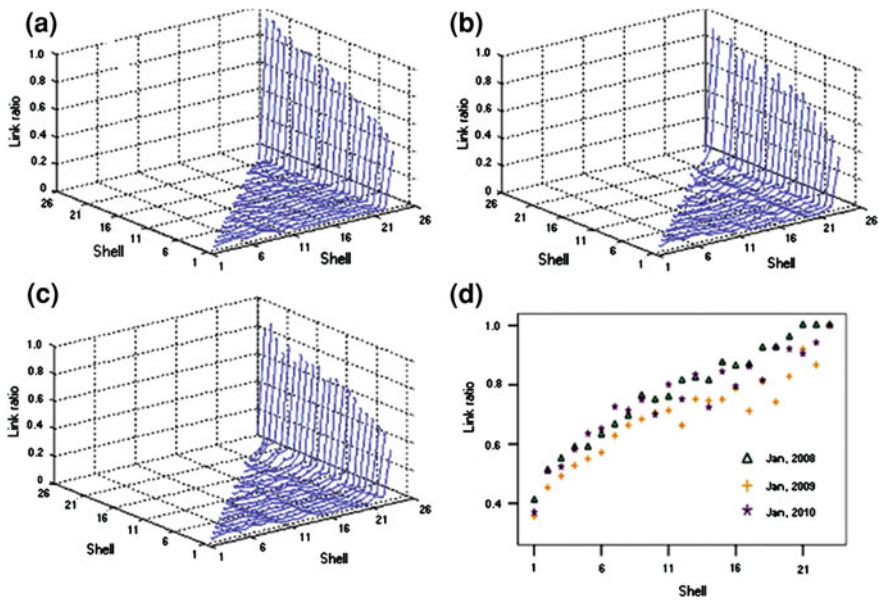


Fig. 34.4 Distribution on links between the nodes in different shells and their higher shells

labels the average value which is about 96.16%. We can see that the difference is little among the data sets.

34.5 The Distribution on Links to Higher Shells

In order to further study the links to higher shells, Fig. 34.4 shows the ratio of the number of the links to higher shells to the total links in each shell. In which, Fig. 34.4a, b and c describe the data sets of Jan, 2008, Jan, 2009 and Jan, 2010 respectively and Fig. 34.4d is the zoom out of the links to the highest shell of Fig. 34.4a, b and c.

Even as Fig. 34.4a, b and c, these data sets behave consistent: the links to the highest shell occupy most links, and the other links to the higher shells distribute equably. The lower is the shell, the lower is the ratio, which we can see clearly in Fig. 34.4d. The ratio of the links to the highest shell is increased with the shell. This is because the lower is the shell, the more is the hierarchies between the highest shell and it. And just as the existence of the hierarchies, the links to higher shells are dispersed. So the links to the nodes in the highest shell in lower shells are lower. As the links in each shell distributes averagely, the effect on the links to the highest shell is changed equably with the change of the shell.

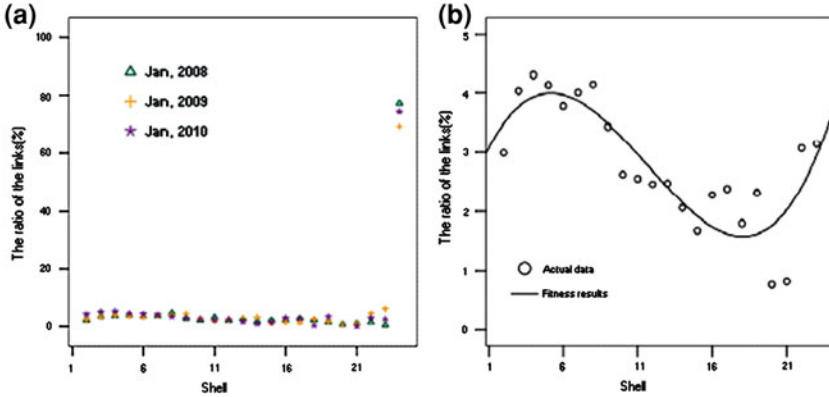


Fig. 35.5 The fitness graphs of distribution on links between the nodes in different shells and their higher shells

Figure 34.5 shows the average value of the above data sets. In which, Fig. 34.5a is the result of the average value computed according to x -axis of Fig. 34.4a, b and c. Fig. 34.5b shows the ratio of the links to higher shells except the highest shell. In which, the vertical axis is labeled by the average of the data in Fig. 34.5a. We can see from it that the ratio in lower shells is higher, then reduces with the increase of the shells, and ascends again when it is close to the highest shell. The Fitness is as followed:

$$Y = 0.030693 + 0.004841X - 0.000722X^2 + 0.000023X^3 \quad (34.2)$$

We can get the ratio of the links to the highest shell in each shell by the following way:

We denote k as one shell of the topology, K as the highest shell, and import k to $(K-1)$ into formula 34.2. Then we can get the ratio of the links to higher shells (except the highest shell) in shell k , denoted as $Y_k, Y_{k+1}, \dots, Y_{K-1}$ respectively. And the ratio of the links to the highest shell in shell k is as followed:

$$R_k^K = 1 - \sum_{i=k}^{K-1} Y_i. \quad (34.3)$$

34.6 The Modeling Algorithm Based on Coreness

Combined with the results we analyzed above and some classical modeling on Internet topology, now we give a new framework on modeling Internet AS-level topology which use k -core decomposition as basement and combine with the property of preferential attachment. The hierarchical property of the network is exhibited in it.

The main thought of the modeling is: The Internet AS-level topology is divided into three parts: the highest shell, the lowest shell and the rest. The modeling begins from the highest shell, then import the second highest shell, and add links between them, and so on, until the lowest shell. Considering most of the existing methods of modeling on topology all adopt preferential attachment as basis, and the only difference is the linking probability function. In order to avoid the similarity in degree of the nodes in different shells, especially for the innermost shell, preferential attachment is also chosen in the model we proposed. But, preferential attachment is only limited to the same shells, and the linking probability function of the nodes which degree is d_i in shell c is:

$$P = \frac{(d_i - c + 1)^{1+0.45\lg(c)}}{\sum_j (d_j - c + 1)^{1+0.45\lg(c)}}. \quad (34.4)$$

In formula 34.4, it chooses the result of the difference of degree d_i and coreness c of the nodes and adds 1 as the base. This is because in the beginning of preferential attachment in the shell, the difference of the degree of the nodes is not evident, and is close to the coreness of them. And this will result in the invalidation of the existing linking probability function. Considering the degree of the nodes is pre-assigned, so we must consider the preference trend of the nodes subtracted the degree. And add 1 is in order to ensure the chosen probability is not zero. The algorithm is as follows:

Input: N -the number of the nodes; K -the highest shell of the network topology;
Output: Network topology.

- (1) Compute the number of the nodes in shell K according to formula 34.1, that is

$$N_K = 0.86 \times N \times K^{-2.319}$$

- (2) Construct the highest core, that is shell K ;
- (3) Construct shell $K-1$ to shell 1 respectively according to the steps as follows:

- (a) Compute the number of the nodes in shell k ($2 \leq k \leq K-1$) according to formula 34.1, that is

$$N_k = 0.86 \times N \times k^{-2.319}$$

and the number of the nodes in shell 1: $N_1 = N - \sum_{i=2}^K N_i$

- (b) For each shell k' higher than shell k , compute the linking probability to them in shell k according to formula 34.2, that is

$$R_k^{k'} = 0.030693 + 0.004841k' - 0.000722k'^2 + 0.000023k'^3 \quad (k < k' < K)$$

- (c) Compute the linking probability to the highest shell in shell k according to formula 34.3, that is

Table 34.1 The comparison of some main features in three data sets

Data set	AS measuring data	BA model	Model in the paper
The number of the nodes	8,120	8,000	8,000
Average degree	5.51	6.01	5.34
The largest degree	1,119	236	1,249
CCDF power exponent	1.14	1.75	1.10
Network coreness	23	5	23

$$R_k^K = 1 - \sum_{k'=k}^{K-1} R_k^{k'}$$

- (d) For the nodes in shell k , choose the nodes that have not got links with others to construct the links in shell k according to the linking probability (3.84%);
- (e) For the nodes in shell k , choose the nodes in higher shells to construct links according to the linking probability computed in steps (b) and (c) with the formula 34.4.

(4) End.

In step 2, because the highest core plays the very important role for the network topology, so it must be constructed separately. We choose 10% of the all nodes, N_K , in the shell to link with the other nodes in order to get the whole connection. For the rest nodes, we assign $(K - 10\% N_K)$ links for them to ensure their roles as to be the highest core.

34.7 The Results and Analysis

To test the validation of our model, we choose the classical model, BA model, to compare in some main features. Table 34.1 illustrates the experiment results.

We can see from table 34.1 that the model we proposed is more close to the actual topology under the same parameters. Especially for the network coreness, its superiority is more evident compare to BA model. And this is the powerful proof that the model we proposed fully exhibits the hierarchy in AS-level.

Figure 34.6 shows the CCDF-degree distribution of the model we proposed and BA model.

Figure 34.6a describes the data in the model we proposed, and Fig. 34.6b describes the data in BA model. We can see from it that although the data in the two models is different in some parts, but they all exhibit power-law in the whole. This proves the model we proposed give prominence to the aspect of power-law, while it also illustrated in the other hand that the power-law of the network does not only lie on the preferential attachment of the nodes.

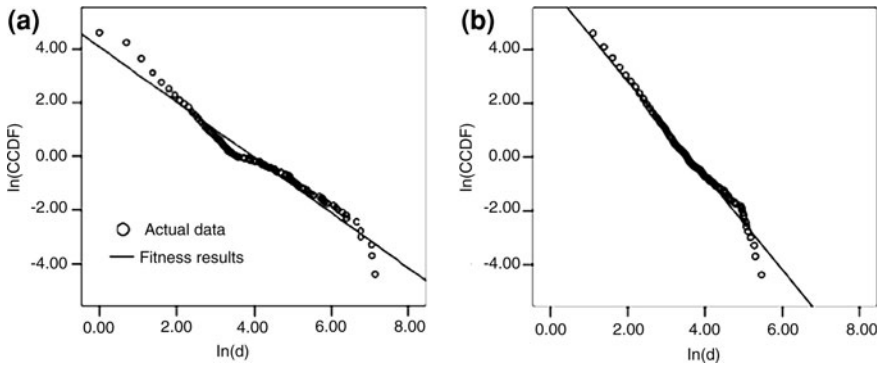


Fig. 34.6 The contrast graph of power exponents of CCDF(d)-degree distribution between the model we proposed and BA model

34.8 Conclusion

In this paper, we first analyzed some crucial properties related with coreness with the actual topology measuring data. Then we designed a modeling algorithm based on coreness according to the statistical rules. At last, we realized the model. By analyzing the experimental data and comparing to the classical BA model, we illustrated the advantages of the model. The model does not ensure the power-law of the network, but also exhibits the hierarchy of the network better. In the meanwhile, it also indicates that the node coreness plays the important role in the measurement of network topology.

References

1. Zhang Y, Zhang HL, Fang BX (2004) A Survey on internet topology modeling. *J Softw* 15(8):1220–1226
2. Albert R, Barabási A (2000) Topology of evolving networks: local events and universality. *Phys Rev Lett* 85(24):5234
3. Barabási A, Albert R (1999) Emergence of scaling in random networks. *Science* 286:509–512
4. Vilhar A, Novak R (2008) Policy relationship annotations of predefined AS-level topologies. *Comput Netw* 52(15):2859–2871
5. Wang J, Liu YH, Zhang C et al (2009) Model for router-level Internet topology based on attribute evolution. *IEEE Commun Lett* 13(6):447–449
6. Fabrikant A, Koutsoupias E, Papadimitriou CH (2002) Heuristically optimized trade-offs: a new paradigm for power-laws in the internet. In: *ICALP '02: proceedings of the 29th international colloquium on automata, languages and programming*, Springer-Verlag, Berlin Heidelberg London, pp 110–122
7. Zhou S, Mondragon RJ (2004) Accurately modeling the internet topology. *Phys Rev E* 70:066108

8. Zhou S, Mondragon RJ (2003) Towards modeling the internet topology-the interactive growth model. *Teletraffic Sci Eng* 5:121–130
9. Sagy B, Mira G, Avishai W (2005) A geographic directed preferential internet topology mode. Arxiv: CS, NI/0502061
10. Alvarez-Hamelin JI, Dall'Asta L, Barrat A, et al (2005) k -core decomposition: a tool for the visualization of large scale networks. Arxiv preprint cs. NI/0504107

Chapter 35

Based on Quasi-IP Address Model of Repeater Coordination

Shaohong Yan, Lihui Zhou, Hong Wang, Yan Li, Lanqian Liu and
Mengyuan Chen

Abstract The paper discusses the structure of a wireless network satisfying the communication demand of different users. The circular system of the network can not only satisfy the demand of between two users, but also between one user and other users. Our work is divided into two steps. At first, we find “quasi-IP address” for every user and repeater to build a “quasi-IP address model”. And then, we distribute labels and positions for each repeater to build a “circular system model” for each user to communicate for one or more users.

Keywords Quasi-IP address model · Circular system · Effective service coverage · Correction factor

35.1 Introduction

The VHF radio spectrum involves line-of-sight transmission and reception [1]. This limitation can be overcome by “repeaters,” which pick up weak signals, amplify them, and retransmit them on a different frequency to avoid self-transmission and self-receiving. Therefore, a pair of special transmitter frequency (TF) and receiver frequency (RF) is set for each repeater [2, 3]. Within the available spectrum of 145–148 MHz, when the transmitter frequency in a repeater is either 600 kHz above or 600 kHz below the receiver frequency, 10 frequency pairs are available as in Table 35.1.

S. Yan (✉) · L. Zhou · H. Wang · Y. Li · L. Liu · M. Chen
College of Science, Hebei United University, Tangshan Hebei
063009 china
e-mail: shaohong@heut.edu.cn

Table 35.1 Frequency pair

RF (MHz)	145	145.6	146.2	146.8	147.4	148	147.4	146.8	146.2	145.6
TF (MHz)	145.6	146.2	146.8	147.4	148	147.4	146.8	146.2	145.6	145

However, repeaters can interfere with one another unless they are far enough apart or transmit on sufficiently separated frequencies. In addition to the two solutions, the “continuous tone-coded squelch system” (CTCSS), sometimes nicknamed “private line” (PL) technology can be used to mitigate interference problems. This system associates to each repeater a separate subaudible tone that is transmitted by all users who wish to communicate through that repeater. The repeater responds only to received signals with its specific PL tone. With this system, two nearby repeaters can share the same frequency pair (for receive and transmit); so more repeaters (and hence more users) can be accommodated in a comparatively small area. For each repeater, other repeaters communicating with it are regarded as users. Because each repeater provides services to all its users, the receiver frequency of the repeater is the transmitter frequency of the user and vice versa. For the sake of convenience, we use a system called “quasi-IP address” to label repeaters and users as in the following.

- (1) For a user, if the receiver frequency available is 145 MHz and transmitter frequency available is 145.6 MHz, the part for frequency pair of the label is 145.1.2; if the receiver frequency available is 145.6 MHz and transmitter frequency available is 146.2 MHz, the part for frequency pair of the label is 145.2.3; if the receiver frequency available is 146.2 MHz and transmitter frequency available is 146.8 MHz, the part for frequency pair of the label is 145.3.4. The rest can be deducted by analogy.
- (2) For a user, if the receiver frequency available is 148 MHz and transmitter frequency available is 147.4 MHz, the part for frequency pair of the label is 145.6.5; if the receiver frequency available is 147.4 MHz and transmitter frequency available is 146.8 MHz, the part for frequency pair of the label is 145.5.4; if the receiver frequency available is 146.8 MHz and transmitter frequency available is 146.2 MHz, the part for frequency pair of the label is 145.4.3. The rest can be deducted by analogy.
- (3) For a user, if the subaudible tone available is 5, the part for subaudible tone of the label is 5.
- (4) Following is the label for a user.

Part for frequency pair			Part for subaudible tone
Basic frequency	Receiver frequency	Transmitter frequency	
145	1 ~ 5	2 ~ 6	1 ~ 54
	6	5	1 ~ 54
	5 ~ 2	4 ~ 1	1 ~ 54

For example, for a user, if the receiver frequency available is 145.6 MHz and transmitter frequency available is 146.2 MHz, and when it transmits signals to its repeater, if the subaudible tone used is 7, the label for the user is 145.1.2.7.

(5) Because for each repeater, other repeaters communicating with it are regarded as users, so the way to label a repeater is the same as labeling a user.

For a circular flat area of 40 miles radius to accommodate 1,000 simultaneous users, in any arbitrary time, every user can communicate and establish a private double-way link (the receiver is the designated repeater or user for the transmitter) with any other arbitrary repeater or user through the network of both. Therefore, the allocation and choice of repeater is the focus of our research.

Based on the above mentioned Quasi-IP Address, we can label all repeaters and users. In the same way, as dealing with the problem of buildup network, we can use loop network topology to establish private link among all repeaters and users in a circular flat area of 40 miles radius.

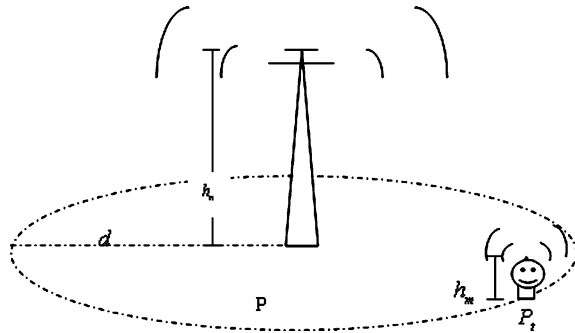
35.2 Model Assumptions

- (1) The circular area in our research is idealistic, flat and without any barrier to communication.
- (2) 1,000 users are uniformly distributed within the area and without any change of position.
- (3) The repeaters in the network is controlled by automatic controlling center, which, following the principle of automatic control, transmits signals combined with subaudible tone to all repeaters and users.
- (4) Both repeaters and user possess the function of coding and decoding signals combined with subaudible tone. When a repeater provides services to a user, both of them will use the same subaudible tone.
- (5) The services for a user are provided by the repeater closest to it.

35.3 Symbols

P_r —Power of User (W), P_{\min} —Protecting Power (W), G_r —Antenna Gain of User (dB), G_r —Antenna Gain of Repeater (dB), L_r —Feeder Antenna Loss of User (dB), h_m —Antenna Height of User (m), h_n —Antenna Height of Repeater (m), d —the Max Communication Distance (km), K —Correction Factor (dB), f —Transmitter Frequency of Repeater (145–148 MHz), x —Radius of Central Circle, r —Service Circle of Outer Repeater, N —Layer of Outer Repeater

Fig. 35.1 Coverage of a repeater



35.4 Coverage of One Repeater

There will be loss when a signal with certain frequency [4], which is related to the signal frequency, the terrain spread, transmission distance, repeater height and so on. Fig. 35.1

By the empirical formula, based on measured data, path loss can be obtained:

$$LM = Pt + Gt + Gr - P \min - Sm - Lt \tag{35.1}$$

Longest communication distance :

$$d = \log^{-1} \{ [LM - 88.1 - 20 \log f + 20 \log(h_m \times h_n) - K] / 40 \} \tag{35.2}$$

We can get the longest communication through (35.1) and (35.2). But the change of antenna height and power the coverage radius of repeater can be adjusted. In order to realize the maximum effect, we can choose the necessary communication distance within the biggest communication coverage.

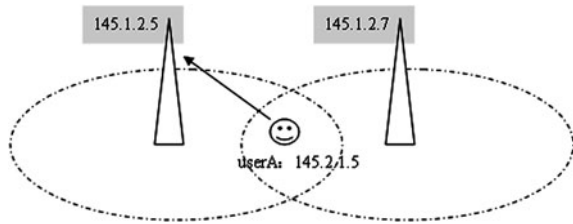
35.5 Number of Users Supported by a Repeater

We have known that when two repeaters use the same frequency with an enough distance from each other or two repeaters use the different frequencies close to each other, there won't be any interference between them. But when two repeaters use the same frequency close to each other, there will be interference, which, can be solved by the system of CTCSS, which is also called "Private Link" (PL).

35.5.1 Differentiation of Repeaters Through the Use of Subaudible Tone

When receiver frequency of two repeaters is 145 MHz, and transmitter frequency of user A is also 145 MHz, the two repeaters can receive the signals transmitted by

Fig. 35.2 Differentiation of repeaters through the use of subaudible tone



user A combined with the same subaudible tone, or the signals can't be received. As in the following chart, signals of user A is received by repeater labeled 145.1.2.5. (Fig. 35.2)

35.5.2 Differentiation of Users Within the Coverage of a Repeater Through the Use of Subaudible Tone

When two users communicate with each other within the coverage of a repeater, because the receiver frequency and transmitter frequency of them are the same and equal to the receiver frequency and transmitter frequency of the repeater, if the repeater and the two users don't use subaudible tone, then signals transmitted by user A will be received by all of the users, as user B or user C, etc., within the coverage of the repeater, which becomes a broadcasting system in reality.

However, if user A and user B use subaudible tone matching with each other, then signals transmitted by user A will be received by user B only, hence is "private line". Of course, if user C and user B simultaneously use the subaudible tone matching with user A, then they can receive signals transmitted by user A simultaneously. Fig. 35.3

A broadcasting system can be used by all of the users within the coverage of a repeater, and subaudible system can be used by 54 users within the coverage [5].

35.5.3 Differentiation of Repeaters and Users Through the Use of Subaudible Voice

Figure 35.4 user A sent to repeater data with subaudible 1, Repeater I's tone is 1, which is activated to start forwarding data of user A's. By the control system, Repeater I transmit tone 9, then user B receive user A's data.

User C want to send data to user D, User C sent to repeater data with subaudible 1 first, Repeater I's tone is 1, which is activated to start forwarding data of user C's. By the control system, Repeater I transmit tone 2, then user D receive user C's data.

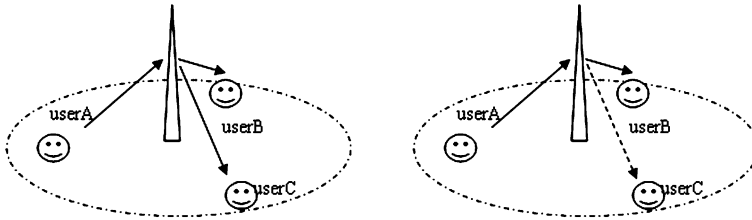


Fig. 35.3 Broadcasting system without the use of CTCSS technology (*left*), One-to-one system with the use of CTCSS technology (*Right*)

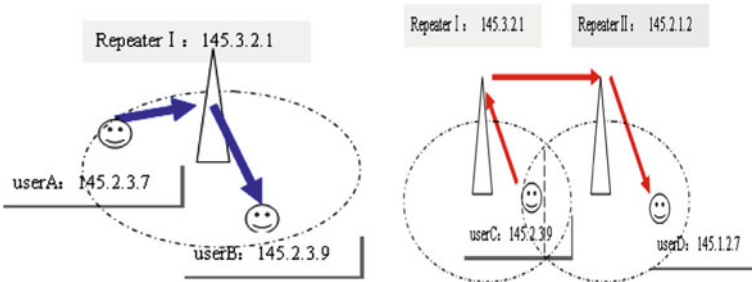


Fig. 35.4 Though one repeater (*Left*), though two repeaters (*Right*)

35.6 Building a Circular System

35.6.1 Service Area

The practical area occupied by users served by the repeater is called service area, which is different from the coverage area of the repeater [6]. Users within the coverage area of a repeater are not necessarily within the service area of the respective repeater.

35.6.2 Calculation Method

The radius of the central circle is x ; the width of the first layer is r ; the width of the second layer is r ; and rest can be deducted from analogy (Fig. 35.5).

From the total radius of R km, the following formula can be deducted out:

$$R = x + nr \tag{35.3}$$

If we make the area of the inner circle equals to the area of the outer “ladder-shaped area”, the following formula can be deducted out:

Fig. 35.5 Circle system

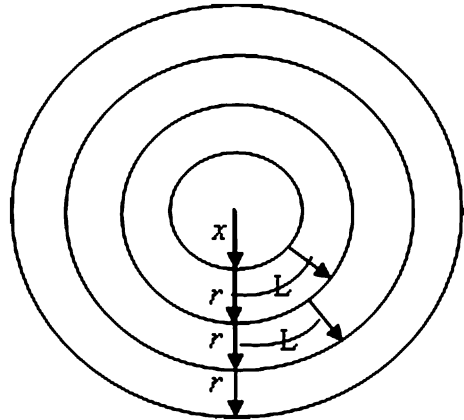
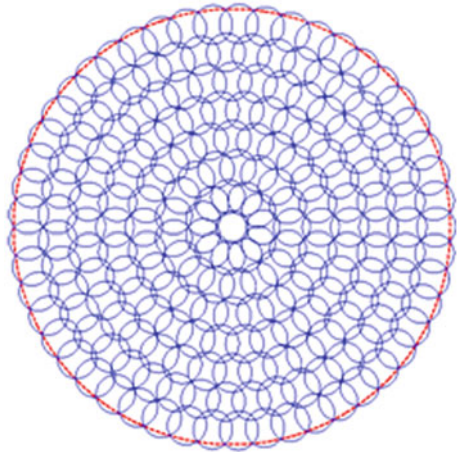


Fig. 35.6 197 repeaters



$$\pi x^2 = rL \tag{35.4}$$

Finally, we should adjust the number of the divided areas in the circle bigger or equal to $\lceil Z/53 \rceil$. (the minimum number from the calculation of number of users/53)

$$1 + \frac{2\pi(x + \frac{r}{2})}{L} + \frac{2\pi(x + \frac{3r}{2})}{L} + \dots + \frac{2\pi(x + \frac{(2n-1)r}{2})}{L} \geq \lceil Z/53 \rceil \tag{35.5}$$

The simultaneous formula of (35.3) ~ (35.5) will be calculated to deduct the optimized result.

For example, $Z = 10000$, $R = 64$, we have $n = 7.1921$, which, after integration, is $n = 7$, i.e. 7 layers.

$r = 8.4609$, $x = 4.7736$, L and r is equal to each other approximately.

In our statistics, the number of repeaters is 197 and the concrete distribution of them is in the following Fig. 35.6

35.7 Summary

- (1) Using “quasi-IP address”, we can not only establish one-to-more communication but also establish one-to-one communication and signals transmitted won't be received by other users.
- (2) “Cellular system” equals the effective service area of repeaters into hexagons, but we use “circular system” to equal the effective service area of repeaters into fan-shaped area, which covers the whole circle of 40 miles.
- (3) When a “Cellular system” uses hexagons to cover a circle, there will be cracks between hexagons, which, compensated only through the addition of more hexagons, will bring waste to the system.
- (4) Combined with “Time Division Multiple Accessing” technology, “quasi-IP address” will satisfy more simultaneous users.

References

1. Meerschaert MM (2005) *Mathematical modeling*, 2nd edn. China Machine Press, China, 6
2. Lee P (2007) *The frequency of cellular relay network planning and protocol design*. Shanghai Jiao Tong University, 04
3. Zhang L, ZongLing C (2007) *Mobile radio propagation prediction model selection founder database*, 11
4. GuoDai C (1995) *Mobile communication network to establish the radio*, Posts Telecom Press, 12
5. XiaoHui L, Ling C (1995) *Series of radio enthusiasts*. Journal of Chengdu University of Science, 04
6. ShaoMing X (1998) *Wireless communication technology*, Shenzhen Huawei Technology 10

Chapter 36

Fast Dynamic Mesh Moving Based on Background Grid Morphing

Tianjun Lin and Zhenqun Guan

Abstract The dynamic mesh moving method based on Delaunay graph is non-iterative and therefore efficient. However, intersections occur occasionally among the background graph elements for complex geometries with large relative movements, especially for the large boundary rotations. It not only consumes more time but also deteriorates mesh quality to regenerate the graph and to relocate mesh points. In this work, a dynamic mesh generation method based on background grid morphing is proposed. By appending some assistant points in the initial graph, a new background graph in conjunction with a new mapping is generated. Then, the ball-vertex method with boundary improvement is employed to move the new graph which drives the moving of the object mesh. The examples demonstrate that the improved method have better ability of keeping the shape for the regions concerned and can improve the intersection problems even with large boundary rotations. In general, the method proposed shows better performance in mesh quality and efficiency.

Keywords Dynamic meshes · Mesh moving · Delaunay graph · Ball-vertex method

T. Lin (✉) · Z. Guan
State Key Laboratory of Structural Analysis for Industrial Equipment
Department of Engineering Mechanics, Dalian University of Technology
Dalian 116024 China
e-mail: tianjun_lin@yahoo.com.cn

Z. Guan
e-mail: guanzhq@dlut.edu.cn

36.1 Introduction

Flow problems with moving boundaries arise in many engineering areas, such as fluid—structure interaction (FSI), free surface flows, etc. When boundaries of the solution domain are changed, the mesh should be updated. The dynamic mesh can be considered as a means to propagate the boundary moving into the whole mesh of solution domain. There are generally two categories of methods to achieve this: (1) by mesh regeneration; (2) by mesh moving. However, the efficiency and robustness of mesh regeneration method are limited. The mesh moving approach is usually adopted. There are mainly two methods for mesh moving: spring analogy method and elastic solid method. The elastic solid method can deal with problems with large movements and maintain the mesh quality. But the efficiency is poor for its large computational cost. On the other hand, the spring analogy method is simple and efficient. The spring analogy scheme, first developed by Batina [1], is widely used [2, 3]. It looks the mesh as a network of linear springs and solves the static equilibrium equations for this network to determine the new locations of the mesh points. However, as there is no constraint on the angles between two adjacent edges in one element, the method suffers problems in robustness for large deformation, such as meshes for viscous flows, resulting in mesh crossing and negative volumes. Farhat [4] proposed a modified spring analogy by adding additional non-linear torsion springs to avoid the negative cell volume problem associated with the linear spring network. Owing to the complexity of the torsion spring, the computational cost is increased and the results are still not satisfactory when it is applied to large movements.

Recently, Liu [5] presented a new mesh moving method based on Delaunay graph mapping. The method is much more efficient, as it is non-iterative. However, when it comes to the applications which have more complex geometries with large relative movements, especially for the large boundary rotations, intersections of the background grid easily occurs. Literature 5 presents the treatment strategy that regenerate Delaunay map and relocate the mesh points, but this will lead to new problems. Firstly, it will spend more computational time; secondly, it will discard the good features of the initial mesh, while retaining the bad information of the deformed one. After that, the mesh quality is even worse. When the boundary backs to the initial position, the mesh quality cannot be restored to the initial state. Moreover, the more times the background graph regenerated, the more serious situation it will be.

This paper presents an improved mesh moving method based on background grid. By appending some assistant points in the initial graph, a new background graph in conjunction with a new mapping is generated. Then, the ball-vertex method with boundary improvement is employed to move the new graph which drives the moving of the object mesh. After relocating the mesh points in the new graph, we obtain the final mesh. The improved method have better ability of keeping the shape for the regions concerned and can improve the intersection problems even with large boundary rotations. It also shows better performance in mesh quality and efficiency.

36.2 Dynamic Mesh Moving Based on Delaunay Graph Mapping

A detailed analysis of the method is presented by Ref. [5]. In summary, the procedure of the moving mesh method can be written as follows:

- (1) Using selected boundary points and the Delaunay method [6, 7] to construct the Delaunay graph;
- (2) Locating the graph elements for all the computational mesh points using the walk-through method [8]; computing and storing the area or volume ratio coefficients for each mesh point using the linear interpolation method;
- (3) Moving the Delaunay graph according to the deformation of the solid surfaces or the relative movements between different bodies;
- (4) Checking for any intersection between the graph elements. And if this happens, go back to step (1). Otherwise go to the next step; If the movement is too large, intersections of the mesh will occur. We need to remesh the domain;
- (5) Relocating the computational mesh points in the graph;
- (6) Repeating step (3)–(5) for the next time step.

This method is simple and much more efficient, as it requires non-iterative algebraic calculations. But, when intersections of the background graph occur, we need to spend more time to regenerate Delaunay map and relocate the mesh points. And also the mesh quality gradually gets worse and worse.

36.3 The Ball-Vertex Method

In this section we briefly review the classical edge spring method whose basic idea is to create a network of springs connecting all vertices in the mesh. Each mesh edge is replaced by a spring. Given two vertices, i and j , the displacements of vertices i and j are noted u_i and u_j , respectively. The resulting force on vertex i is aligned along the unit vector i_{ij} and can be written as

$$f_{ij} = k_{ij}(u_j u_i) \cdot i_{ij} i_{ij} = f_{ji}. \quad (36.1)$$

where i_{ij} is unit edge-vector, K_{ij} is the spring stiffness which is typically chosen as inversely proportional to the edge Length.

The position of each vertex is found by writing its equilibrium under the effect of all its edge connected springs. The resulting linear system of equations can be solved in a variety of ways. In this work we use a Gauss–Seidel iteration method.

While this classical method performs reasonably well in a number of cases, it does indeed fail as soon as the local mesh motion is not small compared to the local mesh size. The method cannot control the collapse mechanisms of the mesh elements. In the presence of nearly flat elements, cross-over is easily obtained. Clearly, it will lead to a locally invalid mesh. For the sake of simplicity, we restrict our attention to the sole more interesting three-dimensional problem in the following.

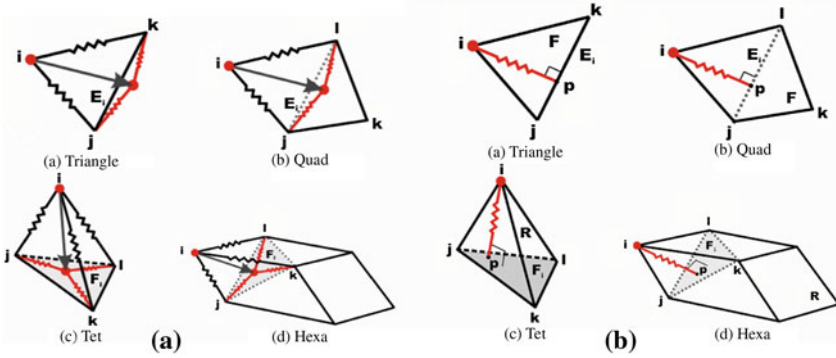


Fig. 36.1 The ball-vertex method. **a** Element collapse, **b** Add a new spring

With the goal of controlling the volume of a mesh element, the ball-vertex method [9] introduces additional linear springs that resist the motion of a mesh vertex towards each one of its region-opposed faces. Once the additional springs are constructed for all tetrahedrons using a vertex i , one has effectively constrained the same vertex not to leave the polyhedral ball $B_i = \text{ball}(i)$ that encloses it (Fig. 36.1).

The position of each mesh vertex is found by writing its equilibrium under the combined effect of its edge-connected springs, together with the additional ball-vertex springs. If u_i and u_p indicate the displacements of vertex i and of its projection p on F_i , respectively, the resulting force on i can be expressed, similarly to the edge spring case, as

$$f_{ip} = k_{ip}(u_p - u_i) \cdot i_{ip}i_{ip} = -f_{pi} \tag{36.2}$$

p is a virtual point, We interpolate the displacement at p based on the displacement values at the three face vertices j, k and l , respectively.

$$u_p = \xi u_j + \eta u_k + (1 - \xi - \eta)u_l \tag{36.3}$$

Where ξ and η are the interpolation coefficients.

The ball-vertex method offers straightforward solutions. It is robust and effective when dealing with large deformations. It can also maintain a better mesh quality.

36.4 Boundary Improvement

The spring stiffness is allowed to change by the introduction of two parameters in the definition of the stiffness [10].

$$K_{ij} = \Phi \left((x_i - x_j)^2 + (y_i - y_j)^2 \right) \Psi \quad (36.4)$$

In the original concept of Batina, the parameter values are $\Phi = 1$ and $\psi = -0.5$. The purpose of the spring analogy is to regularize a mesh after boundary deformation. However, this regularization only takes place close to the deformed boundary. Therefore, the spring analogy can only handle relatively small deformations since the mesh regularizes only locally. In order to improve it, the stiffness near the boundary is increased so that the deformation is spread out further into the mesh. To achieve it, the factor Φ in Eq. 36.4 is increased for a number of element layers adjacent to the boundary.

We will use the improvement in the ball-vertex method in the next section.

36.5 The Improved Method Based on Background Grid

We append some assistant points in the initial graph and a new background graph in conjunction with a new mapping is generated which makes a transition between inside and outside of the background grid. Then, the ball-vertex method with boundary improvement is used to move the new graph which drives the moving of the object mesh. In the background grid, as the layers of the mesh are reduced significantly, the springs are stiffened only at one layer adjacent to the boundary. This has been shown to be sufficient. It should be noted that, due to the small number of assistant points, the efficiency is almost no changed.

The procedure of our improved method can be written as follows:

- (1) Using selected boundary points and the Delaunay method to construct the Delaunay graph;
- (2) Adding assistant points to the center of each element of the background grid. Using the Delaunay method to generate a new layer background grid which improves the connection between the local areas in order to eliminate intersections of the background graph;
- (3) Locating the new graph elements for all the mesh points;
- (4) Moving the new Delaunay graph according to the deformation of the solid surfaces or the relative movements between different bodies and using the ball-vertex method to move the assistant points to get the final background graph;
- (5) Checking for any intersection between the graph elements, and if this happens, go back to step (1). Otherwise, go to the next step; In practice, by adding assistant points to improve the connection between local area and using ball-vertex method to move these points, it can greatly improve the intersection of the mesh;
- (6) Relocating the mesh points in the new graph;
- (7) Repeating step (4–6) for the next time step.

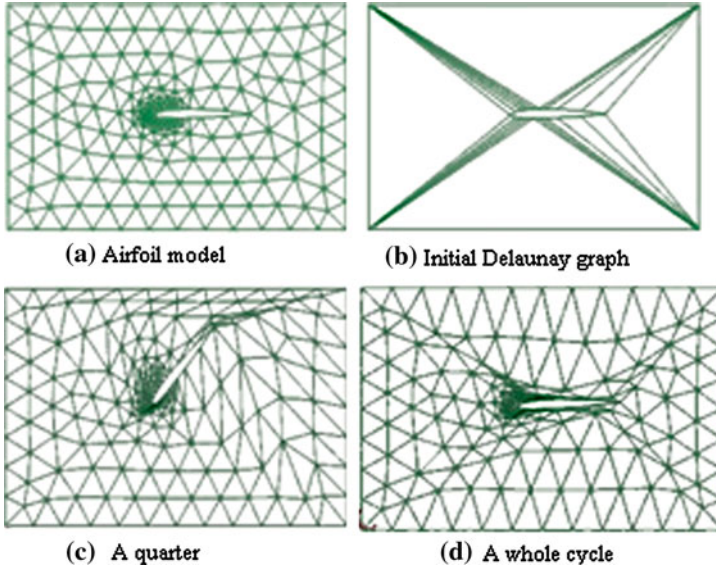


Fig. 36.2 The original method. **a** Airfoil model, **b** Initial delaunay graph, **c** A quarter, **d** A whole cycle

36.6 Numerical Examples

In this section, two test cases demonstrate our method for moving meshes. The improved method is compared with the original one described in [Sect. 36.2](#).

36.6.1 Rotation of the Airfoil

In this example we consider an airfoil with pitching amplitude $h_0 = 50^\circ$, that moves according to

$$\theta_n = \theta_0 \cos(2\pi n/N) \quad (36.5)$$

where n is the generic time step, $N = 100$ is the number of time steps per cycle.

First, the mesh is deformed by the original method.

The result is shown in [Fig. 36.2](#). It is seen that, in one cycle, intersections occur twice when the airfoil rotates to 30° and -25° . Thus we must regenerate the background graph twice. Then it will discard the good features of the initial mesh, while retaining the bad information of the deformed one. After that, the mesh quality becomes even worse.

Here perform the mesh moving by our improved method.

Although the imposed displacements at each time step are rather large, the improved method exhibits a remarkably performance throughout the simulation. In

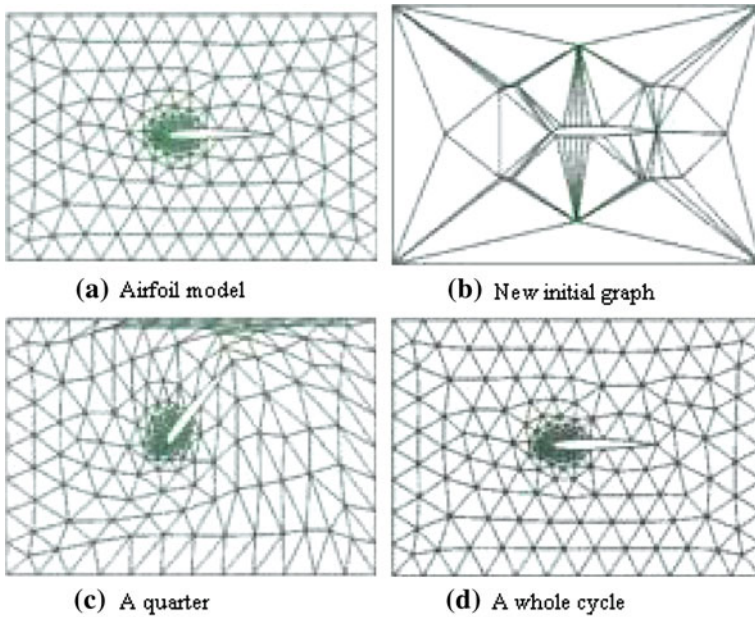


Fig. 36.3 The improved method. **a** Airfoil model, **b** New initial graph, **c** A quarter, **d** A whole cycle

Fig. 36.3, no inversion occurs in a whole cycle. This method is able to maintain a good mesh quality. When the boundary backs to the initial position, the mesh quality can be restored to the initial state.

We consider quality measures based on the ratio of the radii of the inscribed and circumscribed circles, noted $2r/R$. Here, we choose some elements around the inner boundary for test. Figure 36.4 shows that our method has greatly improved the quality of the mesh.

Through the analysis of the ball-vertex method, we can easily draw the conclusion that: The shorter the distance between the point and its opposite edge, the stiffer the spring added is. In the new background grid, the elements around the inner boundary are long and narrow. The length of the spring added is small, so it is stiffer. Therefore, the deformation of the elements near the inner boundary is relatively small. Accordingly, the computational mesh elements near the inner boundary deform less. This is one advantage of our algorithm.

36.6.2 Geometries Composed of Segments

We deform a logo such as the one illustrated in Fig. 36.5. To compare the two methods, we will take two kinds of boundary movement strategy.

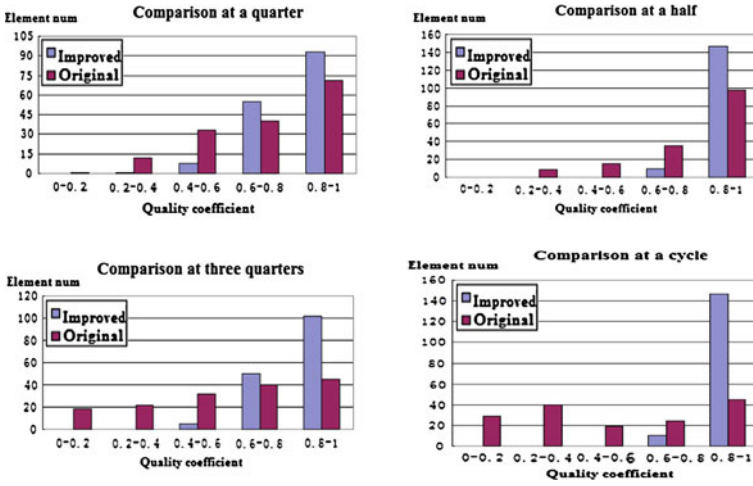


Fig. 36.4 Quality comparison

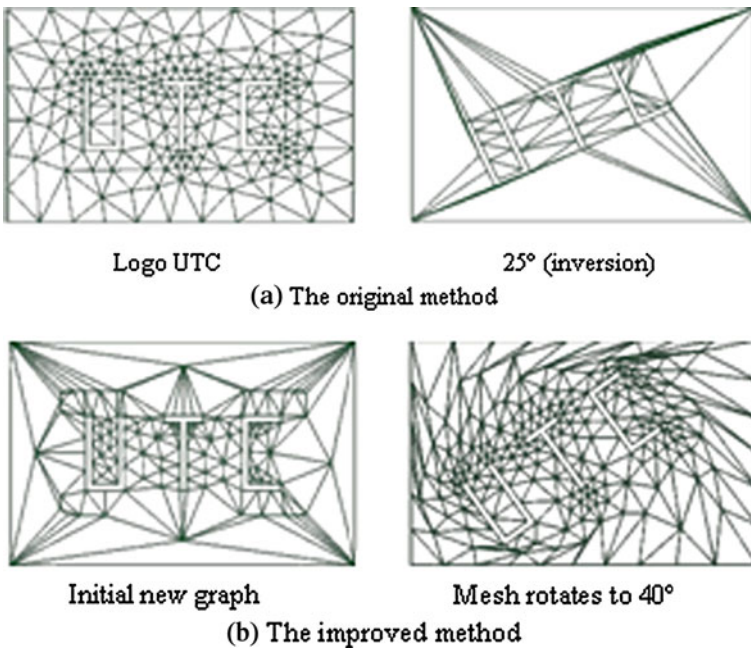


Fig. 36.5 Comparison of strategy one. a The original method b The improved method

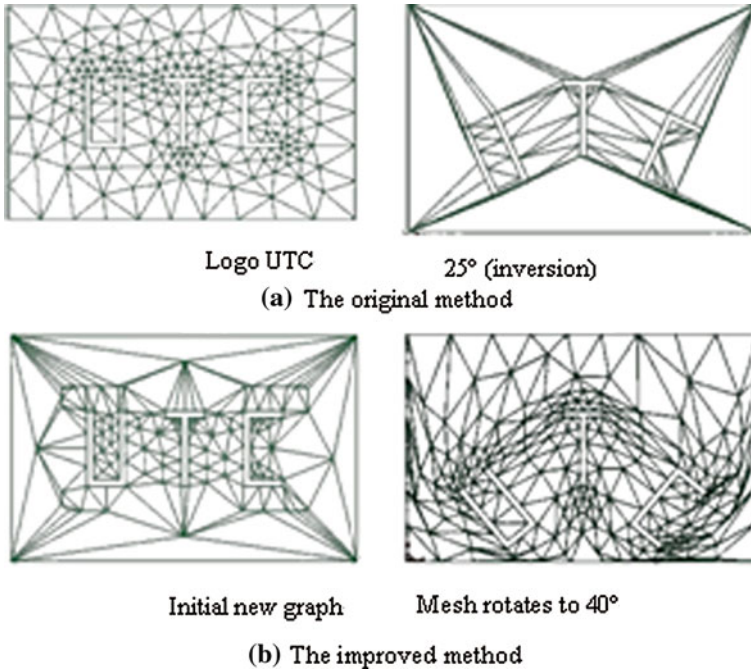


Fig. 36.6 Comparison of strategy two. **a** The original method, **b** The improved method

Firstly, the logo is counterclockwise rotated around a point on T.

Secondly, U is counterclockwise rotated around a point on T, at the same time, C is clockwise rotated around the same point.

As can be seen from Figs. 36.5 and 36.6, in the original method, intersection occurs when the logo rotates to 25°. The improved method can still maintain the mesh valid and ensure the mesh quality when it rotates to 40°.

36.7 Conclusion

This paper presents an improved mesh moving method based on background grid which is simple to implement for arbitrary mesh topologies. By appending some assistant points in the initial graph and using the ball-vertex method, it generally improves the efficiency and quality of the dynamic mesh. Some cases demonstrate that this method have better ability of keeping the shape for the regions concerned and can greatly improve the intersection problems even with large boundary rotations.

References

1. Batina T (1990) Unsteady Euler airfoil solution using unstructured dynamic mesh. *AIAA J* 28(8):1381–1388
2. Robinson AB, Batina T, Yang HTY (1991) Aeroelastic analysis of wings using the Euler equations with a deforming mesh. *J Aircr* 28(11):781–788
3. Kleb WL, Batina T (1992) Temporal adaptive Euler/Navier—stokes algorithm involving unstructured dynamic meshes. *IA-A J* 30(8):1980–1985
4. Farhat C, Degand C, Koobus B et al (1998) Torsional springs for two-dimensional dynamic unstructured fluid meshes. *Comput Methods Appl Mech Eng* 163(1–4):231–245
5. LIU XQ, Qin N, Xia H (2006) Fast dynamic grid deformation based on delaunay graph mapping. *Comput Phys* 211(2):405–423
6. Bowyer A (1981) Computing Dirichlet tessellations. *Comput J* 24(2):162–166
7. Watson DF (1981) Computing the n-dimension-al Delaunay tessellation with application to Voronoi polytopes. *Comput J* 24(2):167–171
8. Devroye L, Mucke E, Zhu B (1998) A note on point location of Delaunay triangulation of random points. *Algorithmica* 22(4):477–482
9. Bottasso CL, Detomi D, Serra R (2005) The ball-vertex method: a new simple spring analogy method for unstructured dynamic meshes. *Comput Methods Appl Mech Eng* 194:4244–4264
10. Blom FJ (2000) Considerations on the spring analogy. *Int J Numer Methods Fluids* 32(6):647–668

Chapter 37

Research on Network Congestion Control System Based on Continuous Time Model

Huang Ze

Abstract The congestion control system, as one of the key parts for the normal operation of computer network, has become a hot research in both domestic and abroad at currently. The preliminary research on congestion control system of computer network was focused on the improvement of algorithm and some researches on network behavior. This paper, based on some relative literatures, carried out a comparatively intensive research on the aspect of network congestion control and built up a continuous time model for TCP congestion control system. TCP congestion control algorithm and RED queue management algorithm, as the two parts of TCP congestion control system, are modeled independently and then cascaded into a unity. Finally, it analyzed the model stability near the stationary points, and pointed out that the stability of TCP congestion control system mainly depended on the RED algorithm.

Keywords Continuous time model · Network · Congestion control · System

37.1 Introduction

With the fast development of Internet, the human society and living style has been greatly changed. A variety of network technologies, such as online bank, e-commerce, e-government, internet video, networking telephone and so on, changed or replaced the traditional methods and made great contributions for the development

H. Ze (✉)
Department of Computer and Information Science,
Hechi University, Yizhou 546300 China
e-mail: aspone@qq.com

of modern society [1]. However, this is just happened in the recent decades, the future of computer network is brilliant. For importance of computer network to modern society, therefore, the research on it has never stopped since its birth. While research on network congestion control algorithm, the core algorithm of whole computer network control system, is always the hot issue. Congestion control is the key factor for maintaining the robustness of computer network, and also the basis of other computer network management and control algorithms such as QoS control, etc [2]. Thus, the research on congestion control algorithm has a decisive meaning. At currently, the computer network in China is developing fast, [3] and it has become an important factor influencing the whole society on all aspects, which not only relate with economic life, but also affect the national fundamental policies like national defense safety, etc. so the research of computer network on its behavior and control methods under the congestion situation, has a significant real meaning.

37.2 The Reasons Caused Network Congestion

At present, there are two congestion control algorithms in computer network, one is end-to-end congestion control algorithm based on flow control, whose representative is the congestion control algorithm of TCP protocol, and the other is routing congestion control algorithm based on queue management, whose representatives are RED, REM, AVQ, etc. [4] these two algorithms used together to form a congestion control system of computer network, and both complement each other. In TCP/IP network, it is the network layer and transport layer that undertake the task of network congestion control. The network layer provides the queue management based routing control algorithm, while the transport layer supplies the end-to-end congestion control algorithm based on flow control.

Network congestion is a phenomenon occurred in computer network, that is, when the amount of data in computer network exceeds the network's maximum load, the computer network starts to discard parts of data packet and resulting in quality deterioration of data transportation. The most serious situation is causing paralysis of the entire computer network, almost all data packets cannot be transported correctly to their destination end. However, this situation is strongly avoided by computer network.

Some causes of network congestion are as follows:

- 1 A large amount of network data packets suddenly intensively flow to one or several network links, which leads to local overload, enters congestion situation, which will even as serious as spreading to the entire network.
- 2 The processing speed of routing nodes in network is not fast enough, leading to large amount of data packets existed in network and then entering congestion situation.
- 3 The too low bandwidth of network can also make the data packets accumulated in network links, which then leads to congestion for the overload of network.

Once the network congestion occurred, may it possibly be further deteriorated and spread, because the routing nodes discard data packet for the queue saturation, and then the sending end resend this data packet for overtime, even resend in many times, which causing a further increase of data amount in network and heavier network congestion. Therefore, one of the important purposes of congestion control is to avoid network entering the congestion situation rather than to “dredge” after network entering the congestion state.

37.3 The Modeling of TCP Congestion Control Algorithm

Suppose the congestion window is $W(t)$, $R(t)$ represents RTT [5], the queue length of routing nodes is $q(t)$, the number of TCP flows accessed is $N(t)$, C is the bandwidth of TCP flows shared link.

Considering the definition of RTT, we can simplify $R(t)$ as the following form

$$R(t) = di + \frac{q(t)}{C} \quad (37.1)$$

Among, di represents the propagation delay time of number I TCP flow.

Meanwhile, it can be considered that the distribution of TCP flows in unity is equilibriums, which means if each TCP flow is distributed independently then can suppose $R(t) = R(t)$. Suppose all TCP flows are TCP Reno, then it can be found according TCP Reno algorithm that after the flow rate of each TCP flow passed their first peak point, they will be stay at CA status rather than entering SS status because of the mechanisms of Fast Retransmit and Fast Recovery.

Based on the above two supposes, we can achieve the differential equations of TCP congestion control:

$$\begin{aligned} \dot{W}(t) &= \frac{1}{R(t)} - \frac{W(t)W(t-R(t))}{2R(t-R(t))}p(t-R(t)) \\ \dot{q}(t) &= \frac{W(t)}{R(t)}N(t) - C \end{aligned} \quad (37.2)$$

Among, $p(t-R(t))$ is the packet-loss probability of routing queue, representing the effect of routing congestion control algorithm on congestion window of TCP flows. If the routing queue management algorithm is DropTail, then $p(t)$ is a $(0, 1)$ distribution. In the above formulas, the first describes the changes of congestion window made up of two parts. The first part is $\frac{1}{R(t)}$, representing the linear increase of TCP at CA stage, while the second part of $\frac{W(t)W(t-R(t))}{2R(t-R(t))}p(t-R(t))$ represents congestion window halving after receiving CN signal; the second formula describes the relationship between TCP congestion window and routing buffer formation, $\frac{W(t)}{R(t)}N(t)$ represents the imputing ability of TCP sending end to routing nodes, and C stands for the ability of routing nodes on data processing.

Suppose the fixed points of the formula as (W_0, q_0) and $N(t) = N$, then (W_0, q_0) meets the following equation system:

$$\begin{cases} \frac{1}{R_0} - \frac{W_0^2}{2R_0} p_0 = 0 \\ \frac{W_0}{R_0} N - C = 0 \end{cases} \tag{37.3}$$

Make a Linearization to the differential equations, and get the following equations:

$$\begin{aligned} \delta W(\dot{t}) &= -\frac{N}{R_0^2 C} (\delta W(t) + \delta W(t - R_0)) - \frac{R_0 C^2}{2N^2} \delta p(t - R_0) \\ \delta q(\dot{t}) &= \frac{N}{R_0} \delta W(t) - \frac{1}{R_0} \delta q(t) \end{aligned} \tag{37.4}$$

So the transmission function of the first equation is shown as:

$$H(s) = -\frac{\frac{R_0 C^2}{2N^2} e^{-sR_0}}{s + \frac{N}{R_0^2 C} (1 + e^{-sR_0})} \tag{37.5}$$

Considering about $W_0 \gg 1$ in the ordinary, and it can be know from the formula that

$$\frac{N}{R_0^2 C} = \frac{1}{W_0 R_0} \ll \frac{1}{R_0} \tag{37.6}$$

So we can get the formula of

$$\frac{N}{R_0^2 C} e^{-sR_0} = \frac{N}{R_0^2 C} (1 - sR_0 + \dots) \approx \frac{N}{R_0^2 C}. \tag{37.7}$$

Thus, the transmission function can be simplified into

$$H(s) = -\frac{\frac{R_0 C^2}{2N^2} e^{-sR_0}}{s + \frac{2N}{R_0^2 C}}. \tag{37.8}$$

And then establish the simplified linearized differential equation system as below:

$$\begin{aligned} \delta W(\dot{t}) &= -\frac{2N}{R_0^2 C} \delta W(t) - \frac{R_0 C^2}{2N^2} \delta p(t - R_0) \\ \delta q(\dot{t}) &= \frac{N}{R_0} \delta W(t) - \frac{1}{R_0} \delta q(t) \end{aligned} \tag{37.9}$$

From this way, we can get the transmission function of TCP congestion control shown as following:

$$P(s) = \frac{\frac{C^2}{2N}}{\left(s + \frac{2N}{R_0^2 C}\right) \left(s + \frac{1}{R_0}\right)} e^{-sR_0}. \quad (37.10)$$

Considering further about the second equation in the equation system, and according to the relationship between the flowing rate at data terminal and the queue length of routing nodes as well as considering the factors of queue capacity and bandwidth, we can establish the following:

$$\dot{q}(t) = -1_{q(t)} \bullet C + \sum_{i=1}^N \frac{W_i}{R_i(q(t))}. \quad (37.11)$$

Among, $-1_{q(t)} \bullet C$ represents the capability of data link in data transmission. When the transmission ability of link is stronger than the data sending ability at terminal, the queue length gets to 0. Therefore, only when $-1_{q(t)} \bullet C + \sum_{i=1}^N \frac{W_i}{R_i(q(t))} > 0$, then $\dot{q}(t) = C + \sum_{i=1}^N \frac{W_i}{R_i(q(t))}$, or else $\dot{q}(t) = 0$.

Suppose σ as the sampling time, and then the relationship between average queue and current queue in RED algorithm is as below:

$$q^{ave}(t + \sigma) = (1 - \omega)q^{ave}(t) + \omega q(t) \quad (37.12)$$

And can get the following after derivation,

$$q^{ave}(t) = \frac{\ln(1 - \omega)}{\sigma} q^{ave}(t) - \frac{\ln(1 - \omega)}{\sigma} q(t) \quad (37.13)$$

Finally according to the relationship of packet-loss probability and average queue among RED algorithm, we can reach the equation of $p(t) = L_{RED}(q^{ave}(t) - q_{\min})$, in which $L_{RED} = \frac{p_{\max}}{q_{\max} - q_{\min}}$.

Therefore, combining all the formulas and designating $K = \frac{\ln(1-\omega)}{\sigma}$, we can obtain the linearized differential equation about $p(t)$ and $q(t)$, that is $\delta p(t) = K \delta p(t) - KL_{RED} \delta q(t)$, whose transmission function is $G(s) = \frac{-L_{RED}}{s/K - 1}$.

The linearized differential equations for the entire system are shown as below:

$$\begin{aligned} \delta W(\dot{t}) &= -\frac{2N}{R_0^2 C} \delta W(t) - \frac{R_0 C^2}{2N^2} \delta p(t - R_0) \\ \delta q(\dot{t}) &= \frac{N}{R_0} \delta W(t) - \frac{1}{R_0} \delta q(t) \\ \delta p(\dot{t}) &= L_{RED} K \left(\frac{\delta p(t)}{L_{RED}} - \delta q(t) \right) \end{aligned} \quad (37.14)$$

The generic transmission function is as below:

$$H(s) = A \bullet \frac{e^{-sR_0}}{\left(1 + \frac{s}{2N/R_0^2C}\right) \left(1 + \frac{s}{1/R_0}\right) \left(1 + \frac{s}{(-K)}\right)} \quad (37.15)$$

in which $A = \frac{L_{RED}(R_0C)^3}{4N^2}$ is the coefficient.

37.4 The Stability Analysis of Congestion Control System

From the formula (6.17), we can get three extreme points of system (6.16) which are $2N/R_0^2C$, $1/R_0$, and $-K$. In RED algorithm, as the weight parameter ω of weighted average queue is extremely small, so the absolute value of K compared with $2N/R_0^2C$ and $1/R_0$ is much smaller. It can be considered that $-K$ is the main extreme point of system (6.16). So the Fourier function of transmission function (6.17) at a frequency of $\nu \leq 0.1 \min\{2N/R_0^2C, 1/R_0\}$ can be simplified as below:

$$H(j\nu) = A \bullet \frac{e^{-j\nu R_0}}{1 + j\nu/(-K)} \quad (6.18)$$

According to stability theory, it is required to meets $|H(j\nu)| \leq 1$ and $\angle H(j\nu) \geq -180^\circ$, and the amplitude should be as:

$$\angle H(j\nu) = \angle \frac{\frac{L_{RED}(R_0C)^3}{4N^2}}{1 + j\nu/(-K)} - \nu R_0 \geq -90^\circ - 0.1 \frac{180^\circ}{\pi} \approx -96^\circ \quad (6.19)$$

So, it required that as following: $\frac{L_{RED}(R_0C)^3}{4N^2} \bullet \frac{1}{\sqrt{1+\nu^2/K^2}} \leq 1$.

Therefore, the followings are necessary when we are determining the stability of TCP congestion control system:

Estimate the stabilized frequency upper limit ν_g according to condition of $\nu \leq 0.1 \min\{2N/R_0^2C, 1/R_0\}$.

After fixing ν_g , then we can suppose $K = 0.1\nu_g$.

Obtain the estimated value of L_{RED} according to formula (6.20), and then establish the transmission function of RED algorithm.

Estimate the sampling interval parameter σ according to the bandwidth of routing nodes existed link, and calculate the value of weight parameter ω in RED algorithm.

Determine the upper limit of packet-loss rate, obtain the differential value between q_{\min} and q_{\max} according to L_{RED} , and finally the parameters in the whole TCP-RED congestion control system have been fixed up.

37.5 Summary

The congestion control system, as one of the key parts for the normal operation of computer network, has become a hot research in both domestic and abroad at currently. The preliminary research on congestion control system of computer network was focused on the improvement of algorithm and some researches on network behavior. This paper, based on some relative literatures, carried out a comparatively intensive research on the aspect of network congestion control and built up a continuous time model for TCP congestion control system. TCP congestion control algorithm and RED queue management algorithm, as the two parts of TCP congestion control system, are modeled independently and then cascaded into a unity. Finally, it analyzed the model stability near the stationary points, and pointed out that the stability of TCP congestion control system mainly depended on the RED algorithm.

References

1. Athuraliya S, Li VH, Low SH et al (2008) REM: Active queue management. *IEEE Netw* 15(3):48–53
2. Kunniyur S, Srikant R (2009) Analysis and design of an adaptive virtual queue (AVQ) Algorithm for active queue management. *ACM Comput Commun Revi* 31(4):123–134
3. Yamagaki N, Tode H, Murakami K (2009) RED method with dual-fairness metrics cooperating with TCP congestion control. Institute of Electrical and Electronics Engineers
4. Durreli A, Barolli L, Sridharan M, Chellappan S et al (2006) Load early detection (LED): a congestion control algorithm based on routers' traffic load. *IPSIJ Digital Carrier*
5. Vishal M, Wei-Bo G, Don T (2008) Fluid-based analysis of a network of AQM routers supporting TCP flows with an application to RED. In: *Proceedings of ACM/SIGCOMM 2008*

Chapter 38

The Realization of OPNET and MATLAB Co-Simulation Based on HLA

Yu Zhang, Zhen Yao and Xiangming Li

Abstract In some cases, it is difficult to simulate the communication behavior of a wireless Network precisely only using MATLAB or OPNET. In this chapter we will use High Level Architecture to achieve the OPNET and MATLAB's co-simulation.

Keywords MATLAB · OPNET · Co-simulation · HLA · Wireless

38.1 Introduction

There are a lot of times you need a very detailed physical layer simulation. It may fail to meet requirements of the simulation accuracy, only by OPNET's Pipeline Stage mechanism. MATLAB facilitate these kinds of simulation using its own simulation models. The High Level Architecture (HLA) is software architecture for creating computer simulations out of component simulations. Here we will discuss how to use HLA to achieve OPNET and MATLAB's co-simulation.

Y. Zhang (✉) · Z. Yao · X. Li
School of Information and Electronics Beijing Institute of Technology, Beijing, China
e-mail: YuZhang@bit.edu.cn

Z. Yao
e-mail: rain_china@yeah.net

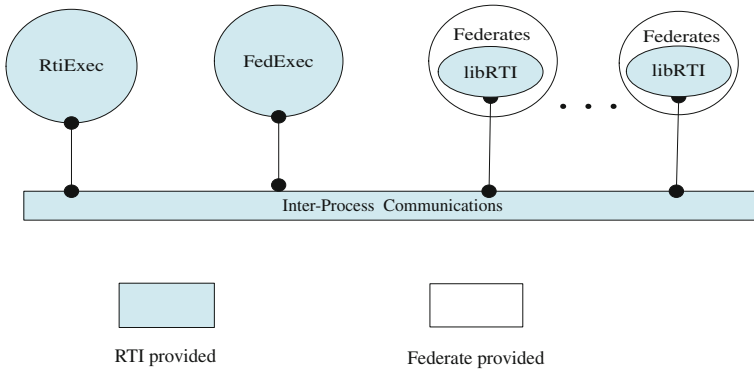


Fig. 38.1 Logical view of an HLA federation

38.2 The Mechanism of High Level Architecture

The HLA provides a general framework within which simulation developers can structure and describe their simulation applications. The HLA consists of three components: HLA Rules, Interface Specification and Object Model Template (OMT). HLA is not an implementation, it provides a framework.

In fact, the Run Time Infrastructure (RTI) is the implementation of HLA. Figure 38.1 shows a high level, logical view of an executing HLA Federation. All of the components shown in the figure are part of a single federation with the exception of the RtiExec.

38.3 The Framework and Federates of Co-Simulation

38.3.1 The Framework of Co-Simulation

OPNET or MATLAB should be as a federate to join the federation execution. In fact, they firstly communicate with RTI through RTIAmbassador and Federat-Ambassador. And RTI will coordinate all federates to transfer messages as Fig. 38.2 shows:

38.3.2 The OPNET Modeler Federate

The High-Level Architecture (HLA) functionality in OPNET Modeler enables you to include a simulation as one of the federates in an HLA federation, and thus to both affect—and be affected by—events in other federates.

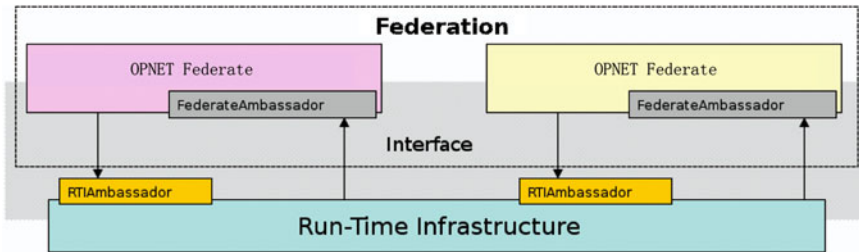


Fig. 38.2 Framework of co-simulation

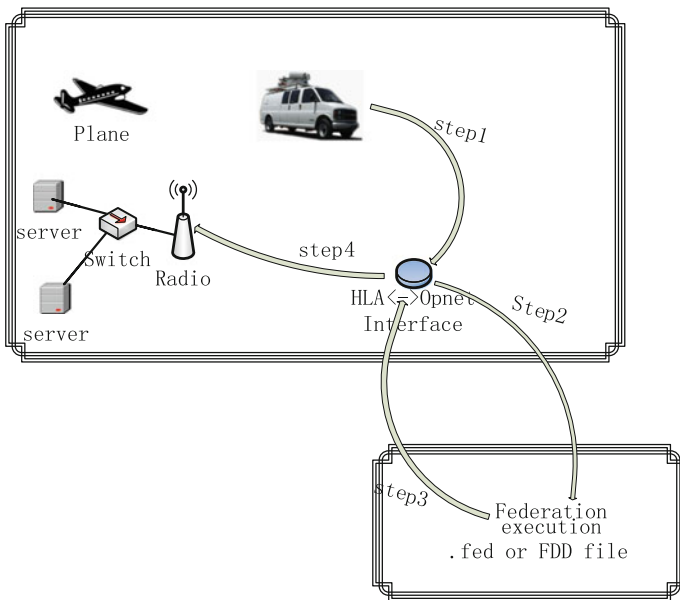


Fig. 38.3 OPNET modeler federate

Figure 38.3 shows the topology of an example OPNET Modeler federate and illustrates how events flow between the OPNET Modeler federate and the federation.

Step 1. Events flow from the OPNET Modeler network nodes to the HLA-OPNET Interface node. Step 2. Events then flow to the federation’s RTI. Step 3. HLA flow from the federation to the HLA-OPNET Interface node. Step 4. HLA events the flow to other nodes of the network model.

An OPNET Modeler federates is composed of the following elements:

A network model containing instance of the node and link models, plus the special HLA interface node. Node and link models implementing the simulation

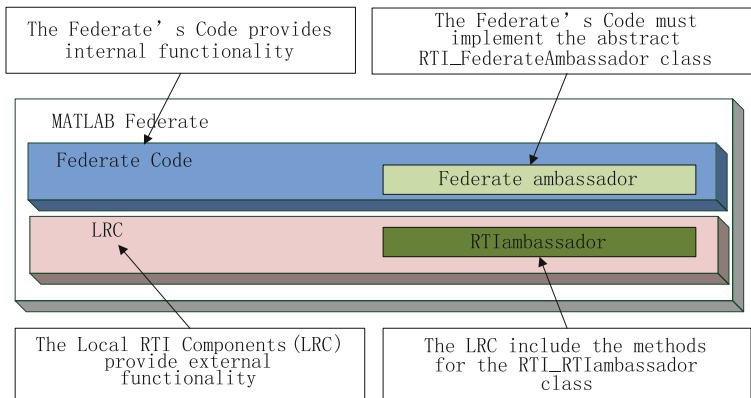


Fig. 38.4 The components of single federate

behavior, HLA interactions, and HLA outgoing messages. A special HLA interface node model. This node contains a factory function responsible for creating an instance of a Federate Ambassador for the simulation.

An Simulation Object Model (SOM) file specifying the HLS interface of the federate.

The .fed or FDD file, shared by all federates.

38.3.3 The MATLAB Federate

MATLAB does not directly provide simulation models to support HLA. But we can use mixed programming of MATLAB and C++ to achieve our aim.

LibRTI(The RTI Library) makes HLA service methods available to federates, and It is written in C++. User's codes for a federate is linked with Local RTI Component Code (LRC) from the C++ library libRTI to form a complete federate. These local RTI components provide the services for the federate through communication with the Rtiexec, the FedExec and other federates. The components of a single federate are shown in Fig. 38.4:

LibRTI contains the RTIambassador class, which is used to provide RTI services to the federate. It also contains the abstract class FederateAmbassador which defines all of the interfaces to the callback functions but cannot be used until actual implementations of these functions are provided by the federate. Class RTI_FederateAmbassador is an abstract class, as it contains (in the C++ implementation) pure virtual functions. So we can use C++ to realize the implementation of the virtual function to get the service.

At the same time, MEX-files in MATLAB are a way to call your custom C, C++ routines directly from MATLAB as if they were MATLAB built-in functions.

The gateway routine to every MEX-file is called `mexFunction`. This is the entry point MATLAB uses to access the DLL. In C/C++, it is always:

```
mexFunction(int nlhs, mxArray *plhs[],
int nrhs, const mxArray *prhs[])
{
//add code here
}
```

An MATLAB federates is composed of the following elements:

Mexw32 file which packages C++ code to get RTI services, and will be called by MATLAB. In fact it is an interface between MATLAB models and RTI.

Mathematical models in MATLAB.e.g: physical layer or link layer of wireless network.

The .fed or FDD file like in OPNET Modeler federate.

38.4 Example of OPNET and MATLAB's Co-Simulation

Here is a simple example using OPNET and MATLAB's co-simulation. The example has two federates: an OPNET Modeler federate (OPNET) and a controller federate (MATLAB). The OPNET federate contains two nodes representing planes. The movements of these planes are controlled by the controller federate.

The controller federate advances federation time, sends interactions that modify the bearing and ground speed of the planes in the OPNET federate, and receives interactions from the OPNET federate.

38.4.1 The Realization of OPNET Federation

Launch OPNET Modeler and make sure that Wireless functionality is available. Then create a new project with an empty scenario, world scale. Add two plane nodes and a HLA interface to the scenario.

Create a new FDD file with two interaction classes: Order and Order_Reply.

Create a mapping file in order to map Order class to Plane_Order and Order_Reply to Plane_Ack.

Each of the node could get the order and then give the control a replay. The code just like blow:

```
/* extract information from packet */
pk= op_pk_get (/*parameters of stream index*/);
op_pk_nfd_get_dbl (pk, /*parameter*/, /*parameter/);
op_pk_nfd_get_dbl (pk, /*parameter*/, /*parameter/);
op_pk_destroy (pk);
```

```

/* update node attributes */
op_ima_obj_attr_set_dbl (/*parent id*/,/*parameter*/, /*parameter/);
op_ima_obj_attr_set_str (/*parent_id*/,/*parameter*/, /*parameter/);
/*send interaction*/
pk = op_pk_create_fmt (/*name of ack*/);
op_pk_nfd_set_int32 (pk, /*parameter*/, /*parameter/);
op_pk_nfd_set_dbl (pk, /*parameter*/, /*parameter/);
op_pk_nfd_set_dbl (pk, /*parameter*/, /*parameter/);
op_hla_interaction_pk_send(pk,/*time*/);

```

38.4.2 The Realization of OPNET Federation

Use C++ to realize the related services. We need MATLAB to create the federation, then join it, publish and subscribe interaction class. After the initialization, MATLAB could declare its Time Strategy. Then we need mexFunction to package all the services in C++.

The code just like below:

```

void mexFunction(int nlhs, mxArray *plhs[], int nrhs, const mxArray *prhs[])
{
/*Judge the parameters*/
int num_parameters;
num_parameters = (int) mxGetScalar(prhs[0]);
switch (parameters)
{
/*call proper Func.
Func should have the rule like example below */
}
return;
}
/*for example: how to join the federation in C++ code*/
void connecthla(int nlhs, mxArray *plhs[], int nrhs, const mxArray *prhs[])
{
    (*rtiAmb).createFederationExecution(tFederationName, tFedFile);
    (*rtiAmb).joinFederationExecution(tFederateName,tFederationName,
    federateAmb);
}

```

In MATLAB we add the interface to its search path. So MATLAB could call the functions by the interface.

The code in MATLAB likes below:

```

/*create federation*/
e=createFedExec(FederationName,FedFile);
/*we can use the function in interface just like MATLAB inner functions*/

```

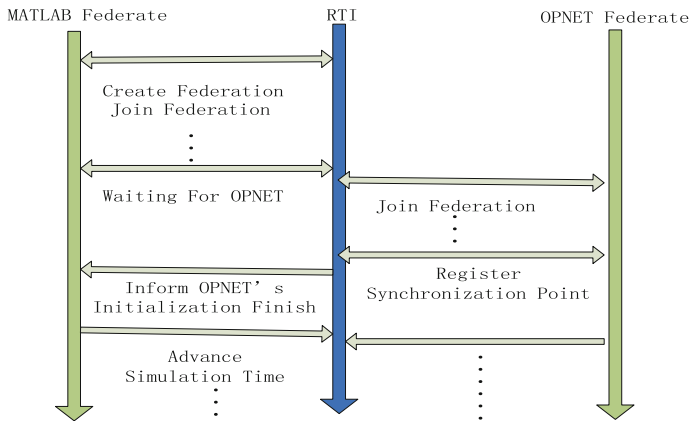


Fig. 38.5 The process of co-simulation

38.4.3 The Process of Co-Simulation

At first MATLAB create the federation; finish the initialization and then wait for OPNET to join the federation. When OPNET finish the initialization, they start to advance simulation. The process just like Fig. 38.5:

Acknowledgments Thanks my advisor, Yu Zhang, who has been very nice and patient to me during the process of writing. Also, I would thank my fellow classmates, Zhiqiang Sun, and friends, Peng Wang. They give a lot of help when I met difficulties.

References

1. OPNET (2003) Modeler Product Documentation Release 10.0
2. U.S. Department of Defense (DMSO) (1998) High level architecture rules, high level architecture federate interface specification high level architecture object model template specification version 1.3
3. DMSO (1999) Federation development and execution Process Model Version 1.5
4. Bates Jon, Tompkins Tim (2000) Applied Visual C++6.0 tutorial. Tsinghua University Press, Beijing, pp 103–105
5. Li C (2004) The interface between Matlab and C Language. University of Posts and Telecommunications Press, Beijing, pp 92–95
6. Dahmann JS (1997) High level architecture for simulation. In: Proceedings of the 1st international workshop on distributed interactive simulation and real time application, p 9–14
7. Harding C, Griffiths A, Yu H. An interface between MATLAB and OPNET to allow simulation of WNCS with MANETs, IEEE

Part IV
Data Modeling and Programming
Languages

Chapter 39

Design and Applications of Third Party Integrated Information Security Management System

Yuefeng Fang, Xiaoyong Wang and Fei Gao

Abstract This chapter mainly analyzes the current status of information security, discusses the domestic and international trends of its technology development, and proposes the third-party integrated information security management system. And by analyzing the services used currently, it also generally introduces the application design of information security management system, such as planning and design, whole framework design, and so on. Finally the planning and the design of the system is presented.

Keywords The third party · Integrated information security management system · Designs

39.1 Introduction

Information Security Managed Service is the comprehensive and systematic service that fulfill internal information system in enterprise with a third-party service [1–4]. According to the monitoring report by CNERT/CC, in 2007 the growth rate of phishing incidents in Chinese Mainland was 140%, that of vicious webpage code incidents was 260%, that of the tampered-web sites was 150%, and that of the Trojan-implanted hosts was 2,200%. The security threats are increasingly serious and the complexity of security management increases synchronously. The traditional single and simple network security concepts and tools can not keep

Y. Fang (✉) · X. Wang · F. Gao
Department of Computer and Information Science, No.8 Qianhunanlu Road, Ningbo,
315000, Zhejiang, People's Republic of China
e-mail: fyfywf@sina.com

the information system from the new, dynamic and intelligent network security threats well. All of those show the importance of establishing the Information Security Management System [5]. The third-party integrated information security management system is mainly to monitor and manage the incidents of users' information security punitively, and to provide users with means to build various information security services.

39.2 Technology Trends at Home and Abroad

To build an intelligent active defense system the following key technologies must be involved: security-evaluation on information system, real-time processing ability to mass events, intelligent aided decision-making ability, and so on. Those key technologies are recently the important research directions of information security fields. From the practical point of view, the biggest problem to construct the security system currently is not the lack of security product, but that the products cannot make a timely reasonable response to security situation of the network. Domestic and foreign experts and scholars have done extensive researches and explorations on information security management and monitoring, and the content of service, which are concentrated in the following areas:

39.2.1 Security Risk Analysis Service

National certification organizations and risk identification and certification systems are founded in 1970 and 1980s in these IT-developed countries like America and Canada. They are responsible for researching and developing the evaluation criterion, evaluation and certification methods and evaluation techniques, and do information security identifications and certifications based on evaluation standard [6, 7]. In China, the study on information system risk evaluation has just been started in recent years. It focuses mainly on the establishment organizational structure and services system. The relevant standard and technology systems are still in research stage.

39.2.2 Security Evaluation Services

Nowadays, the popular security evaluation services are ISO/IEC TR 13335, Operational Key Threat, Assets and Vulnerability Evaluation Methods and Processes (OCTAVE), Comprehensive Analysis of Network Security based on index analysis, Network Security Evaluation based on fuzzy evaluation and Evaluation Methods based on knowledge [3, 4].

39.2.3 Network Vulnerability Analysis Services

From a technical perspective, in the field of Network Vulnerability Analysis and Information System Security Strengthening, the existing vulnerability scanners, such as NESSUS and ISS Internet Scanner, are used to vulnerability-scan the network for specified single or multiple hosts. But they lack of lack of a comprehensive analysis of network security situation. Therefore, we need to integrate the comprehensive information of security configuration, such as the vulnerabilities of target network, network services, physical links and access rights, to construct the possible systematic attack scenarios (i.e. the attack path set), and to automates the generation of attack scenarios and analysis procedure.

39.2.4 Network Security Incidents Processing Services

Security incidents processing is the core of network security services. Security incidents reflect the real-time status of the intrusion existing in the network. The following are the current security incidents processing:

Security Situation Evaluation

It organically integrates the protection, monitoring, and emergency response of the network security system. The main function is to comprehensively evaluate the network security status and transforming trends, that is to provide users with an accurate evaluation of network security status and its developing trend according to the history record of network security properties, and to enable the network managers to make decisions and protection purposefully.

Security Incidents Correlation

There are many intrusion detections, which are based on alert correlation in research. The representatives are EMERALD from America and MIRADOR System from France. EMERALD has Alert thread and Meta alert. The former is made up with the alarms which belong to the same attack and sensor. The later is constituted with one or more of the alarm correlation from heterogeneous sensors [5].

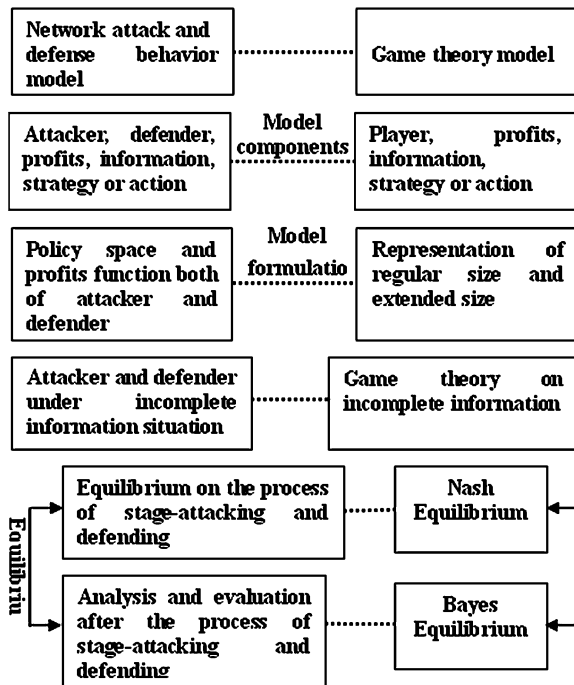
39.2.5 Intelligent Decision-Making and Active Defense Services

It includes the technologies as active acquisition of attack information, automatic decision-making based on artificial intelligence, and security enhancement, and so on.

Active Security Pre-alarming Technology

It includes the following tow technologies: attack information active acquisition technology which is based on Honey pots and active pre-alarming and defense

Fig. 39.1 Corresponding graph of network attack and defense



technology for large-scale malicious attacks. There are only a few comprehensive studies on the early warnings of large-scale DDOS attacks and worm attacks at home and abroad. They just design appropriate early warning protection methods mainly for the a particular characteristic of large-scale malicious attacks, such as SYN Caches and SYN Cookies methods for SYN Flood Attack [4].

Security Resources Active Response Technology Based on a Game Theory

The behavior models of network attacking and defenses based on game theory explore how to use the principles and methods of game theory. By establishing the profitability function of offence and defenses and related personals, and analyzing the possible balance and strategy choice of all the players in the offence and defenses game, the theory has certain strategically directive significance to network security management. The direct study of game theory in the field of network security applications has just begun at home and abroad. Researchers take the network attack and defenses as infinitely repeated incomplete information game with perfect information, and get the following correspondences (Fig. 39.1).

Active Security Enhancement Technology

It includes security patches automated publishing technology and key service survivability enhancement technology. All of these patch managements and distribution systems are based on Microsoft’s SUS, and can basically meet needs of patch distribution, testing and management. But information source server has become the main security and efficiency bottleneck in large-scale network, because the whole structure of the system are based on C/S Mode. The key service

survivability enhancement technology mainly focuses on the service survivability enhancement technology based on Dynamic Drift. Its main idea is to transfer user's service requests to other similar server components by drift technology, when the components that provide services break down, such as meeting attack, error, natural disasters. That will allow the system to provide services incessantly.

39.3 Application Design of Information Security Management System

39.3.1 Planning and Design of the System

Information Security Management Comprehensive System realizes its management on the following three levels from top to bottom: Security Management Center, Security Management Domain and Management Sub domain. Management Sub domain is responsible for collecting and managing the information of security devices and servers in the managed sub domain. Security Management Domain arranges security management information collected from multiple management sub domains. Security Management Center real-timely monitors and manages all of the security equipments and the core servers in the business network through Security Management Domain and Management Sub domain, such as firewall, intrusion detection systems, virus prevention systems, network audit system, host audit system, data security systems, and so on. The total Security Management Server is placed center machine room. And security engine is placed in the network segment monitored by user. The following Fig. 39.2 describes the architecture of information security management system.

Information security management system is hierarchically deployed in the security management servers of each security management sub domains, security management domains and security management center. It is a distributed processing system based on Web processing mode. Security management server is the core of the system, and runs the background functional modules of event-collection, correlation, matching, trend calculation, and so on. Through the four types of interfaces, security management server allows the higher level to issue security notice and tactics, and do information query to the lower level. It also makes sure the lower level to get notice and tactics, response the query requests, and send event alarm, early warning, situation data and security reports to the higher level. By the further secondary development, security management server can exchange information with the external systems, such as GIS systems, network management systems, command and control system, and so on.

The server of security management center provides the professional security management staff with management and control platforms: Security Management Client and Situation Awareness Client. The former is a browser faces to management and maintenance of the security manager, it provides the user with the

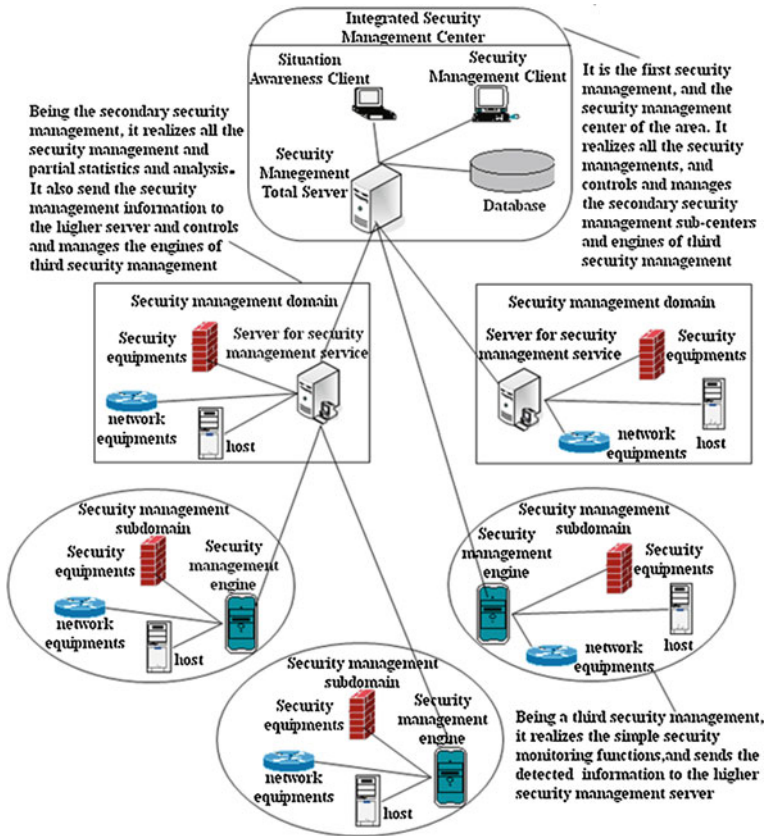


Fig. 39.2 Deployment diagram of information security management system

functional interfaces forewent alarms, query, strategy development, deployment and change. The later provides user with trend analysis, display, alarm, which realizes the functions of strategy development, deployment and change, and so on.

39.3.2 Whole Framework Design of System

The whole framework design of information security comprehensive management system is showing as Fig. 39.3.

As Fig. 39.3 shows, information security comprehensive management system provides the following functions:

Topology Discovery

Automatic Topology Discovery Module can realize automatic topology discovery, and organizes the formation of the topology automatically. Network

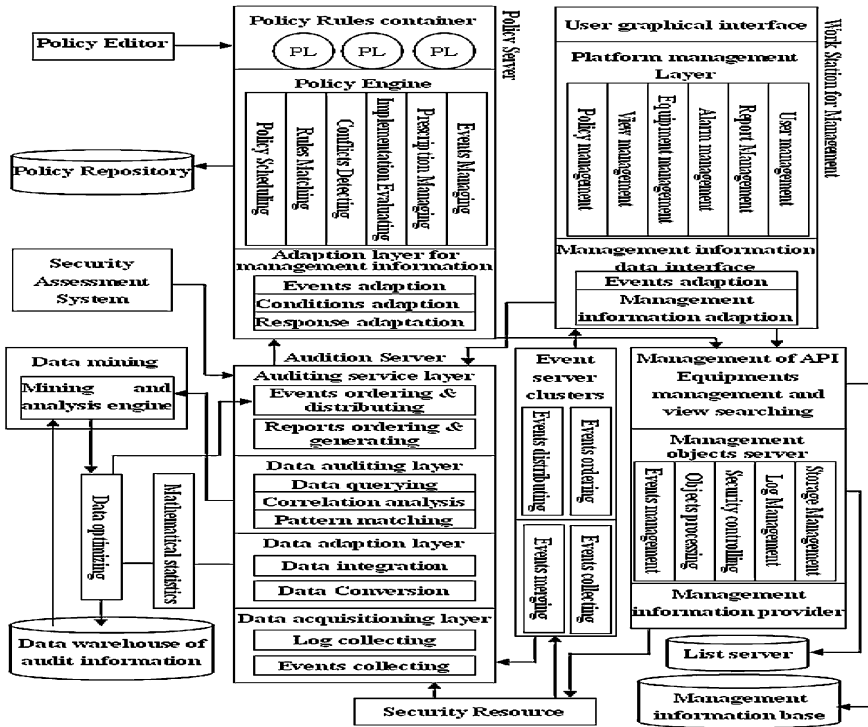


Fig. 39.3 Functional structure chart of information security management system

topology shows operation conditions of the entire network clearly, distinctly and hierarchically. It also and provides information for assets management according to network topology discovery.

Assets Management

Assets Management includes all management of the following hardwires and software's: network equipment, security equipment, servers, computer terminals, software products installed, and so on. Combined with topology management, it manages the relevant network equipment, security equipment, terminals and servers in network topology. Device management supports the combination of automatic discovery and manual input. It automatically collects all kinds of hardware and software configuration information of each equipment. It also records changes in hardware and software equipment and timely records the asset movements to provide the customer with real-time, detailed hardware and software information of network assets. By statistical analyses, a unified asset usage report is formed to provide the reference for hardware and software upgrades. It gives fund input for rationalization information developments to avoid duplication of investment and construction. As to each device, assets management module manually inputs the asset tag number, equipment procurement time, production units and contact methods, and the name, e-mail address and telephone number of

the users and organizations. It provides users of different levels with assets inquiries within corresponding operation right and the reports related to various assets. It provides scientific basis and decision support for the customer's normalization, too.

Equipment Monitoring

Equipment monitoring module timely monitors the the network equipment, safety equipment, some of the main server and host resources within management domain. Its monitoring range covers servers, part of the hosts, routers, switches, IDS, firewalls, vulnerability scanners, and so on. By updating the information of equipments running status and coordinating with policy management system, it responses automatically and timely when device status information exceeds the set point. Through a lot of analysis and integration of the data collected, various reports are formed, and large-screen displayed in graphics, such as pie chart, bar graph, curve, and so on.

An equipment resource includes the followings (the corresponding parameters should be confirmed by the equipment providers): Server, Service survivorship, Host, Switch, Router, IDS, Firewall, Service monitoring.

Security Situation Analysis

The main function of security situation analysis is to format-unify and standardize the massive original alarm data reported by distributed heterogeneous security devices, and to do correlation statistical analysis and data mining. The original information is from management equipments, security devices, monitoring device, an so on, in the network. Proper mathematical treatment and integration to the data can provide the number and chart that reflects the running status of network. Situation analysis module of information security integrated monitoring management system can provide the following functions:

- (1) It provides user with an accurate assessment to the current state of network security, and helps administrators to know the security threat status of system.
- (2) It provides the historical evolution of security status, and helps the user to find the rules of security evolution according to the security status changes over time.
- (3) It analyzes the development trend of network security, and helps network administrator to predict potential network security problems in the future.
- (4) It provides the evaluation index and relevant evaluation algorithm of security status. A multi-level, multi-angle security status evaluation platform is built to reduce the heavy burden of data analysis, so as to release administrator from the mass, noise-charged data.

Event-handling

Event-handling includes the security events discovered by equipment monitoring and reported by security equipments. By analyzing the reported security events, event-handling module can do a series of measures, such as pretreatment, fusion and correlation analysis. So that the redundant security events can be removed, and reconstruct the attack scenario by identifying the multi-step attacks. It timely detects and alarms the known security events. As for some potential

security risks and anomalies, it can notify the relevant managers promptly. The later can maintain and update security systems. Alert information does separate alarm and automatic processing according to different administrative domains, types and threats, to provides events query and statistical analysis.

Log Audit

Log audit module provides log collection tools to all kinds of network equipments, security devices, operating systems and application systems. It also provides synthetically auditing operations to all kinds of log information, such as uniform query, stats and ranking, etc. By uniform querying the log information, administrators can view the running status of monitored equipments under specified circumstances, and do further analysis on network security status associated with the operating conditions of equipments. By stats and ranking, administrators can easily access statistical ranking of network addresses and security events that expiated security records during query time, so as to guide the work of the relevant technical staff. Through the log cross-query among the equipments, alliance of each equipment can be confirmed to provide references for upgrades and replacement of network equipments. Log audit module combines all algorithms of event-handling module, and does deep audit analysis on log information, such as merging, fusion, correlation. It provides the administrators with a method to command network security situation from the point of auditing analysis. It also provides a reasonable “downsizing” program for log information if the historical the log information is too bulking, which is to statistically compress the temporarily unneeded information. By setting appropriate security policy, it send alarms to the administrator timely when a suspicious or abnormal condition happens to the audited results. The audited results output log-query results in the form of report.

Reporting Function

Reporting function module can analyze the latest dynamic data real-timely, by depth statistical analysis, it generates intelligible and straightforward graphical reports. It also provides flexible data so that the final generated report can easily represent wide spectrum statistical results of source data. Reporting function module can provide definable reports, which is to say, the time, objects and statistical results of the report is definable. So that the final generated report can produce appraisal results according to the characteristics of statistical information, such as pie chart, bar charts, graphs, etc.

Policy Management

The integrity of network security needs a unified security policy and workflow-based management. With policy management, the unified security policy can be provided for whole network security administrators so as to guide security management agencies at all levels to do the deployment of security policy in the light of local conditions and improve the whole network’s security and defense capability. Meanwhile, it can also further improve the construction of security policy system for IP network and provide guidance to all security work, which effectively avoids the security risks caused by lacking of password, authentication and access control policy, and so on.

39.4 Conclusion

To sum up the above arguments, the application design of integrated information security management system should fully consider the current advanced concept of information security management, monitoring and security services, and adopts the core technologies with independent intelligent property, such as management, monitoring and services technologies. It uses the output of the information system's daily monitoring, information collection and risk assessment as the priori knowledge for the whole system and uses the event processing as the principle of decision-making active defense. At the same time, the appropriated intelligent decision-making mechanism is established to achieve integrated management and monitoring to network security events combining with the knowledge base. And with the establishment of professional information security services, it provides comprehensive management and service to the security events beforehand, middle and afterwards. It builds the active defense system and provides the user with multi-angle, multi-faceted information security management, monitoring and emergency response services.

Acknowledgments Author Introduction Yuefeng Fang (1965–), male, born in Chaohu Anhui, graduated from Zhe Jiang Industrial University with the major of Information Science. Being a senior engineer, and mainly engaged in information engineering study.

References

1. Levin D (2003) Lessons learned in using live red teams in IA experiments. DARPA information survivability conference and exposition (DISCEX'03), vol 1. IEEE Computer Society, Washington DC, pp 110–119
2. Noel S, Jajodia S (2004) Managing attack graph complexity through visual hierarchical aggregation. In: Proceedings of ACM CCS workshop on visualization and data mining for computer security, Fairfax, pp 109–118
3. Fox KL, Henning RR, Farrell JT et al (2003) A prototype next generation information assurance tool: a data fusion model for information system defense. In: Proceedings of 7th world multiconference on systemics, cybernetics and informatics (SCI 2003), vol 1. Orlando, pp 284–289
4. Ramakrishnan CR, Sekar R (2002) Model-based analysis of configuration vulnerabilities. *J Comput Secur* 10(12):189–209
5. Zou CC, Towsley D, Gong W (2007) Modeling and simulation study of the propagation and defense of internet e-mail worm [J]. *IEEE Trans Dependable Secur Comput* 4(2):105–118
6. Cavusoglu H (2004) A model for evaluating IT security investments. *Commun ACM* 47(7):87–92
7. Chan MT, Kwok LF (2001) Integrating security design into the software development process for e-commerce system [J]. *Inf Manag Comput Secur* 9(3):112–122

Chapter 40

A Protocol for a Message System for the Tiles of the Heptagrid, in the Hyperbolic Plane

Maurice Margenstern

Abstract This chapter introduces a communication system for the tiles of the heptagrid, a tiling of the hyperbolic plane. The chapter focuses on a short account of an experiment obtained by running a simulation program.

Keywords Hyperbolic tilings · Cellular automata · Applications

40.1 Introduction

In this chapter, we present a protocol to manage communications between tiles of a tiling of the hyperbolic plane. Why the hyperbolic plane? Because the geometry of this space allows to implement there any tree structure. The reason for that is that the simplest tilings defined in this plane are spanned by a tree. We refer to the author's chapter and books on this topic, see [5, 6]. Tree structures are already intensively used in computer science, especially by operating systems. But the fact that trees are naturally embedded in this geometrical space, especially on tilings living there, was never used. This chapter proposes to take advantage of this property.

In [Sect.40.2](#), we briefly indicate how to model the tiling of the hyperbolic plane in which we make the simulation announced in the title. We sketchily describe the navigation technique with a mention to a new aspect. In [Sect.40.3](#), we present the protocol which allows us to improve the system briefly mentioned in [7].

M. Margenstern (✉)

Professor Emeritus of Université Paul Verlaine—Metz LITA, EA 3097 UFR-MIM,
and CNRS, LORIA Campus du Saulcy, 57045 Metz, Cédex, France
e-mail: margens@univ-metz.fr

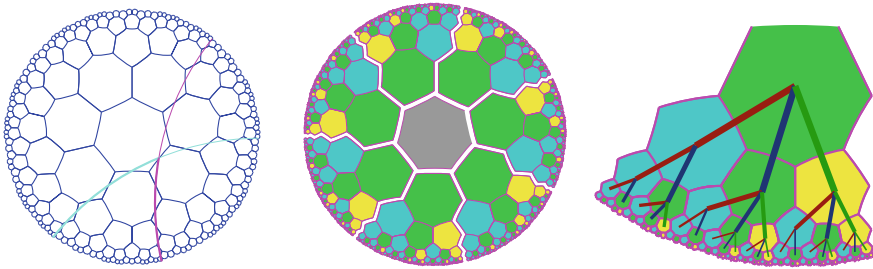


Fig. 40.1 *Left-hand side*: the mid-point lines. This tool which shows how a sector spanned by the tree is defined in the heptagrid. *Middle*: first part of the splitting, around a central tile, fixed in advance, seven sectors. Each of them is spanned by the tree represented in the right-hand side. *Right-hand side*: the tree which spans the tiling

In [Sect.40.4](#) we present the experiment which was performed by a simulation program. In [Sect.40.5](#), we conclude with further possible development of the scenario implemented by the protocol.

40.2 The Heptagrid and How To Navigate There

Discovered around 1829, hyperbolic geometry put an end to the more than two thousand years search of mathematicians to derive one axiom of Euclid's *Elements* from the others, see [\[6\]](#). Around 50 years later, Poincaré devised a model for this new geometry which is still in use nowadays, the so called Poincaré's disc model. Poincaré also proved that there are infinitely many tilings in the hyperbolic plane constructed by the following simple process: take a convex regular polygon P , replicate P in its sides and, recursively, the images in their sides. For any number of sides at least five and infinitely many values of the interior angle between two sides of P , the process covers the plane and tiles never overlap.

We shall consider a particular tiling constructed on the heptagon with an interior angle of 120° , the **heptagrid** [\[3\]](#). The left-hand side and middle pictures of [Fig. 40.1](#) give a representation of this tiling.

The tiling looks very much the regular hexagonal tiling of the Euclidean plane. However it seems difficult to locate the tiles, even on this small figure. How to overcome this problem? In fact, the tiles of each sector defined in the middle picture of the figure are in bijection with the nodes of the tree represented in the right-hand side picture. This key property and specific features of this tree allowed the author to introduce a system of coordinates for the tiles of this and other tilings of the hyperbolic plane with the following property: for each tile, it is possible to compute in linear time with respect to the size of its coordinate a path from the tile to the root of the sector in which the tile lies, see [\[2, 5, 6\]](#). We have a stronger result:

Theorem 1 (see [6])) *There is an algorithm which computes a shortest path between two tiles of the heptagrid which is linear in the size of the coordinates of the tiles.*

The computer program already mentioned in the abstract implements an algorithm satisfying the statement of Theorem 1.

40.3 The Communication Protocol

In Sect.40.1 we mentioned that in [7], we already proposed a communication protocol for the tiles of the heptagrid. This protocol was based on a specific system of coordinates, inherited from [4, 6]. We briefly describe this system in Sect 40.3.1 Then, in Sect.40.3.2 we define the new protocol.

40.3.1 Absolute and Relative Systems

The **absolute** system is based on a numbering of the sides for the tiles of the heptagrid. For each tile, we number the sides from 1 to 7 in this order while counter-clockwise turning around the tile. Now, how to fix side 1? We again take the situation of the middle picture of Fig. 40.1: a central cell surrounded by seven sectors, each one spanned by a copy of the Fibonacci tree. Now, side 1 is fixed once and for all for the central tile. For the other tiles, side 1 is the side shared by the tile and its father, considering that the central cell is the father of the root for each copy of the Fibonacci tree spanning the sectors.

Now, a side always belongs to two tiles and so it receives two numbers. This is why this numbering is called **local**. However, the association between both numbers is not arbitrary. We give here the possible associations:

(1,3), (1,4), (1,5), (2,6), (2,7), (3,7), (3,1), (4,1), (5,1), (6,2), (7,2), (7,3), where given the left-hand side number of the parenthesis, the right-hand side number is the number in the other tile.

The local numbering gives a way to encode a path between two tiles A and B . Let $\{T_i\}_{i \in [0..n]}$ be a shortest path from A to B . This means that $T_0 = A$, $T_n = B$ and that T_i and T_{i+1} for $i \in [0..n - 1]$ share a side. Denote this side by s_i . Let a_i be the number of s_i in T_i and b_i be its number in T_{i+1} . Then we say that the sequence $\{(a_i, b_i)\}_{i \in [0..n-1]}$ is an **address of B from A** . The reverse sequence gives an address of A from B . To distinguish the ends of the path, we proceed in a slightly different way. We associate the above numbers to the tiles: the tile T_i has an entry side denoted by en_i and an exit one denoted by ex_i . Accordingly, an address or a **coordinate** of B from A is the sequence $\{(en_i, ex_i)\}_{i \in [0..n]}$, with $en_0 = 0$, $ex_n = 0$, $ex_i = a_i$ for $i \in [0..n - 1]$ and $en_i = b_{i-1}$ for $i \in [1..n]$.

Now, how to define a shortest path between A and B ?

There are two ways: the first way is given by Theorem 1.

The second way consists in the following. If A sends messages to every tile, it considers itself as the central tile, taking its own number 1 as the number 1 of the central cell. Remember that all tiles have the same size, the same shape and the same area, this is why each tile may feel ‘equal’ to the others. When it sends the message to its neighbours, it also sends them the information that it is the central cell and it sends $(0, i)$ to its neighbour i . And so, the neighbour receives its coordinate from A . By induction, we assume that each tile T which receives the message from A also receives its coordinate from A and its status in the relative tree to A in which T is. From this information, and as T knows from which neighbour it receives the message, T knows which of its neighbours are its relative sons and so, it can append the element (en_T, ex_T) to the address it conveys to the corresponding son together with the relative status of the son as the tree has two types of nodes. And so, we proved that each tile receiving a message from A also receives its address from A and its relative status with respect to A . The local numbering attached to A as a central tile is called the **relative** system. Now, from its coordinate from A , T may compute the coordinate of A from T in a linear time in the size of the coordinate, as follows from what we already have noticed. And so, if T wishes to reply to the message sent by A it can do it easily. Moreover, from the properties we have seen in Sect.40.2, we can see that, proceeding in the just described way, a public message is sent to every tile once exactly, which is an important feature.

40.3.2 The Protocol

We distinguish two types of messages, **public** ones and **private** ones. By definition, a public message is a message sent by a tile to all the other tiles. A private message is a message sent by a tile to a single other one. This distinction belongs to the sender of the message. In this protocol, we assume that we have a global clock defining a discrete time and that a message leaving a tile T at the time t can reach only a neighbour of T at time $t + 1$. We say that the maximal speed for a message is one.

The public message makes use of the relative system of the sender. However, in the coordinates constructed by the tiles which relay the message, the numbers en_i and ex_i computed by the relaying tile are defined according to the absolute system: the tile does not know where is the sender and its own local numbering is defined by the absolute system.

A private message is either a reply to a message, either public or private, or a message sent to a single tile according to the following procedure. Each tile T has a direct access to the system. Given the coordinates of a tile N as defined in Sect.40.2, the central cell being that of the absolute system, the managing system gives to T a shortest path from T to N which is a coordinate of N from T . And so, a private message is defined by the fact that it has the address of the receiver.

Now, a private message from A to B stores the coordinate of B as two stacks a and r . The stack a is for the direct run from A to B , the stack b is for the way back. Each tile T on the path conveys the message to the next one N on the path, in the direction from A to B . To perform this, T reads the top of a , say (en, ex) . It knows that ex is the number of N from itself. Just before sending N the message and the stacks, T pops the top (en, ex) of a and pushes (ex, en) on the top of r . In this way, T knows that it is the receiver if $ex = 0$. When this is the case, T pops the top (en, ex) of a , pushes (ex, en) onto r but does nothing else. When T is ready to answer, it exchanges a and r and so, the same process allows the message to reach A together with a coordinate of B from A .

It is easy to see that this process is linear in time with respect to the coordinate of the receiver, assuming that all messages travel at a constant speed.

However, we cannot assume that all tiles send messages at any time. This would not be realistic. Also, we cannot assume that public messages are sent for ever to cover the whole plane which would also not be realistic. Indeed, in case of public messages sent without stopping, the number of messages at any tile at each time would increase to infinity at an exponential rate with time.

We limit the scope of a public message by defining a **radius** of its propagation. If a public message is sent from A , it will reach any tile whose distance from A is at most the radius. The distance between two tiles A and B is the length of a shortest path between A and B . How to implement the limitation? One solution is that the message carries the delay which would be decremented by 1 each time it reaches a new tile, and the message would destroy itself when the delay would be 0. With this solution, we have to transport the delay to all tiles within the radius and at each time, we have to perform this decrementing at each tile.

There is another solution. When A sends a public message with radius r , the message is sent at the speed $\frac{1}{2}$. As we have a global clock, we shall distinguish between odd and even times. A public message travels at odd times and remains on the tile at even times. This time, only the sender has to remember the delay and has to decrease it at each time until the delay is 1. At this moment, an **erasing** message is sent at speed 1. If the delay is r , the erasing message reaches the corresponding public message r times later and it then destroys the message together with itself. See the illustration by Fig. 40.2. With this, we completed the description of the process concerning public messages. Private messages always travel at speed 1.

A last point for the simulation: as a tile does not send a message at any time, we have to decide when it sends a message. For this purpose, we use a Poisson generator, both for the decision of sending a message and, in the case of a public message, for defining the radius of the propagation of the message. For each parameter, we use a different coefficient for the generator. We shall see the values in [Sect.40.4](#)

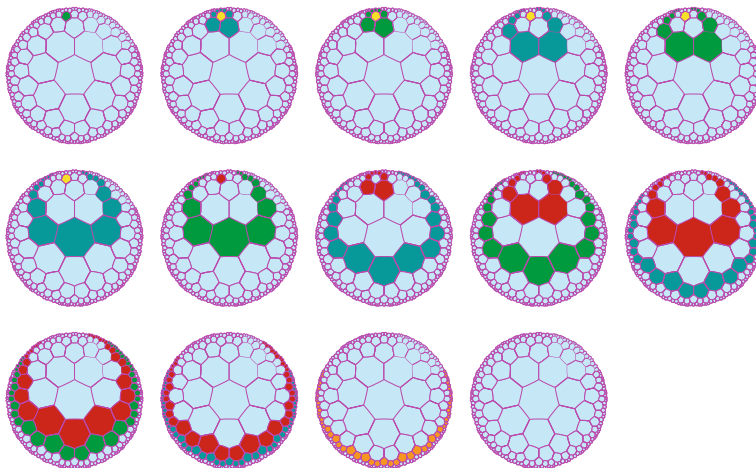


Fig. 40.2 The emission of a public message and of its erasing signal

40.4 The Experiment

The experiment was performed by running the program on a simple laptop. The laptop is a **Lenovo** one, with two Intel processors, both working at 2 GHz, and Linux Mandriva as operating system. The used *ADA*-compiler belongs to the **gnu**-family, version 4.4.1. In the first section, we describe the experiment and we give an account of the results. In the second section, we give an interpretation of the results.

The program was run for six values of the depth of the Fibonacci tree, ranging from 5 to 10. Denote by \mathcal{S} the observation space. The size of \mathcal{S} is defined by the depth of the Fibonacci tree which spans the seven sectors displayed around the central cell. The area of \mathcal{S} in number of tiles is $7f_{2d+2} - 6$ for a depth d , where f_n is the n^{th} term of the Fibonacci sequence defined by $f_0 = f_1 = 0$. Table 40.1 gives the area of \mathcal{S} and the number n_T of messages emitted up to the indicated time T . The line mean of the table indicates the mean value of $\frac{n_T}{T}$ on the interval $[1..T]$.

The parameters of the Poisson law for the different decisions regarding a tile were the following:

- 0.005 for emitting a public message,
- 0.0025 for replying to a public message,
- 0.001 for a private message directly to a single tile,
- 5.00 and 10.00 for the radius of propagation of a public message.

Another important feature regarding private messages is that in the experiment, it was assumed that once a communication has started it goes on endlessly: if

Table 40.1 The number of tiles and the total number of messages

Depth	5	6	7	8	9	10
Tiles	1,625	4,264	11,173	29,261	76,616	200,593
Sent, 5	1,101	3,308	8,636	21,797	49,295*	60,453
Time, 5	168	168	168	168	142	69
Ratio	3.00454	2.61064	2.52396			
Mean, 5	6.81949	19.13115	50.10053	128.31127	342.79960	877.83673
ratio	2.80536	2.61879	2.56108	2.67162	2.56079	
sent, 10	2,173	8,289	13,687*	13,784*	19,167*	30,164*
time, 10	168	168	92	41	30	24
mean, 10	11.21332	40.57043	101.35430	197.08219	405.77815	965.53752
ratio	3.618057	2.50356	1.94449	2.058929	2.37947	

A replies to a message sent by B, either public or private, then B replies to A which again replies to B and this process goes on periodically. Moreover, the reply was always assumed to be immediate.

Table 40.1 indicates the number of messages issued during the whole time of the simulation, lines sent, measured by the number of iterations of the main loop of the computer program, lines time. The number after sent and time indicates which was the mean value for the radius of propagation. Table 40.1 also indicates the ratio between the numbers of messages sent when the radius of propagation is 5 and when the depth of the Fibonacci tree is 5, 6 and 7 as for these depths, we have the same duration 168. Under the line indicating the time for radius 5, we indicate the ratio between the number of messages for consecutive depths as long as the overall duration is the same. The lines mean indicate the mean value of $\frac{L_t}{t}$ on the interval $[1..T]$, T being indicated by the lines time.

From the data of Table 40.1, it seems to us that we can draw the following conclusion:

Assumption 1 For any t, the number of messages issued at the time t in S is proportional to the number of tiles belonging to S.

We have not the room to produce other data collected during the simulation. In particular, the proportion between public message, replies to them and direct private messages to a single tile were analyzed. Also the maximal number of messages passing at a tile was studied. As we cannot display the data, we can just indicate that the experiment shows a slow increasing of the maximal number passing at a tile with time.

At this point, we can also indicate that our choice of a Poisson law to model this message system was driven by the geometry of the space. Indeed, a uniform distribution would give much more weight to distant tiles by the simple fact that their number increases exponentially: the farther they are the more message they would send to the centre. If no limitation would be put, the number of messages received at any point would grow exponentially with time by the just mentioned argument. Now, the Poisson law also gives the possibility to obtain big values with respect to the mean value. Simply these extremal values are very rare, the more rare they are higher.

40.5 Conclusion

It is the place here to discuss how these results can be interpreted in a more qualitative way. The number of iterations suggests that the unit of time is an hour. The tiles can be interpreted either as individuals or as groups of individuals in a given constant area, the one defined by the area of a tile. Remember that in the space we consider, all tiles have the same area. The limitation of the public messages can be interpreted as a natural limit due to the conditions in which the message is sent and also depending on the intentions of the sender.

Two important points should be noted. The first one is the property of the public messages to cover all the tiles of a given area for each tile to receive the message once exactly. The second point is the mechanism to limit the propagation of a public message. This mechanism needs no centralization. It is monitored by the sender and, a priori, each tile can be a sender. To be a sender is defined by a probability which is the same for every tile. The third point is that in some sense, the indicated scenario is a worse case one with respect to the traffic load supported by each tile. Indeed, the fact that once a communication is established between two tiles goes on endlessly contributes to increase the traffic with the time. There is room here to tune the modeling by introducing various ways to delay answers or to limit the number of contacts of a tile with others: here also we could consider that this number is a Poisson random variable whose mean can be fixed uniformly or depending on other criteria which we have not considered here. A last point is the possible improvement of the program in order to obtain more data and to go further in the exploration of the simulation space. Note that the file which records all the communications when the depth is 5 and the radius of propagation is 10 and the number of iterations is 168 has a size of around 164 MB. Note that increasing the depth by 1 multiplies the area by around 2.618. Accordingly, the depth which defines the area of S cannot be extended very much. Already depth 10 with radius 10 requires a machine more powerful than a simple laptop.

We are convinced that there is further work ahead to better analyze the data already obtained, to improve the program in order to go further in the exploration of the simulation space. There is also room to tune the basic parameters in order to get a picture closer to real networks as, for instance, social networks.

References

1. Bonola R, *Non-Euclidean Geometry*. Dover, 0-486-60027-0, New York, p 389
2. Margenstern M (2000) New tools for cellular automata in the hyperbolic plane. *J Univers Comput Sci* 6(12):1226–1252
3. Margenstern M. (2003) Implementing cellular automata on the triangular grids of the hyperbolic plane for new simulation tools. *ASTC'2003 Orlando*, March 29–April 4.
4. Margenstern M (2006) On the communication between cells of a cellular automaton on the penta- and heptagrids of the hyperbolic plane. *J Cell Autom* 1(3):213–232

5. Margenstern M (2007) Cellular automata in hyperbolic spaces vol 1, pp 422, Theory, OCP, Philadelphia
6. Margenstern M (2008) Cellular automata in hyperbolic spaces. Implementation and computations, vol 2, p 360, OCP, Philadelphia
7. Margenstern M (2011) Possible applications of navigation tools in tilings of the hyperbolic plane. Lect Notes Electr Eng 70:217–229

Chapter 41

Monitoring Information System Quality: Between Reality and User Expectation at Bina Nusantara University

Anderes Gui, Cecilia Sabrina, Suryanto, Verina Kristian,
Vina Margareta and Noviyanti Hadiwidjaja

Abstract The objective of this research is to determine the user's satisfaction towards Monitoring Information System Quality and to see user's satisfaction level from two dimension, system quality and information quality. Research methods used are data collection method using questionnaire, data processing method using statistical tools such as pearson product moment correlation coefficient and cronbach's alpha, and data analysis method using gap analysis and Cartesians diagram. The expected result is user's expectation is greater than user's real experience; therefore, system quality and information quality have to be improved. Result shows that user is dissatisfied with applied monitoring information system right now.

Keywords System · Information · Quality · Monitoring

41.1 Background

The development in information system has helped and facilitated many aspects of human's life. The greater the needs in business world, information system application is more needed by the company [1], [2], [3]. Information system is an important system to improve business operation. Company chooses to use an information system because information system can help it to maintain its competitiveness. The significance of information system for the company is reflected

A. Gui (✉) · C. Sabrina · Suryanto · V. Kristian · V. Margareta · N. Hadiwidjaja
Computerized Accounting Department, Computer Science Faculty, Bina Nusantara
University, 9, Jl Kh Syahdan, Jakarta, Indonesia
e-mail: anderesgui@binus.edu; anderesgoei@yahoo.com

by investing 50% of its capital only for information system [4]. Information system is an important investment because it presents company's value chain activities and generates business value throughout time [5], [6]. Right now many organizations depend on information system to maintain and increase their competitive advantage [4], [7], [8].

The purpose of this research is to determine whether the existing monitoring information system quality has given the satisfaction for user or not.

41.2 Methodology

The methods used are place and time of the research, variable constellation of the research, definition of satisfaction operational, population and sample of the research, table of distribution, data collection method, processing data method and data analysis method.

41.3 Theory

The American Society for Quality (ASQ) describes quality as product or services' capability to give value to needs, expectation, and consumer's satisfaction. Defines quality as totality of characteristics from several entities which give all abilities on values of needs or values of satisfaction [9].

System is a group of interconnected or interactive components which form a unity. In information system field, system is more accurate to be defined as a group of interconnected component cooperating to achieve common goals by receiving input and giving output in organized transformation process [10]. A system has three basic function components which is related each other:

Input: involves catching and assembling various elements which enter the system to be processed. For example, raw materials, energy, data, and human's effort have to be secured and regulated for processing.

Processing: involves transformation process which change input into output. For example, manufacturing process, human respiration process, or mathematical calculation.

Output: involves the element movement which is produced by transformation process to its final destinations. For example, final goods, services done by human, and information management which have to be shifted to the users.

System is a group of integrated elements with the intention to achieve one objective. The resources flow from input element, passing through transformation element, to output element. A control mechanism monitors transformation process to ensure that the system meets its objective [11].

By information, data that have been shaped into a form that is meaningful and useful to human being, also states that information is data being processed in order

Table 41.1 Indicator of system quality

Research variable	Dimension	Indicators
Quality of information system	System quality	(1). Usable (2). Secure (3). Efficient (4). Correct (5). Reliable (6). Maintainable (7). Comprehensible (8). Portable (9). Interoperable

Source Mathiassen, object oriented analysis and design, (Marko Publishing ApS,Denmark), p 178

to be meaningful. While, that information is a data being changed to be a meaningful context and to be useful for the final users, also has similar view, Information is data that have been organized and processed to provide meaning to a user [10], [12].

Information system is an organized combination from people, hardware, software, communication network, and data source which collect, change, and distribute information in an organization [10]. Information system is a conceptual system which enable management to control physical system operational of the company and information system can be defined technically as a set of interrelated components that collect (or retrieve), process, store, and distribute information to support decision making and control in an organization [11], [13].

Monitoring is a sequence of activities being done to supervise or observe the process and progress on implementation of a program or activity. Monitoring is an essential aspect of control, where the monitoring term has the meaning that monitoring is an important aspect from a control, quality system is a system characteristic feature of the desired quality of the information system itself and the quality of information desired characteristics of the product information [14], [15]. The quality of an information system measures success technically, communication technical level is interpreted as the accuracy and efficiency of the communication system which produce information [14].

Information quality depends on three things:

Accurate Information has to be precise and free from error because the source of information going to the recipient has high possibility to change the information.

Timeliness of the information being received by the receiver must not be late because information is the principal basis of decision making.

Relevance of information is advantageous in the usage of application; information relevance is different for each person [16].

There are 12 indicators of System Quality, but in this research there will be 9 indicators of System Quality which are appropriate, it consist: (1) usable, (2) secure, (3) efficient, (4) correct, (5) reliable, (6) maintainable, (7) comprehensible, (8) portable, and (9) interoperable. (Table 41.1) [17]

Table 41.2 Indicator of information quality

Research variable	Dimension	Indicators
Quality of information system	Information quality	(1). Accuracy (2). Relevant (3). Timeliness (4). Currency (5). Frequency (6). Time Period (7). Completeness (8). Conciseness (9). Scope (10). Clarity (11). Detail (12). Order (13). Presentation (14). Media (15). Confidence In Source (16). Reliability (17). Appropriate (18). Received by Correct Person

Source Mosweu, evaluation of the effectiveness of information system: IS planning, project governance, IS/IT governance, change management. (Verlag Dr. Muller) p 6

Then, for Information Quality consist 19 indicators but in this research there'll be only 18 appropriate indicators such as (1) accuracy, (2) relevant (3) timeliness (4) currency (5) frequency (6) time period (7) completeness (8) conciseness (9) scope (10) clarity (11) detail (12) order (13) presentation (14) media (15) confidence in source (16) reliability (17) appropriate (18) received by correct person. (Table 41.2) [18]

41.4 Research Result

The research was done by spread the questionnaire as much as 122 questionnaires with the return of the result totally 87 questionnaires. The respondent of the questionnaire are divided by the positions, which are 17.24% (15 persons) as a head of department, 82.76% (72 persons) as Concentration Content Coordinator/Specialist Content Coordinator/Subject Content Specialist. For divided by the faculty, which are 11.49% (10 persons) from Language and Cultural Faculty, 13.79% (12 persons) from economic and business faculty, 27.59% (24 persons) from Computer Science Faculty, 18.39% (12 persons) from Communication and Multimedia Faculty, 8.05% (7 persons) from psychology and technology faculty. In other way, this questioner also divided by training, which are 88.51% (77 persons) ever done the training and 11.49% (10 persons) didn't do the training. For divided by the frequency of using the monitoring system, which are 67.82% (59

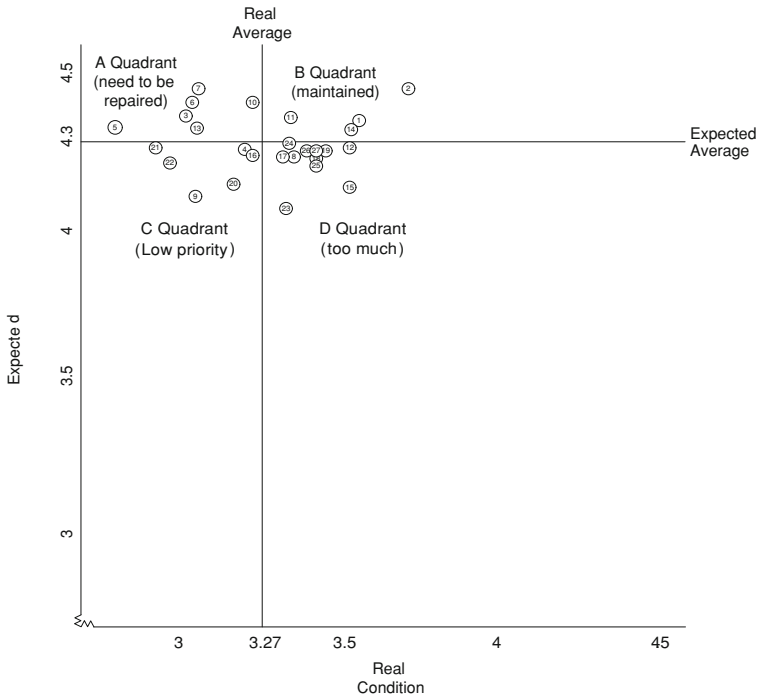


Fig. 41.1 Diagram cartesian between reality and user expectation

persons) used 1–5 times per week, 16.09% (14 persons) used 6-10 times per week, and 16.09% (14 persons) used more than 10 times per week.

From the questionnaire result, it can give the recommendation from the monitoring information system, such as: interface from the system are user friendly, then the system that have been made can be accessed everywhere because as known so far, the system just can be accessed in the campus area, beside that the monitoring information system for the lectures and need to be combined with the university’s website and to make the monitoring process become more easier, it’s better id there’s an option to choose in given the layout of the class that haven’t monitored and gave the sign to the class that haven’t or already been monitored. Also for the update at once meeting, it have to stay In that meeting, no excuse to start from the beginning and when the process monitoring happened, it have to cultivated no need to submit one by one but just combined it and submit all of them. After the premonitory send the message to the lecture, so the lecture can reply or give the comment in the university’s website that will also get into the monitoring system too. Beside that, it’s necessary to display the picture of the lecture and prepare the help menu in the monitoring information system.

Beside the questionnaire, analysis was done by the Cartesians diagram (Fig. 41.1) that divided into four quadrant, they are A, B, C, and D. In A quadrant, it shown a factor that affect the user satisfaction that become the main priority that

Table 41.3 Result based on gap analysis

Point	Reality	Expected	Gap per point	Description
5	2.80	4.33	-1.53	A Quadrant
7	3.07	4.47	-1.40	A Quadrant
6	3.05	4.42	-1.37	A Quadrant
21	2.93	4.29	-1.36	C Quadrant
3	3.04	4.39	-1.35	A Quadrant
13	3.07	4.35	-1.28	A Quadrant
22	2.99	4.24	-1.25	C Quadrant
10	3.24	4.42	-1.18	A Quadrant
4	3.21	4.27	-1.06	C Quadrant
9	3.07	4.12	-1.05	C Quadrant
16	3.23	4.26	-1.03	C Quadrant
11	3.37	4.39	-1.02	B Quadrant
20	3.19	4.17	-0.98	C Quadrant
24	3.35	4.30	-0.95	B Quadrant
17	3.34	4.26	-0.92	D Quadrant
8	3.37	4.26	-0.89	D Quadrant
2	3.62	4.49	-0.87	B Quadrant
26	3.41	4.28	-0.87	D Quadrant
12	3.44	4.29	-0.85	D Quadrant
27	3.43	4.28	-0.85	D Quadrant
18	3.43	4.25	-0.82	D Quadrant
19	3.46	4.28	-0.82	D Quadrant
1	3.57	4.38	-0.81	B Quadrant
25	3.42	4.22	-0.80	D Quadrant
14	3.53	4.32	-0.79	B Quadrant
23	3.34	4.09	-0.75	D Quadrant
15	3.43	4.16	-0.73	D Quadrant
Average	3.27	4.30		

need to be repaired, the factor in this quadrant are, point 3, 5, 6, 7, 10, and 13. In B quadrant show the factor that affect to the user satisfaction that have to maintained or become the second priority because the fact of the user felt was meet the expectation of the user, the factor in this quadrant are 1, 2, 11, 14, and 24. There's also in the C quadrant that show the fact of user in the variable are in the lower place and doesn't need the improvement so far, the factor in this quadrant are 4, 9, 16, 20, 21, and 22. The last but not least, in the D quadrant that show the factor that affected the satisfaction in this quadrant and it assessed that it was excessive in the reality because the user necessary in this variable was low, the factor in this quadrant are point 8, 12, 15, 17, 18, 19, 23, 25, 26, and 27.

There's also the research using Gap analysis based on the questioners and it divided as a point in the quadrant, as a dimension, and the total gap that done to know the gap between mean form the fact and the hope. Table 41.3 show all negative gap, it means average user didn't satisfy for every factor with the quality of monitoring information system.

For the research with the Gap per dimension, it divided into two such as the quality of system and quality of information. For the quality of system, the gap is -1.15 which in this section defined that the quality of the system haven't in the good condition so the user didn't satisfy with the quality of monitoring system. Then for the quality of information, the gap is 0.96 which it means that user didn't satisfy with the quality of monitoring information that exist in the university. Beside the analysis gap from the point and dimension side, the analysis also done by total gap of the all side, where the mean of the fact and the expected of the quality of information system gap is -1.02 that shown the user didn't satisfy of the monitoring activity of the information system.

41.5 Conclusion

From this research, we can make a conclusion that the users didn't satisfy with the quality of monitoring information system in the lectures. Based on analysis quadrant in the Cartesian diagram, there are 6 indicators that become a main priority (A quadrant), they are: Efficient, Reliable, Maintainable, Comprehensible, Accuracy, Currency. That indicators considered as the main priority because haven't meet the user expectation and need to be repair. For the 5 indicator that become a second priority (B quadrant) are Usable, Secure, Accuracy, Frequency, Confidence in Source. This indicator become a second priority because the performance of this section already get the expectation of user and it need to be maintained.

References

1. Gover V, Straub D, dan Galluch P (2009) Turning the corner: the influence of positive thinking on the information systems field. *Manag Inf Syst Q* 33(1):3–8
2. Al-Zeer NM, dan Areiqat AY (2010) The role of information systems in scientific research. *Interdiscip J Contemp Res Bus* 2(3):329–330
3. Hemmatfar M, Salehi M, dan Bayat M (2010) Competitive advantages and strategic information systems. *Int J Bus Manag* 50(7):158–159
4. Lee CY (2003) Total manufacturing information system: a conceptual model of a strategic tool for competitive advantage. *Integr Manuf Syst* 14(2):114–122
5. Menon NM, Ulku Y, dan Cezar A (2009) Differential effects of the two types of information system : a hospital-based study. *Manag Inf Syst J* 26(1):297–316
6. Rao B, dan Angelov B (2009) Pervasive systems value chain—a services perspective. *Int J Innov Technol Manag* 6(1):17–40
7. Johnston HR, dan Vitale MA (2008) Creating competitive advantage with interorganizational information systems. *Manag Inf Syst Q* 12(2):153–165
8. Fadel JK, dan Brown SA (2010) Information systems appraisal and coping: the role of user perceptions. *Commun Assoc Inf Syst* 26(6):527–529
9. Hidayat A (2007) Six sigma strategies. Elex Media Komputindo, Ltd., Jakarta
10. O'Brien JA (2005) Introduction to information system. (12th ed). Mc Graw-Hill/Irwin, New York

11. McLeod R Jr, dan Schell GP (2007) Management information system (10th ed). Pearson Prentice Hall, New Jersey
12. Romney M, dan Steinbart P (2006) Accounting information system. Pearson Education Inc, New Jersey
13. Laudon KC, dan Laudon JP (2010) Management information system: managing the digital firm. Prentice Hall, New York
14. DeLone W, McLean ER (2003) The DeLone and McLean model of information system success: a ten year update. J MIS 19(4):9–30
15. Bateman TS, dan Snell SA (2004) Management: the new competitive landscape (6th ed). McGraw Hill, New-York
16. Jogiyanto HM (2006) Analysis and design information system. Andi Offsets, Yogyakarta
17. Mathiassen L, Madsen AM, Nielsen PA, dan Stage J (2000) Object oriented analysis and design (1st ed). Marko Publishing ApS, Denmark
18. Mosweu TI (2009) An evaluation of the effectiveness of information system : IS planning, project governance, IS/IT governance change management. Verlag Dr. Muller

Chapter 42

Ordinal-Set Pair Analysis Prediction Model and Application in Liao River Basin

Jianxin Xu, Yindi Liu, Zezhong Zhang and Yanbin Li

Abstract Based on the combination of the ordinal theory and the set pair analysis theory, a novel method called ordinal set pair analysis model (O-SPAM) is proposed to simplify the process of runoff prediction and improve the prediction precision. The annual runoff data of fifty-two years from 1956 to 2007 of a certain reservoir in Liao River basin are used to prove the prediction effect of O-SPAM. The results indicate that seven relative errors which account for 70% of the total prediction number are less than 5%; one which accounts for 10% of the total number is between 5 and 10%; while the others which account for 20% of the total number are between 10 and 15%. All the relative errors are less than 20%, which meets the required precision of the annual runoff prediction in “Standard for hydrological information and hydrological forecasting”, indicating that the model can predict the runoff accurately. The study demonstrates that O-SPAM is of clear principle, simple structure, easy calculation, and good prediction precision, and the model is applied to the annual runoff prediction.

Keywords Annual runoff prediction · Set pair analysis (SPA) · Ordinal

J. Xu · Y. Liu (✉) · Z. Zhang · Y. Li
North China University of Water Conservancy and Electric Power,
Zhengzhou 450011, China
e-mail: ranma629@126.com29132973@qq.com

Z. Zhang
e-mail: zhangzezhong78@126.com

42.1 Introduction

Reservoir runoff is the scientific basis of the water conservancy measures, such as flood control and drought relief, developing and utilizing water resources comprehensively, compiling industrial and agricultural water use plans and so on. Reservoir runoff prediction plays a vital role in reservoir operation and regulation. Because of the complexity and variability of the hydrometeorology conditions and human activities, the runoff process presents uncertainty and fuzziness, which leads the mid-long term runoff prediction to become a difficulty in hydrology [1]. There have been many researches about runoff prediction, such as Support Vector Machine (SVM), Artificial Neural Network (ANN), and Singular Spectrum Analysis (SSA) and so on [2]. Although these methods have their own advantages, complex solution procedures as well as specific functions are necessary. In view of this, we introduced the ordinal theory into the Set Pair Analysis (SPA) theory, and improved the existing SPA prediction model [3]. Based on the ordinal, this chapter presents a new SPA runoff prediction model (Ordinal-Set Pair Analysis model O-SPAM), and applies it to predict the annual runoff of a certain reservoir in Liao River basin.

42.2 Set Pair Analysis Principle

Set pair analysis theory (SPA) is a theoretical method which uses connection degree to intensively deal with the fuzziness, randomness, and intermediary of the uncertain system. The core of this theory is to treat certainty and uncertainty as a system. It synthetically depicts connections between things from identity discrepancy contrary (IDC) angles [4]. The set pair is composed of two relative sets. According to specific problems, such connection can be quantitatively characterized by the following connection degree:

$$u = \frac{S}{N} + \frac{F}{N}i + \frac{P}{N}j \quad (42.1)$$

where N is the total number of the set characteristics; S is the number of the identities; F is the number of the discrepancies; P is the number of the contrarieties; i is the uncertain coefficient of the discrepancies, $i \in [-1,1]$, its value depends on conditions; j is the coefficient of the contraries, j is a constant -1 ; $a = S/N$, $b = F/N$, $c = P/N$, and they meet the normalization restriction $a + b + c = 1$, $a, b, c \geq 0$, a, b, c is respectively called identical degree, discrepancy degree, and contrary degree, then formula (42.1) can be written as:

$$u = a + bi + cj \quad (42.2)$$

Table 42.1 Annual runoff time series sets

Historical set	Elements of set				Subsequent value
A_1	x_1	x_2	...	x_m	x_{m+1}
A_2	x_2	x_3	...	x_{m+1}	x_{m+2}
...
A_j	x_j	x_{j+1}	...	x_{m+j-1}	x_{m+j}
...
A_{n-m}	x_{n-m}	x_{n-m+1}	...	x_{n-1}	x_n
B	x_{n-m+1}	x_{n-m+2}	...	x_n	x_{n+1}

Set is the basis of SPA, and connection degree is the core; a, b, c in formula (42.2) respectively denotes the degree of identity, discrepancy and contrary of two sets, they intensively reflects the uncertain relationship between two sets.

42.3 Annual Runoff Prediction Based on the Ordinal Set Pair Analysis

42.3.1 Annual Runoff Prediction Model Based on the Ordinal Set Pair Analysis

The annual runoff time series is given $\{x_1, x_2, x_i, \dots, x_n\}$, and x_i is relative to its adjacent historical value $x_{i-m}, \dots, x_{i-2}, x_{i-1}$ (the number is m). A_j which is generated by the sliding of the runoff time series is named historical set, and its capacity is m . The annual runoff time series $\{x_1, x_2, \dots, x_i, \dots, x_n\}$ generates a total of $n-m$ historical sets; x_i is the subsequent value [5] of each historical set $A_j = \{x_{i-m}, x_{i-m+1}, \dots, x_{i-1}\}$ ($i = m + 1, m + 2, \dots, n; j = 1, 2, \dots, n-m$). If given the current set $B = \{x_{n-m+1}, x_{n-m+2}, \dots, x_n\}$, to predict its subsequent value x_{n+1} , the following steps can be carried out: In the historical sets A_j , find a set A_k or several sets $A_{k1}, A_{k2}, \dots, A_{kl}$ which is similar to set B ; then the subsequent value of set A_k or the weighted average of subsequent values of sets $A_{k1}, A_{k2}, \dots, A_{kl}$ could be considered as the predicted value of x_{n+1} . The annual runoff time series sets are shown in Table 42.1.

In this study, the ordinal analysis was combined with SPA to determine the similarity of sets. The ordinal shows the relative magnitude position of the elements in the sets. For a basin, the influencing factors of the annual runoff are numerous and complicated, yet time variables can be regarded as the comprehensive effect of the entire extrinsic factors [6]. The ordinal of the runoff time series partly reflects its change process. So the similarity of the runoff time series can be compared according to ordinals. Firstly, respectively sort and number the elements of set A_j ($j = 1, 2, \dots, n-m$) and B to obtain the new set A'_j ($j = 1, 2, \dots, n-m$) and B' ; then make set B' and set A'_j compose a total of $n-m$ set pairs

(B', A'_j) ; finally, analyze and calculate the identities, discrepancies and contrarities of these set pairs. That is to calculate the connection degrees u_{B', A'_j} of these set pairs. u_{B', A'_j} is called prediction proximity index in this paper. The larger the index, the higher the similarity of the two sets; then the subsequent value of the historical set is closer to the predicted value of the current set. So the subsequent value could be regarded as the desired predicted value; in case of equal index values, the weighted average of the corresponding subsequent values could be adopted as the desired predicted value. This method is the ordinal set pair analysis (O-SPA) prediction method.

42.3.2 Steps of Annual Runoff Prediction Based on the Ordinal Set Pair Analysis

- (1) Given the annual runoff time series $\{x_1, x_2, \dots, x_i, \dots, x_n\}$, construct the historical sets A_j ($j = 1, 2, \dots, n-m$) and the current set B in Table 42.1, and list the subsequent values x_t ($t = m+1, m+2, \dots, n+1$). Then analyze the periodicity of the annual runoff time series resorting to the spectral analysis method, and adopt the marked period length as the value of m .
- (2) Sort and number the elements of sets A_j and B in descending order, to get the new sets A'_j and B' (in case of equal elements, number them according to the time sequence).
- (3) The new sets A'_j and B' constitute a total of $n-m$ set pairs (B', A'_j) ; then the elements of set B' divided by the corresponding elements of set A'_j , and f_{ji} ($j = 1, 2, \dots, n-m; i = 1, 2, \dots, m$) denotes the quotients. Quantitative criteria are given here: When $f_{ji} = 1$, it denotes 'identity'; when $0 \leq f_{ji} < 1$, it denotes 'discrepancy'; when $f_{ji} > 1$, it denotes 'contrary'. Calculate the numbers of identities, discrepancies and contrarities of each set pair, and write out the expression of the connection degree u_{B', A'_j} ; then evaluate i, j , to figure out the prediction proximity index.
- (4) Select the set A'_k whose prediction proximity index value is the maximum as the similar set of B' , accordingly set A_k is the similar set of B , and there may be one or more similar sets. If the subsequent value of the similar set A_t ($t = k_1, k_2, \dots, k_l$) is x_{m+t} , then the predicted value x_{n+1}' of x_{n+1} could be obtained by the following formula (42.3) and (42.4):

$$x'_{n+1} = \frac{1}{l} \sum_{t=k_1}^{k_l} w_t x_{m+t} \tag{42.3}$$

$$w_t = \frac{1}{m} \sum_{j=1}^m x_j \bigg/ \left(\frac{1}{m} \sum_{i=1}^m x_{ti} \right) \tag{42.4}$$

where w_t denotes the weight; x_{it} and x_j is respectively the element of set A_t ($t = k_1, k_2, \dots, k_l$) and set B ; l is the number of the similar sets of B . If there is only one similar set A_k , $x_{m+t} = x_{m+k}$, $l = 1$, $w_t = w_k$, $x_{n+1}' = w_k \times x_{m+k}$.

42.4 Example

We collected and collated the fifty-two-year annual runoff data from 1956 to 2007 of a certain reservoir in Liao River basin, then applied the ordinal set pair analysis (O-SPA) to predict the ten-year annual runoff from 1998 to 2007. In order to ensure the prediction accuracy and make full use of information, we predicted year by year: For example, to predict the runoff of 1998, we used the annual runoff data from 1956 to 1997; and to predict the runoff of 2006, we used the annual runoff data from 1956 to 2005. Resorting to the spectral analysis method, we analyzed the periodicity of the annual runoff time series, and found that the marked period length was five to seven years, so we adopted six as the value of m .

Take the annual runoff prediction of 1998 as an example.

- (1) We made the forty-two-year annual runoff data from 1956 to 1997 slide to generate a total of thirty-six historical sets (like Table 42.1), and each set whose capacity was six was respectively denoted by A_1, A_2, \dots, A_{36} . And the current set $B = \{x_{37}, x_{38}, x_{39}, x_{40}, x_{41}, x_{42}\} = \{46.70, 51.65, 60.71, 81.03, 67.95, 44.55\}$.
- (2) We respectively sorted and numbered the elements of sets A_j ($j = 1, 2, \dots, 36$) and B , to change them into new sets A_j' and B' . Then we constituted a total of thirty-six set pairs using set B' and A_j' ($j = 1, 2, \dots, 36$). And we calculated the connection degree of the set pairs. Here take set A_{13} and B as an example: $A_{13} = \{53.83, 62.09, 62.75, 71.27, 52.97, 70.43\}$, $B = \{46.70, 51.65, 60.71, 81.03, 67.95, 44.55\}$, then set $A_{13}' = \{5, 4, 3, 1, 6, 2\}$, set $B' = \{5, 4, 3, 1, 2, 6\}$, the two sets were constituted the set pair (B', A_{13}') ; next the elements of set B' divided by the corresponding elements of set A_{13}' , to obtain $f_{13,1} = 1.00$, $f_{13,2} = 1.00$, $f_{13,3} = 1.00$, $f_{13,4} = 1.00$, $f_{13,5} = 0.33$, $f_{13,6} = 3.00$, so there were four 'identities' ($f_{ji} = 1$), one 'discrepancy' ($0 \leq f_{ji} < 1$) and one 'contrariety' ($f_{ji} > 1$). Therefore we obtained that: $a = 4/6 = 0.66$, $b = 1/6 = 0.17$, $c = 1/6 = 0.17$, then the connection degree expression of the set pair (B', A_{13}) was that $u_{B', A_{13}'} = 0.66 + 0.17i + 0.17j$. According to Ref. [7], $i = 0.618$, $j = -1$, we calculated the connection degree $u_{B', A_{13}'} = 0.60$. The results are shown in Tables 42.2 and 42.3.
- (3) In the study, if the connection degree $u_{B', A_j'}$ was over 0.5, then set A_k' was similar to set B' . And the corresponding historical set A_k was similar to the current set B . The results are shown in Table 42.4.

Table 42.2 Set B' and its similar set A_j'

New set	Ordinal	'Identity' number	'Discrepancy' number	'Contrary' number
A_{13}'	5 4 3 1 6 2 4		1	1
B'	5 4 3 1 2 6			

Table 42.3 Connection degree $u_{B'}', A_j'$ calculation of set pair (B', A_j')

New set	f_{ji}	Identical degree	Discrepancy degree	Contrary degree	$u_{B'}', A_j'$
A_{13}'	1.0 1.0 1.0 1.0 0.33 3.0	0.66	0.17	0.17	0.60

Table 42.4 Relative calculations of set B and its similar set A_k

Original set	Elements of set ($m^3 \cdot s^{-1}$)							Elements mean ($m^3 s^{-1}$)	Subsequent value ($m^3 s^{-1}$)	Prediction proximity index
A_{13}	53.83	62.09	62.75	71.27	52.97	70.43	62.22	62.92	0.60	
B	46.70	51.65	60.71	81.03	67.95	44.55	58.76	57.77		

Table 42.5 Set B' and its similar sets A_j'

New set	Ordinal	'Identity' number	'Discrepancy' number	'Contrary' number
A_7'	4 5 1 6 2 3 4		1	1
A_{14}'	5 4 1 6 2 3 3		2	1
A_{34}'	4 2 1 6 5 3 3		2	1
B_9'	4 6 1 5 2 3			

Through Table 42.4, it was clear that $l = 1$. We respectively applied formula (42.4) and (42.3) to obtain the weight w_1 and the annual runoff predicted value of 1998: $w_1 = 0.944$; $x_{1998} = 59.43$. And the relative error was $e = - 2.87\%$.

Similarly, to predict the annual runoff of 2005, we made the forty-nine-year annual runoff data from 1956 to 2004 slide to generate a total of forty-three historical sets and a current set (like Table 42.1), they were respectively denoted by A_1, A_2, \dots, A_{43} and B . According to the previous calculation processes, we obtained the similar sets of set B' . All the relative calculations can be seen in Tables 42.5, 42.6 and 42.7.

Through Table 42.7, it was clear that $l = 3$. We respectively applied formula (42.4) and (42.3) to obtain the three weights and the annual runoff predicted value of 2005: $w_1 = 0.834, w_2 = 0.831, w_3 = 0.945$; $x_{2005} = 59.68$. And the relative error was $e = - 2.87\%$.

We applied the ordinal set pair analysis (O-SPA) method to predict the ten-year annual runoff from 1998 to 2007. And the prediction results are shown in Table 42.8.

We analyzed the relative errors of the prediction results based on O-SPAM: seven relative errors which accounted for 70% of the total prediction number were

Table 42.6 Connection degree u_B , A_j' calculation of set pair (B', A_j')

New set	f_{ji}						Identical degree	Discrepancy degree	Contrary degree	$u_{B',A_j'}$
A_7'	1.0	1.2	1.0	0.83	1.0	1.0	0.66	0.17	0.17	0.60
A_{14}'	0.8	1.5	1.0	0.83	1.0	1.0	0.50	0.33	0.17	0.54
A_{34}'	1.0	3.0	1.0	0.83	0.4	1.0	0.50	0.33	0.17	0.54

Table 42.7 Relative calculations of set B and its similar sets A_i

Original set	Elements of set ($m^3 \cdot s^{-1}$)						Elements mean ($m^3 \cdot s^{-1}$)	Subsequent value ($m^3 \cdot s^{-1}$)	Prediction proximity index
A_7	60.71	60.13	86.34	48.68	62.71	62.44	63.50	53.83	0.60
A_{14}	62.09	62.75	71.27	52.97	70.43	62.92	63.74	69.21	0.54
A_{34}	51.65	62.68	62.92	46.70	51.65	60.71	56.05	81.03	0.54
B	50.82	44.61	63.09	48.65	59.05	51.65	52.98	60.40	

Table 42.8 Prediction results based on ordinal set pair analysis

Year	Measured value ($m^3 \cdot s^{-1}$)	Predicted value ($m^3 \cdot s^{-1}$)	Relative error e (%)
1998	57.77	59.43	-2.87
1999	50.82	45.41	10.64
2000	44.61	45.49	-1.97
2001	63.09	64.43	-2.13
2002	48.65	49.44	-1.62
2003	59.05	59.27	-0.38
2004	51.65	55.12	-6.71
2005	60.40	59.68	1.19
2006	57.49	50.15	12.76
2007	50.89	53.08	-4.30

less than 5%; one which accounted for 10% of the total number was between 5% and 10%; while two relative errors which accounted for 20% of the total number were between 10% and 15% (Table 42.8). According to the required precision of the mid-long-term runoff prediction in “Standard for hydrological information and hydrological forecasting” (SL250-2000) [8], prediction errors within 20% of perennial variation are qualified, and then the qualified rate of the relative errors in the study was 100%. It demonstrates that O-SPAM has good prediction precision; and it is feasible and effective for runoff prediction.

42.5 Conclusions

In this study, the ordinal analysis was introduced into the set pair analysis (SPA) theory, and a new annual runoff prediction model called ordinal set pair analysis model (O-SPAM) was proposed. We adopted the annual runoff time series from

1956 to 2007 of a certain reservoir in Liao River basin as the study object, and applied O-SPAM to predict the annual runoff from 1998 to 2007. Having analyzed the relative errors, we obtained the following conclusions:

- (1) The proposed annual runoff prediction model based on O-SPA can predict annual runoff well in the study. And differing from the existing runoff prediction models, it obviates the need for specific functions and complex solution procedures.
- (2) The results demonstrate that the annual runoff prediction based on O-SPA had high precision, and the prediction results were satisfying. It indicates that the model is applicable to annual runoff prediction, and it provides a new approach for runoff prediction.
- (3) The data on which O-SPA depended was nothing but the historical runoff data, and the abnormalities in runoff resulting from the climate jump were not considered. Therefore, the study remains to be improved, to apply to the runoff prediction under abnormal weather conditions.

References

1. Xiaoyan Q, Jun S, Peng Y et al (2010) Mid-long term runoff forecasting base on EMD and LS-SVM [J]. *Water Resour Power* 28(4):11–13 (in Chinese)
2. DuKui R, Min L, Jiancang X (2010) Application of comprehensive analysis method in the study of models predictive results [J]. *J Xi'an Univ Technol* 26(2):161–164 (in Chinese)
3. Yuan O, Qiong Z, Wensheng W et al (2009) Annual runoff forecasting model based on rank set pair analysis [J]. *Yangtze River* 40(3):63–65 (in Chinese)
4. Wanjun W (2008) Exploration of uncertain set pair analysis [J]. *J Hebei Polytech Univ (Nat Sci Ed)* 30(3):67–69 (in Chinese)
5. Wensheng W, Honglian X, Jing D (2001) Application of nearest neighbor bootstrap regressive model in predication of hydrology and water resources [J]. *Water Resour Power* 19(2):9–14 (in Chinese)
6. Yawei L, Weihua Z, Dongshan W et al (2010) Application of singular spectrum analysis to annual runoff time series in river basin [J]. *Water Resour Power* 28(10):19–22 (in Chinese)
7. Jinzhong S, Shiquan C (2009) A kind of set pairs prediction method based on fuzzy set-valued statistics [J]. *Fuzzy Syst Math* 23(3):56–60 (in Chinese)
8. Water conservancy information center of the ministry of water resources, Hydrology bureau of Yangtze River water conservancy committee of the ministry of water resources, Hydrology bureau of Yellow River conservancy committee of the ministry of water resources, et al (2001) Standard for hydrological information and hydrological forecasting (SL250-2000) [S]. China Waterpower Press, Beijing (in Chinese)

Chapter 43

Integrated RMS Layout and Flow Path Design: Modelling and a Heuristic method

Xianping Guan, Baijing Qiu and Hongbing Yang

Abstract This paper proposed a new model for the integrated layout and flow path design problem of reconfigurable manufacturing system (RMS) using AGV and developed a heuristic method for it. In the proposed model, the layout and the flow path were considered at the same time. And the objective was to minimize the total material handling cost. A heuristic method, named EMLG, was developed to solve the proposed model. Variable neighbourhood search was adopted in the local search procedure of EMLG. The forces of particles were calculated separately and the particles were moved separately. Several computation cases were carried out. The computation results show that the solution quality of the integrated model is much better than that of the step by step one and the proposed method is able to complete computation in limited time and gets high quality solutions. This indicates that the proposed method is effective.

Keywords Electromagnetism-like mechanism • Layout design • Flow path design

X. Guan (✉) · B. Qiu
Key Laboratory of Modern Agricultural Equipment and Technology, Ministry of Education, Jiangsu University, Zhenjiang, China
e-mail: xpguan@ujs.edu.cn

B. Qiu
e-mail: qbj@ujs.edu.cn

H. Yang
School of Mechanical and Electric Engineering, Soochow University,
Suzhou, China
e-mail: tonyyhb@gmail.com

43.1 Introduction

The layout design and the flow path design are key issues for AGV based manufacturing system design and control [1, 2]. This work deals with the integrated layout and flow path design problem for reconfigurable manufacturing system (RMS) using AGV. The layout design problem influences the total material handling cost greatly and it has been studied widely [3]. The flow path design is also an important aspect [4]. These two issues correlate each other and influence the total material handling cost at the same time. Seo et al. [5] considered the integrated design problem, and an iterated tabu search method was employed to solve it. But in previous works, the empty travels of AGVs are often neglected and the reconfiguration cost was seldom taken into account. And the distance between the centres of workstations is often taken as the material distance, which is not the real distance that AGV travels. As the layout and flow path design problem is very complex, optimal methods cannot solve large scale problems. Hence heuristic methods are desirable. Electromagnetism-like mechanism (EM) is a new heuristic method for global optimization [6, 7]. The original EM is designed for continuous optimization. For discrete optimization, some revision is required. Guan et al. [4] proposed a revised version of EM for the flow path design for AGVS. In this chapter, we aim at utilizing the EM for solving the integrated layout and flow path design problem for RMS.

43.2 Integrated Design Modelling

The layout design concerns determining the location of workstations in the shop floor and the flow path design concerns determining the path directions. Suppose the workstation set is W and the site set is S . H_{ws} equals 1 if workstation w is allocated to site s otherwise 0. The loaded flow from workstation w to workstation u is f_{wu} . The shortest path length from P_w to D_u is $L_{P_w D_u}$. C_R is the unit workstation reconfiguration cost. The object of the integrated layout and flow path design is to minimize the total material handling cost J , which includes the cost caused by loaded and empty travels of AGVs and the workstation reconfiguration cost:

$$\min J = \sum_{w \in W} \sum_{s \in S} \sum_{u \in W} \sum_{t \in S} (H_{ws} H_{ut} f_{wu} L_{P_s D_t} + H_{ws} H_{ut} f_{wu}^E L_{D_w P_t}) + C_R \sum_{w \in W} \Delta_w \quad (43.1)$$

With Δ_w indicates whether workstation w is reconfigured or not:

$$\Delta_w = \frac{1}{2} \sum_s |H_{ws} - H_{ws}^0|, \quad \forall w, \quad (43.2)$$

where H_{ws}^0 is the workstation layout of the previous production period.

This problem subjects to many constraints such as site constraints, path constraints and flow constraints. The site constraints mean that one and only one workstation can be allocated to one site:

$$\sum_w H_{ws} = 1, \quad \forall s, \quad \sum_s H_{ws} = 1, \quad \forall w, \quad (43.3)$$

Path length constraints state that the shortest path length between the load/delivery location of workstation w and the delivery/load location of workstation u is the sum of the selected path segments:

$$L_{P_w D_u} = \sum_i \sum_j d_{ij} X_{ij}^{wu}, \quad L_{D_w P_u} = \sum_i \sum_j d_{ij} Y_{ij}^{wu}, \quad (43.4)$$

Path direction constraints call for the path segment selected by loaded/empty flow must satisfy the path direction constraint:

$$X_{ij}^{wu} \leq Z_{ij}, \quad Y_{ij}^{wu} \leq Z_{ij}, \quad Z_{ij} + Z_{ji} = 1 \quad (43.5)$$

Flow constraints mean that for any workstation, the total leaving loaded/empty flow equals the total arriving empty/loaded flow:

$$\sum_{u \in W} f_{wu} = \sum_{u \in W} f_{uw}^E, \quad \forall w, \quad \sum_{w \in W} f_{wu} = \sum_{w \in W} f_{wu}^E, \quad \forall u \quad (43.6)$$

The model given above is very complex, commonly used optimal methods such as branch and bound are not able to solve large scale problems. So heuristic methods are desirable and we adopt electromagnetism-like mechanism (EM) to solve this model.

43.3 Integrated Heuristic Design Method

43.3.1 Algorithm Description

As the original EM is designed for continuous system optimization, we proposed an improved Electromagnetism-like Mechanism for integrated Layout and Flow path design (EMLG). In EMLG, the particles are encoded with two parts $\varphi = (\pi, \gamma)$, with π for the layout, and γ for the flow path. And the total material handling cost is taken as the function value to compute the charges of particles. Variable Neighborhood Search (VNS) strategy is adopted in the local search procedure [8]. The general scheme of EMLG is shown in Algorithm 1.

Algorithm 1

EMLG($m, MaxIt, LSIt, n_l, p_F, n_t$)
 $(\Phi, \varphi_{best}) = Initialize(m)$

```

it = 0
while it < MaxIt do
    (Φ, φbest) = Local(Φ, φbest, LSIt, nl)
    F = CalcF(Φ, φbest, PF)
    (Φ, φbest) = Move(Φ, φbest, F, nl)
    it = it + 1
end while
    
```

43.3.2 The Major Procedures in EMLG

Initialization Procedure In the procedure $(\Phi, \varphi_{best}) = Initialize(m)$, m feasible particles are generated randomly, and the their function values are calculated. The layout is generated by randomly generating the permutation of number from 1 to N_W . And the flow path is generated by randomly generating zero one encoding with length of N_B . If the generated flow path is infeasible, the encoding is regenerated.

Local Search Procedure Reduced VNS search strategy is adopted in the local search procedure $(\Phi, \varphi_{best}) = Local(\Phi, \varphi_{best}, LSIt, n_l)$. The neighborhood is defined according to Hamming distance. The layout is searched with shifting according sequencing method, and the flow path is searched by change the encoding with k positions.

For an encoding φ , the layout neighborhood is defined as:

$$N_k^\pi(\pi) = \{\pi' | H^\pi(\pi, \pi') = k\} \tag{43.7}$$

and the guidepath one as:

$$N_k^\gamma(\gamma) = \{\gamma' | H^\gamma(\gamma, \gamma') = k\} \tag{43.8}$$

To reduce computation time, only the current best particle is searched. The local search is divided into two layers, with the layout as outside, and the flow path as inside. LSIt is set as the number of local search times and n_l is the number of retrying times for infeasible flow path network.

Total Force Calculation Procedure The charge of a particle is determined by its function value, which is calculated as follows.

$$q^i = \exp\left(-m \frac{J(\varphi_i) - J(\varphi_{best})}{\sum_{k=1}^m (J(\varphi_k) - J(\varphi_{best}))}\right), \quad \forall i. \tag{43.9}$$

The total force of layout and flow path is calculated separately, the total force of the layout is calculated according to following equation:

$$F_i^\pi = \sum_{j \neq i}^m \left\{ \begin{array}{l} (\pi_j - \pi_i) \frac{q^i q^j}{\|\pi_j - \pi_i\|^2} \text{ if } J(\varphi_j) < J(\varphi_i) \\ (\pi_i - \pi_j) \frac{q^i q^j}{\|\pi_j - \pi_i\|^2} \text{ if } J(\varphi_j) \geq J(\varphi_i) \end{array} \right\}, \quad \forall i \quad (43.10)$$

the total force of the flow path is calculated according to following equation:

$$F_i^\gamma = \sum_{j \neq i}^m \left\{ \begin{array}{l} (\gamma_j - \gamma_i) \frac{q^i q^j}{\|\gamma_j - \gamma_i\|^2} \text{ if } J(\varphi_j) < J(\varphi_i) \\ (\gamma_i - \gamma_j) \frac{q^i q^j}{\|\gamma_j - \gamma_i\|^2} \text{ if } J(\varphi_j) \geq J(\varphi_i) \end{array} \right\}, \quad \forall i \quad (43.11)$$

Similar to Birbil et al. [7], we change the force direction of the particle farthest to current particle with a certain probability p_F , so as to enlarge the search space.

Moving Procedure In the moving procedure $(\Phi, \varphi_{best}) = Move(\Phi, \varphi_{best}, F, n_t)$, the layout and the flow path are moved seperately. The layout encoding is moved according to the layout force and the initial encoding. We add the total layout force of a particle to its layout encoding, and sort the resulting encoding. Then a new layout encoding can be got according to the order. The flow path encoding is moved according to the ordering strategy, see [4] for more details. If the new flow path network is infeasible, it is retried at most n_t times.

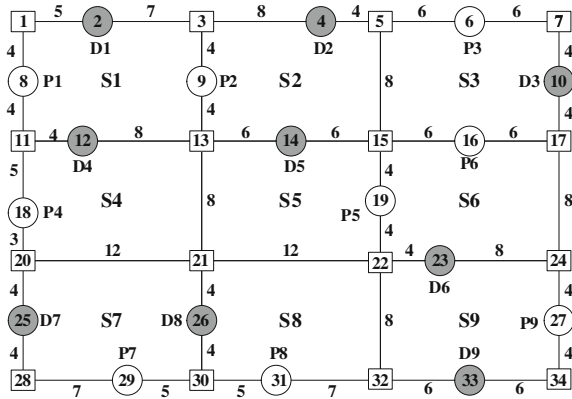
We set the maximum iteration $MaxIt$ as the stop condition.

43.4 Computation Cases

To test the effect of the proposed method, several computation cases are carried out. We compare the integrated design method and the step by step method (GAL/EMG). In GAL/EMG, we utilize the genetic algorithm to get the first k best layout. Then we use the EM method to design the flow path for each resulting layout [4]. When the problem scale is small, we solve the best layout and flow path by enumerating all the possible layout first and using the branch and bound method to get the best flow path for each layout. We call such method as the improved branch and bound method (IBB). All the programs are written in Matlab7.0TM and the running environment is Windows XPTM with processor as IntelTM Core2 2.4 GHz and 2 GB RAM.

At the initial production period, we have to design the layout and flow path with the initial loaded flows (Flow 1). Then at the new production period, we are required to design the system with the new loaded flows (Flow 2) under previous system settings. Then the workstation reconfiguration cost should be taken into account. We consider two different unit reconfiguration costs: zero and 900. We compare the difference of the average material handling cost (AMHC) to the best

Fig. 43.1 The site layout for Case 1



material handling cost (BMHC), and the average computation time (ACT) for each method.

43.4.1 Case 1

This is a case of nine workstations. The site layout is given in Fig. 43.1. The initial loaded flows are given in Table 43.1 and the new loaded flows are given in Table 43.2.

In this case, we first run IBB method once for each setting and then run EMLG method and GAL/EMG method ten times for each setting. The computation results are given in Table 43.3.

From Computation result analysis. In this case, the computation time of heuristic methods (EMLG and GAL/EMG) is much less than that of the optimal method (IBB). The solution quality of GAL/EMG is much worse than that of IBB. And the solution quality of EMLG is near to that of IBB. We can see that the IBB method is not suitable for large scale problems due to large amount of computation time and the solution quality of the step by step method is worse than that of the integrated one.

43.4.2 Case 2 and 3

These are cases with 15 and 25 workstations. As these cases are too complex for IBB method to complete computation in limited time, we only compare EMLG method and GAL/EMG method in this case. We run EMLG method and GAL/EMG method ten times for each setting in these cases. The computation results are given in Table 43.4.

Table 43.1 The initial loaded flows for Case 1

From	To								
	1	2	3	4	5	6	7	8	9
1	0	20	0	30	0	0	50	0	20
2	0	0	10	0	0	5	0	0	5
3	0	0	0	0	0	0	0	10	10
4	0	0	0	0	20	10	0	0	0
5	0	0	0	0	0	0	20	30	0
6	0	0	0	0	0	0	0	30	0
7	0	0	0	0	25	15	0	30	0
8	0	0	0	0	0	0	0	0	0
9	0	0	10	10	5	10	0	0	0

Table 43.2 The new loaded flows for Case 1

From	To								
	1	2	3	4	5	6	7	8	9
1	0	0	40	10	60	0	0	10	0
2	10	0	10	10	0	50	0	0	5
3	20	0	0	20	0	30	0	0	10
4	30	0	0	0	0	30	0	40	0
5	0	20	0	0	0	0	10	0	20
6	50	0	20	0	0	0	0	0	10
7	0	20	0	0	5	0	0	30	10
8	70	0	10	0	30	0	0	0	0
9	0	0	0	10	50	10	0	20	0

Table 43.3 The computation results for each method in Case 1

Flow	URC	IBB method		EMLG method			GAL/EMG method		
		BMHC	CT(s)	AMHC	ACT (s)	DTB(%)	AMHC	ACT (s)	DTB(%)
1	0	15,480	100,569.5	15,836	189.0	2.3	19,960	289.4	28.9
2	0	26,370	120,673.2	26,732	192.6	1.4	30,278	251.6	14.8
2	900	32,945	72,503.5	33,307	178.3	1.1	36,860	282.5	11.9

Computation result analysis. The computation time of GAL/EMG method is more than that of EMLG method and the solution quality is much worse. This shows that the result of step by step method is worse. As the first k layouts are searched, the computation time is large.

From the above computation cases, we can see that the step by step method is worse than the integrated method. The optimal method is not suitable for large scale problems. The solution quality of the proposed EMLG method is better than that of GAL/EMG method.

Table 43.4 The computation results for Case 2 and 3

N _w	Flow	URC	BMHC	EMLG method			GAL/EMG method		
				AMHC	ACT (s)	DTB(%)	AMHC	ACT (s)	DTB(%)
15	1	0	76,900	80,530	593.5	4.7	108,400	1548.7	41.0
	2	0	82,410	84,219	555.6	2.2	95,823	1759.2	16.3
	2	900	93,080	94,637	561.9	1.7	97,103	2114.8	4.3
25	1	0	236,300	248,020	2,284.4	5.0	282,280	3357.2	19.5
	2	0	256,335	264,365	2,712.8	3.1	295,364	3645.7	15.2
	2	900	270,265	272,872	2,802.9	1.0	308,257	3962.3	14.1

43.5 Conclusions

According to the requirement of reconfigurable manufacturing system, an integrated layout and flow path design model is proposed and a fast heuristic method EMLG is developed in this paper. Several computation cases are carried out to verify the effect of the proposed model and method. The computation results show that the solution quality of the integrated method is much better than that of the step by step one. And the proposed method is able to find near optimal solution in limited computation time. This indicates that the proposed method is effective.

As the requirement maybe uncertain, we will consider stochastic loaded flow matrix in future works. There maybe different system requirements, so the multi-objective method is a possible future direction.

Acknowledgments This work was a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions, and was partially supported by the Project of High-tech Key Laboratory of Agricultural Equipment and Intelligentization of Jiangsu Province under grant: BM2009703, the School Foundation of Jiangsu University under grant: 11JDG043, the National Natural Science Foundation of China under grant: 51005160, and the Collegiate Natural Science Foundation of Jiangsu Province under grant: 10KJB410001.

References

1. Le-Anh T, Koster MBMD (2006) A review of design and control of automated guided vehicle systems. *Eur J Oper Res* 171(1):1–23
2. Vis IFA (2006) Survey of research in the design and control of automated guided vehicle systems. *Eur J Oper Res* 170(3):677–709
3. Drira A, Pierreval H, Hajri-Gabouj S (2007) Facility layout problems: a survey. *Ann Rev Control* 31(2):255–267
4. Guan X, Dai X, Li J (2011) Revised electromagnetism-like mechanism for flow path design of unidirectional agv systems. *Int J Prod Res* 49(2):401–429
5. Seo Y, Sheen D, Moon C, Kim T (2006) Integrated design of workcells and unidirectional flowpath layout. *Comput Ind Eng* 51(1):142–153
6. Birbil SI, Fang S-C (2003) An electromagnetism-like mechanism for global optimization. *J Glob Optim* 25(3):263–282

7. Birbil SI, Fang S-C, Sheu R-L (2004) On the convergence of a population-based global optimization algorithm. *J Glob Optim* 30(23):301–318
8. Hansen P, Mladenovic N (2001) Variable neighborhood search: principles and applications. *Eur J Oper Res* 130(3):449–467

Chapter 44

Design and Implementation of Distributed Remote-Reading Water Meter Monitoring System Based on SaaS

Junjie Li, Jianjiang Cui, Lilong Jiang, Zhijie Lu and Zhenyu Tan

Abstract The system has features of loose coupling, service oriented and distributed. The communication subsystem is based on C/S architecture. It copes with highly concurrent access by using asynchronous socket communication. The monitoring subsystem is based on B/S architecture and SaaS (software as a service). It uses WCF (Windows Communication Foundation) web services to accomplish receiving, transmission and storing of data. This system implements functions of remote meter reading, water selling and making statistics of data. The design based on SaaS reduces the difficulty and risk for water enterprises' informatization.

Keywords SaaS · WCF · Distributed · Remote-reading water meter

44.1 Introduction

With the pace of Chinese water enterprises' informatization speeding up, there is daily increasing urgency for digital and networking monitoring system for remote-reading water meters, and there is also higher standard demand for reliability, interoperability and low cost of the system [1]. It is the direction of development for Chinese water enterprises' informatization to gather, store and make statistics

J. Li (✉) · L. Jiang · Z. Lu · Z. Tan
Software College, Northeastern University, Shenyang, China
e-mail: junjieli1213@126.com

J. Cui
College of Information Science and Engineering, Northeastern University,
Shenyang, China

of data in real-time and develop effective, practical and distributed monitoring systems for remote-reading water meters.

At present, the development of remote-reading water meter monitoring system in China is still in the exploratory and experimental stage [2]. On the one hand, due to the various progress of informatization in different regions, the way of meter-reading as well as the environment of hardware and software is different. The interoperability of system, namely cross-regional and cross-platform, must be considered when designing the remote-reading water meter monitoring system. On the other hand, the consumption of water produces a large amount of data. The performance of the single server is highly demanded if traditional centralized storage is adopted. At the same time, it takes them much money to purchase IT products [3], technology and maintenance for water enterprises because of high requirements and risk.

The distributed remote-reading water meter monitoring system based on SaaS is deployed in network. Users can lease and customize it. The communication subsystem can gather and store the data of water consumption in real time. The monitoring subsystem can make statistics of the data. The design and implementation of the remote-reading water meter monitoring system significantly enhances the efficiency and real-time of monitoring and management for water meters. By applying the pattern of software as a service (SaaS) [4], the investment as well as the risk is reduced for water enterprises' informatization. So the system has broad prospects and far-reaching social benefits. This chapter focuses on the overall architecture, key technologies and the implementation methods of the distributed remote-reading water meter monitoring system based on SaaS. Both the idea of SaaS and the technical application of Windows Communication Foundation (WCF) are fully embodied in this chapter [5].

44.2 Overall Architecture of the System

The overall architecture of distributed remote-reading water meter monitoring system based on SaaS is showed in Fig. 44.1. The remote-reading water meter is equipped with GPRS transceiver which can upload data to the communication server at the predefined time. The communication subsystem deployed in the communication server carries out two-way communication with water meters and it exchanges data with water meters by invoking relative services in the services host through WCF proxies [6]. Web application is deployed in the Web server as client end. The application invokes relative services by sending messages to the services host through WCF proxies. Water enterprises use the system by leasing these services.

44.3 SaaS Pattern and WCF Communication Model

SaaS (software as a service) is a completely innovative software application mode arising in the 21st century. Software developers deploy software on the network.

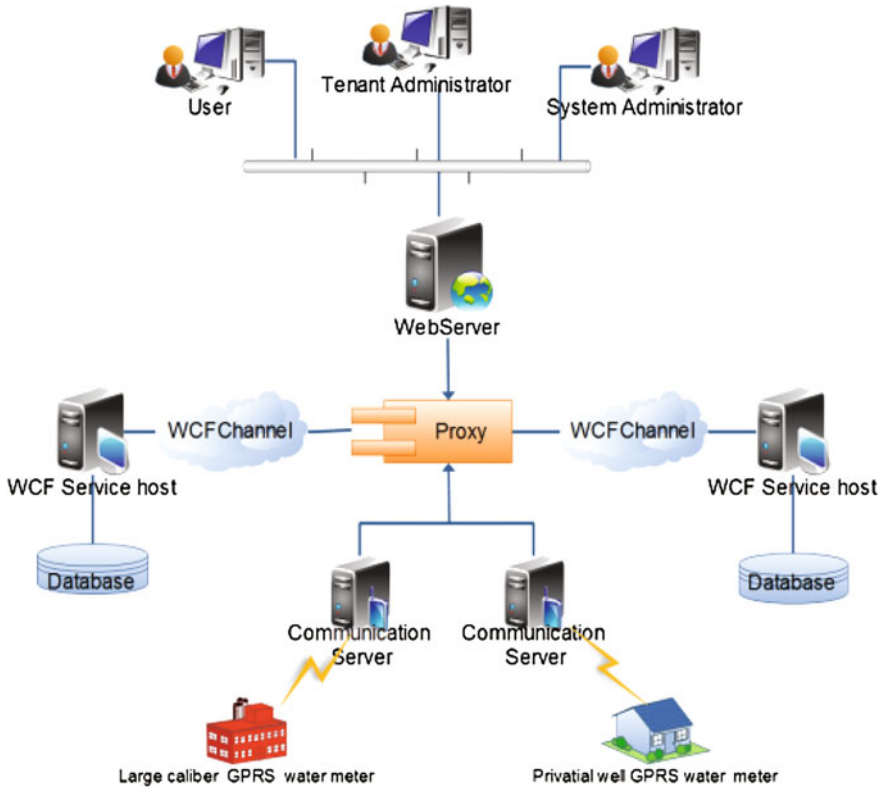


Fig. 44.1 The overall architecture of the system

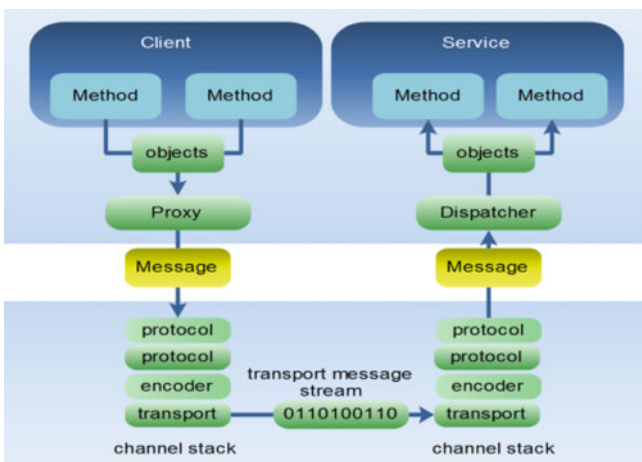


Fig. 44.2 The communication process of WCF

Users use customized software by leasing it. What's more, the software developer is responsible for the upgrade and maintenance of the software.

Windows Communication Foundation (WCF) is Microsoft's unified programming model for building service-oriented applications [7, 8]. It enables developers to build secure, reliable, transacted solutions that integrate across platforms and interoperate with existing investments. In WCF model, contracts define the way of communication for both ends. The security layer which both ends comply with defines the communication security. Contracts are in the form of interfaces and actual service code must be derived from these contracts and implement these interfaces. The communication process of WCF is showed in Fig. 44.2.

44.4 Design and Implementation of Key Functions in the System

44.4.1 Design and Implementation of the Communication Subsystem

In the system, meters and the server communicate with each other by creating socket connections. If multi-threading technology is used, the server have to create multiple threads to handle the connections when a lot of meters access the server concurrently, which will consume a lot of resources and even result in packet loss. Fortunately, this problem can be well solved by using asynchronous socket communication.

The `SocketAsyncEventArgs` class in Microsoft's .NET Framework is used in asynchronous socket communication. This class is designed for the needs of high-performance network server applications, by saving I/O operations context, reducing objects allocation and garbage collection. The communication diagram is showed in Fig. 44.3. First, a `SocketAsyncEventArgs` queue is created to handle the socket connections between water meters and the server. In the communication process, main thread is responsible for listening on the port by means of asynchronous monitoring. When a new socket connection is created between the server and one meter, an object is selected from the `SocketAsyncEventArgs` queue to store the state of this connection. All the parameters of the socket's operations can be set by this object's properties and methods. Then the event handler calls a callback method to receive the data uploaded by the meter. Finally the server sends a message to the meter to achieve setting its parameters including time and spare water.

There may be a large number of meters accessing the communication server concurrently, and the stress can be quite high. To copes with the highly concurrent access, WCF is used in when designing the communication server. Services for data storage and reading are deployed in the services host. After creating a connection with one meter, the communication server receives the data uploaded by

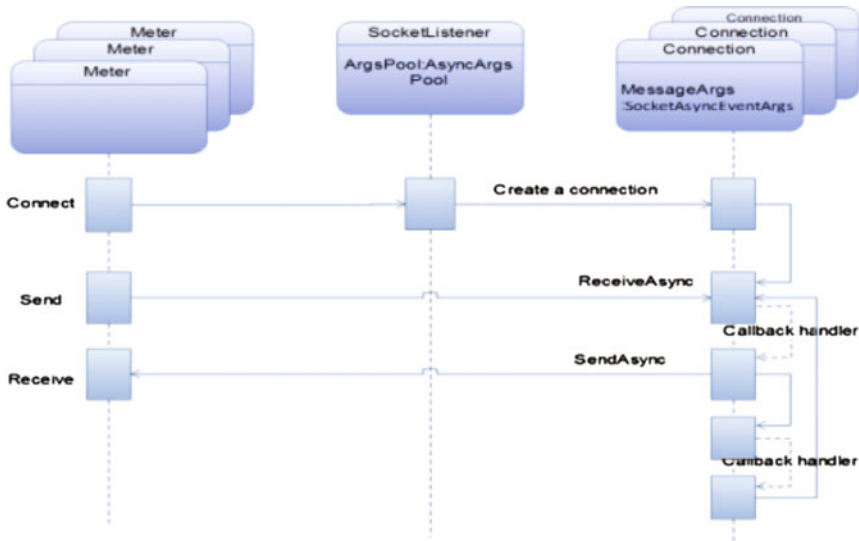


Fig. 44.3 The asynchronous socket communication

the meter, then it invokes the related services in the services host to store the data and get the data it need to send back to the meter. In this way, communication pressure can be eased by increasing the number of communication servers when thousands of meters access concurrently, which improves the reliability, availability, and scalability of the system.

In the system, a SMS notification model is developed for users to keep abreast of their spare water. When the remaining volume is below the alarming volume, the system can send a short message to the related user reminding him to purchase water.

44.4.2 Design and Implementation of Authentication and Authorization

The authentication and authorization of the system is implemented by adopting Lightweight Directory Access Protocol (LDAP) protocol and Role Based Access Control (RBAC). Active Directory Application Mode (ADAM) provides LDAP directory service for applications supporting the directory. In the system, ADAM stores the information of users' authentication and authorization, and the Identity database stores tenants' archives and mapping information of users (mapping of ID and user's name). The information in ADAM and identity database accomplishes the authentication jointly. The authentication and authorization in multi-tenant architecture is showed in Fig. 44.4.

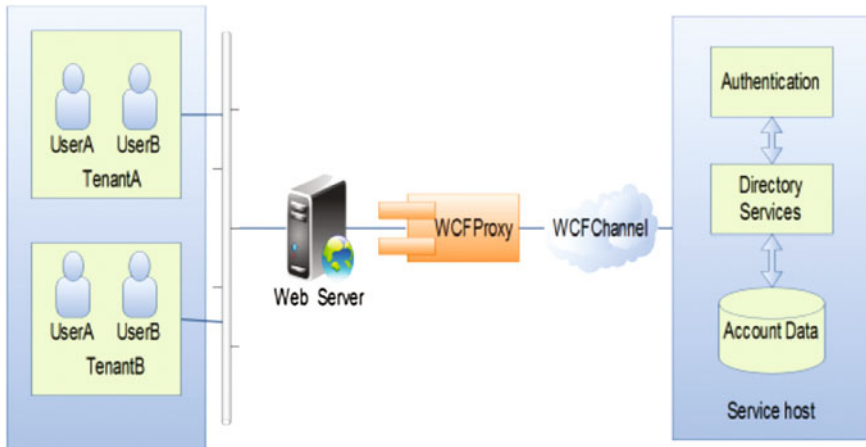


Fig. 44.4 The authentication and authorization in multi-tenant architecture

In the system, class `MembershipProvider` and `RoleProvider` in .NET Framework are implemented. Authentication and authorization in the web application is implemented by using `Membership` in ASP.NET 3.5.

44.4.3 Architecture of Multi-Tenant

The distributed remote-reading water meter monitoring system adopts the architecture of shared database and shared data schema. The information of tenant is identified by increasing the `Tenant` table, `TenantUser` table, and adding `TenantID` field in other tables. The `TenantID` field uses type `GUID` to ensure unique. Tenants' individual needs of data, such as charging different water fees, are met based on the concept of metadata name-value pair method. Web Part interface technology is used to meet Tenants' individual needs of interface.

44.4.4 Design and Implementation of Data Statistics

`Fusionmaps` is introduced as a solution to generating statistical charts. The statistical charts include chart of water consumption density in different regions and chart of business data for different tenants. The web application gets serialized objects named `TenantBusiness` and `WaterSale` through the WCF proxies, and then these objects will be sent to the client's browser along with flash files to generate flash charts.

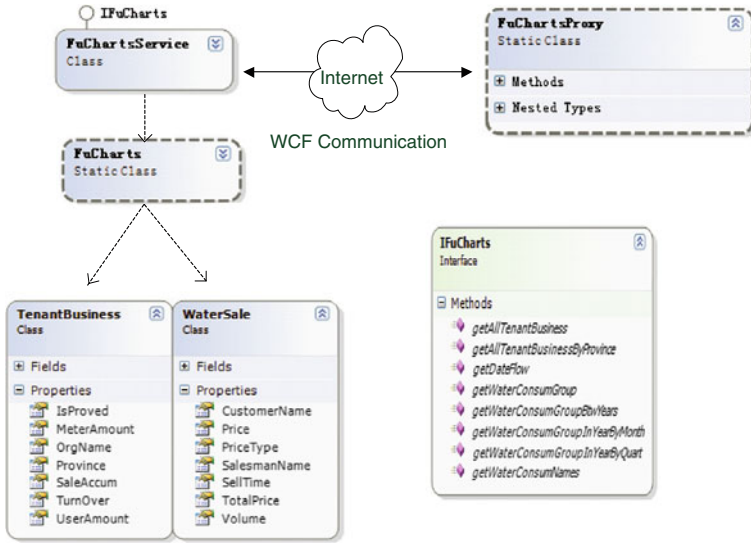


Fig. 44.5 The class relationship in data analysis module

Fig. 44.6 The density of domestic water consumption



The relationship between classes in data analysis module is showed in Fig. 44.5. Class IFuCharts works as message contract and class TenantBusiness and class WaterSale work as data contract. Class TenantBusiness represents tenants’ business data and class WaterSale contains information of water sales. Fig. 44.6 shows the density of domestic water consumption and Fig. 44.7 shows the comparison of different tenants’ business data.

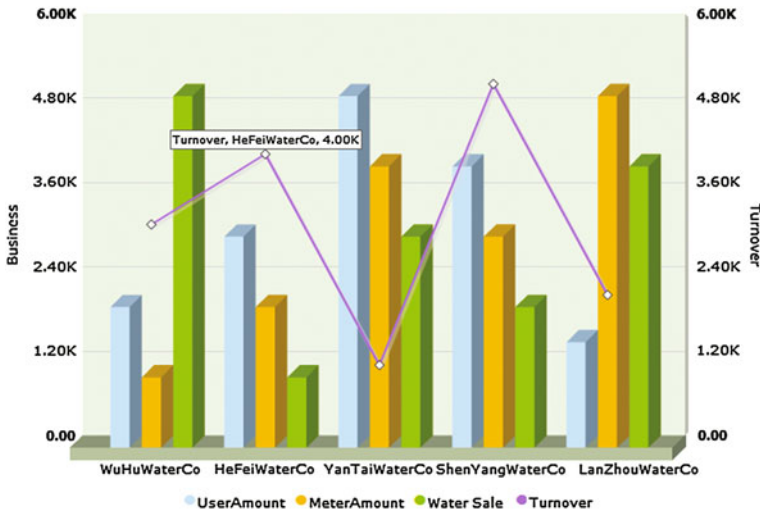


Fig. 44.7 The comparison of different tenants' business data

44.5 Summary

It is a new design approach to introduce the SaaS pattern into the design and implementation of remote-reading water meter monitoring system. It provides an efficient and real-time solution for water enterprises' informatization, which achieves the goal of meter reading, meter monitoring, water selling and data statistics as well as reducing investment under the Internet environment. The communication subsystem based on asynchronous socket communication and WCF model provides a solution coping with highly concurrent access when using wireless devices. The system also has reference value for distributed monitoring of wireless devices and architecture of systems based on SaaS.

References

1. Xiaolan M (2011) *Bulletin of Science and Technology* 27:114–119
2. Suojun L, Wei D (2010) *Automation Panorama* 27:86–90
3. Wei Y (2009) *Revolution in the age of internet: architecture and design of SaaS software*. Electronic Industry Press, China
4. Klein S (2007) *Professional WCF Programming*. Wrox, Birmingham
5. Nagel C, Evjen B, Glynn J, Skinner M, Watson K (2008) *Professional C# 2008*, Pap/Psc edition. Wrox, Birmingham
6. Evjen B, Hanselman S, Rader D (2009) *Professional ASP.NET 3.5 SP1 Edition: In C# and VB*, Har/Cdr edition. Wrox, Birmingham
7. MSDN. <http://msdn.microsoft.com/en-gb/default>
8. Stunning charts for Web and Enterprise. <http://www.fusioncharts.com>

Chapter 45

Enabling the Traceability of Web Information Access in Laboratory Management via E-CARGO Model

Bin Xu

Abstract Internet information security has become increasingly prominent. In this research, E-CARGO collaboration model is suggested in the environment of laboratory in high education. The involved staff and devices has been modeled as roles, the relationship has been modeled as services. Algorithms for harmful web information reporting, gathering and tracing are presented.

Keywords Internet information security · E-CARGO · Information tracing

45.1 Introduction

Though China formally connected to the Internet in 1994, the number of netizens has reached 360 million after a short 15 years of development. At the same time, Internet applications including e-commerce, enterprise information technology and community networks boom rapidly. The Internet has penetrated deeply into all areas of social life, profoundly changes people's lifestyles and modes of production. While the Internet has brought tremendous changes to the social life, gave rise to a lot of convenience, government officials and industry experts believe that, network security has become increasingly prominent [1, 2].

Wei Zi-chuan, vice president of Xinhua, said that in recent years, a handful of people seek profit in a variety of ways to spread of pornography, defamation and disinformation on the Internet.

B. Xu (✉)
College of Computer and Information Engineering, Zhejiang Gongshang University,
Hangzhou 310018, China
e-mail: xubin@zjgsu.edu.cn

The China Internet illegal and unhealthy information report center released a set of startling data: the Center received a total of 294,18 thousand times from the public, account for more than 15,000 phone calls, average 1,009 times daily. According to the report from Internet users, 291 illegal websites clues had been handed over to national law enforcement and administrative authorities after verification, including 164 cases of pornographic and vulgar, 99 cases of sales of prohibited category, and 28 cases of fraud.

Lee Ka-ming, Deputy director of the Center, said the past year, the Center had verified 142,057 web sites, found 603 of them had totally 527,931 illegal and harmful information, and 191 search engine service providers has 914 pornographic keywords.

Young people up counts for 175 million in total 360 million Internet users, while many young people are being harmed by network vulgar, obscene and terrorist violence and other illegal and unhealthy information.

Regarding the impact to high education, Internet brings forth many good ideas and wonderful materials and is very useful in high education, including global knowledge and education resource sharing [8, 10, 11], online training and online talent development [3]. However, bad materials which contain unhealthy content are contaminating the study environment, wasting the valuable time and harming the health. While there is an access policy for such harmful information, traceability is essential for further management and controlling. Therefore, in laboratory management, the access to web information should be kept in repository and be traced to the person who published/accessed the material.

Since internet is an open environment which everyone may publish and share the material freely, there is large effort for the laboratory management to keep the record of information access. This research is to suggest an efficient method based on E-CARGO collaborative model to overcome the problem.

The rest of this paper is organized as following. [Section 45.2](#) introduces E-CARGO model and presents the related definitions. [Section 45.3](#) describes the problem and proposes the solution based on E-CARGO model. The discussion and conclusion are made in [Sect. 45.4](#).

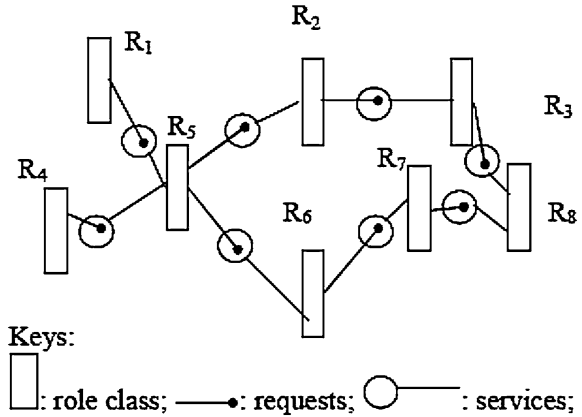
45.2 E-CARGO Model and Related Definitions

45.2.1 E-CARGO Model

Role-based collaboration model E-CARGO (Fig. 45.1) is introduced by Dr. Zhu in [4, 5] to establish the development/business environment as a role net. Each role provides a certain services and applies a certain services in the proposed role net.

Role based collaboration and its kernel mechanism were introduced in [6]. E-CARGO is helpful to build a more efficient collaborative system, e.g. roles can be regarded as agent dynamics in multi-agent systems [7]. Role transfer in

Fig. 45.1 E-CARGO model. Roles provide different services to satisfy the requests from other roles, which clarifies the relationship and enhance the collaboration



emergency management systems [8] is also a regular activity in education management.

45.2.2 Related Definitions

There are kinds of roles in the web information access, including students, teachers, audit, door checker, and web information storage. Teachers guide the laboratory and may also access the web information. The students access the web information so as to finish the laboratory. The door checker is a device to check the identification of the students and teachers before they may access the computers. The web information storage is also a device to record all the accessed web information. The audit will analyze the recorded information and find out the harmful information. There can be several relationships between all these roles. According to the E-CARGO model, all the personnel and devices can be modeled as agents, positions (student, teacher, audit, etc.) can be modeled as roles, and information access, record and analysis can be modeled as services, the web information can be modeled as class/object data. The detail definitions are provided as follows.

Definition 1 *Role* A role is the service template. A role is defined as $r:: = \langle id, ca, S \rangle$, where:

id is the identification of the role *r*;

ca is the category of *role*, there are five different roles considered in this research work, including students, teachers, audit, door checker, and web information storage;

S is a set of services provided by role *r* (Table 45.1).

Table 45.1 Roles and responsibilities

Role	Responsibility
Teacher	Participate the laboratory Provide advice, or technique support in a laboratory Gather harmful information report from student and report to audit
Students	Participate the laboratory Access the web information to finish the excise Report harmful web information
Audit	Analyze the web information access log Gather report from teachers and report harmful web information Trace harmful web information
Door checker	Verify the identification of the students and teachers Gather the time of system logon
Web information storage	Record the accessed web information

Laboratory can be considered as a laboratory project, within which there are teachers to serve as the instructors, who is to provide advice and technique support when the students is in trouble, and gather the harmful record and hand over to the audit; the students who access the web information so as to finish the laboratory and report harmful web information to the teachers; audit who gather the harmful information report and hand over to the security department and trace the harmful information in the storage; door checker verifies the identification of students and teachers and record the time when they logon on the system; and web information storage stores all the accessed web information for the possible traceability.

All these positions can be modeled as roles in the E-CARGO model, and the related responsibilities can be modeled as services.

Definition 2 *Agent*. An agent is a role player (i.e., people). It can refer to a role demonstrated in Fig. 45.1 and is defined as $ag:: = \langle id, name, rid, avm \rangle$ where:

id is the identification of the agent ag ;

$name$ is the name of the agent ag ;

rid is the role which the agent ag belongs to;

avm is a matrix of the available time of agents, it is defined as $\{\langle s, e, location \rangle\}$,

indicating that the agent ag is available at the location between the start time s and the end time e .

Definition 3 *Service*. A service is a responsibility of a role. It is defined as $s:: = \langle id, r, pr \rangle$ where:

id is the identification of the service s ;

r is the role who provide the service s ;

pr refers to the laboratory project.

Definition 4 *Project*. A project is a special service, which refers to the collaboration of roles and information. It is defined as $pr:: = \langle id, R, I, H, s, e \rangle$ where:

id is the identification of the laboratory project pr ;

R is a set of roles in the laboratory project pr ;

I is a set of information accessed in the laboratory project pr ;

H is the harmful report in the laboratory project pr , as matched pair of the time, agent, information and location. It is defined as $H = \{ \langle t, a, i, l \rangle \}$, which expresses that the agent a reports harmful information i at the location l at time t ;

s is the start time of the laboratory project pr ;

e is the end time of the laboratory project pr .

Definition 5 *Information Type*. Type is the category of web information accessed in the laboratory. It is defined as $tp:: = \langle id, name, keys, solution \rangle$ where:

id is the identification of the information type tp ;

$name$ is the category of the information, for instance, name can be *healthy*, *pornographic*, *vulgar*, *sales of prohibited category*, and *fraud*;

$keys$ is the key words of the information type tp ;

$solution$ is the following up actions to the information type tp .

Definition 6 *Information*. Information refers to the web information accessed in the laboratory. It is defined as $inf:: = \langle id, tid, sites, keys, alog \rangle$ where:

id is the identification of the information inf ;

tid is the identification of its category of inf ;

$sites$ is a set of web sites contain the information inf ;

$keys$ is a set of keywords in the information inf ;

$alog$ is the access log of the information inf , which is a matrix of the access of the information at different location. It is defined as $\{ \langle t, location \rangle \}$, indicating that the information inf was accessed at the $location$ at the time t .

Definition 7 *Information Security System*. The laboratory information security management system can be defined as $tds = \langle ROLES, AGENTS, SERVICES, TYPES, INFORMATIONS, PROJECTS \rangle$, where *ROLES* refer to all the positions, *AGENTS* refer to all involved personnel, *TYPES* refer to all the information categories, *INFORMATIONS* refer to all information accessed, and *PROJECTS* refer to all laboratories.

45.3 Problem in Laboratory Information Security Management and the related Algorithms

During the laboratory, the identification will be verified by the door checker. The door checker will assign a unique ID for each person who accesses a computer. After a person logon a computer successfully with the ID, door checker will be noticed with the person id and the IP address of the computer. When students or teachers happen to find a harmful website, he/she may report it so that it will be filtered by the audit during the next web access. All the access to web information will be recorded by web information storage. Of course, the log will also be saved

by web information storage if someone who publish harmful information to the web.

Assume that there is an automatic algorithm to get the keywords from the web information, the problem in laboratory information security management is how to trace back to the right person who published harmful information to the web. In order to do so, the algorithms for harmful information reporting and gathering are required to make the tracing happen.

45.3.1 Algorithm of Harmful Information Reporting

Input: TYPES, PROJECTS, INFORMATIONS; AGENTS, project pr; information inf; pid, site, t; //PROJECTS is the projects to be handled; TYPES is the information types be accessed, pid is the agent id of the reporter, t is the reporting time and site is the site which contains harmful information;

Output: PROJECTS with harmful information reporting.

Find the project pr in PROJECTS;

IF (inf is not exists in INFORMATIONS) {

 Get the keywords to inf.keys from inf;

 Fetch the inform.category to inf.tid from TYPES;

 inf.sites = site;

 Fetch the location of current agent pid;

 inf.alog.add(< t, loc >);

 pr.I.add(inf); //add the harmful information in the access list

 pr.H.add(< t, pid, inf, loc >);

 INFORMATIONS.add(inf);

}

ELSE {

 Fetch the information to inf from INFORMATIONS;

 inf.sites.add(site);

 Fetch the location of current agent pid;

 inf.alog.add(< t, loc >);

 pr.I.add(inf); //add the harmful information in the access list

 pr.H.add(< t, pid, inf, loc >);

}

Algorithm 1 is used to report harmful information in a laboratory. It will first check if the information has ever been reported in that laboratory and then will set the value of web site, key words, and access log accordingly.

45.3.2 Algorithm for Harmful Information Gathering

Input: PROJECTS, HarmList = ϕ ; //PROJECTS is the projects to be handled; HarmList contains a set of harmful information in the projects;

Output: HarmList with harmful information gathering.

```
FOR (each pr in PROJECTS)
  FOR (each h in pr.H)
    HarmList.add(h.i);
```

Algorithm 2 is used to gather all the harmful information together to hand over to the audit. Since all the characteristics in the information have been updated in each reporting, the summary is not needed in this gathering process. However, if there is no such summarization in each reporting phase, the summarization should be done in gathering phase.

45.3.3 Algorithm for Harmful Information Tracing

Input: TYPES, PROJECTS, INFORMATIONS; AGENTS, inf, PList; keys; //PROJECTS is the projects to be handled; TYPES is the information types be accessed; INFORMATIONS is a set of all information and AGENTS contains all the agents, PList is a matrix of the access of the information at different location, it is defined as $\{< t, prj, ag, loc >\}$, indicating that the information inf was accessed by agent ag at the location loc at the time t in the laboratory prj; keys is used to store the temporary key words value of the information inf.

Output: PList return the possible agents who accessed the information inf with the time and location.

```
Fetch keywords to keys from inf;
found = false; //a temporary Boolean;
FOR (each oinf in INFORMATIONS and not found) {
  IF (oinf.keys = keys) {
    found = true;
    inf = oinf;
  }
}
IF (found) {
  FOR (each ilog in inf.alog) {
    aT = ilog.t;
    aL = ilog.location;
    FOR (each agent in AGENTS)
      FOR (each iavm in agent.avm)
        IF (iavm.location = aL
            and aT >= iavm.s and aT <= iavm.e) {
          FOR (each pr in PROJECTS) {
```

```

        IF(agent in pr.R and aT >=pr.s
           and aT <=pr.e)
            PList.add(<aT,pr,agent,aL>);
        }
    }
}
}
}

```

Algorithm 3 is used to trace back the access log to the harmful information. Typically it is done on event that there is someone publish harmful information, and the performance is not very important.

45.4 Conclusion and Status of Research

The problem in laboratory information security management is introduced in laboratory management in the high education environment. A research is being conducted to adopt E-CARGO model to report and gather the harmful information report, and then enable the traceability of harmful information with the roles collaboration. Here in this paper, E-CARGO model is briefly introduced and the definitions of the model have been presented. The algorithms of harmful information reporting, gathering and tracing have been proposed.

When a student checks out at the door checker, the brief account status of the students will be displayed at the welcome screen. In this research, the logon and log out status is kept by door checker.

Currently, we are going to develop a framework with the involved algorithms to guarantee the information security in laboratory management. The benefit and weakness of the framework and related algorithms may be analyzed during the practice and will be improved or optimized accordingly.

Acknowledgments This research is supported by Zhejiang Gongshang University, China with No. XGZ1102 to Dr. Weigang Chen. This research was financial supported by National Natural Science Foundation of China with No. 60873022 to Dr. Hua Hu, The Science and Technology Department of Zhejiang Province, China with No. 2008C11009 to Dr. Bin Xu, and No. 2008C13082 to Dr. Xiaojun Li, and Education Department of Zhejiang Province with No. 20061085 to Dr. Bin Xu. Thanks to the students in development team. Their prototype partly verified the suggested algorithms.

References

1. Xu B, Zhou HY, Ge YJ (2010) Quick tendency forecast on internet public opinions. In: Proceedings of 7th international conference on fuzzy systems and knowledge discovery, pp 1493–2497

2. Zhu JH, Xu B, Jiang B, Chen WG (2010) Identifying harmful web pages in laboratory information security management. In: Proceedings of 2010 2nd international workshop on intelligent systems and applications, pp 1–4
3. Yu B, Xu B, Ling Y (2009) Equipment conflict checking and removal for innovation talent development in IT undergraduate education via E-CARGO collaborative model. In: Proceedings of 4th international conference on computer science and education, pp 1669–1672
4. Zhu H (2006) Role mechanisms in collaborative systems. *Int J Prod Res* 41(1):181–193
5. Zhu H, Zhou MC (2006) Role-based collaboration and its kernel mechanisms. *IEEE Trans Sys Man Cybern Part C* 36(4):578–589
6. Zhu H (2006) A role-based approach to robot agent team design. *IEEE Int Conf Sys Man Cybern*, Taipei, China, pp 4861–4866
7. Zhu H (2007) Role as dynamics of agents in multi-agent systems. *Sys Inf Sci Notes* 1(2):165–171
8. Zhu H, Zhou MC (2008) Role transfer problems and algorithms. *IEEE Trans Sys Man Cybern Part A* 38(6):1442–1450
9. MESA International. Collaborative manufacturing explained, available online <http://www.mesa.org>
10. Xu B, Yang XH, Shen YH, Li SP, Ma A (2008) A role-based SOA architecture for community support systems. In: Proceedings of the international symposium on collaborative technologies and systems. (CTS'08), Irvine, CA, pp 408–415
11. Xu B (2009) Reflection evaluation in online e-learning system via hyperlinks analysis. In: Proceedings of 4th international conference on computer science and education, pp 1673–1677

Chapter 46

An Improved GVF Snake Model Using Magnetostatic Theory

Bingyu Chen, Jianhui Zhao, Erqian Dong, Jun Chen, Yang Zhao
and Zhiyong Yuan

Abstract In this paper, we propose an improved gradient vector flow (GVF) snake to address the problem of poor convergence of GVF snake for the deep concavities in image segmentation. The major contribution is that new external force field is proposed combining the properties of GVF snake model and magnetostatic active contour (MAC) model, and the new external force field can help move the snake contour into deep boundary concavities. The proposed algorithm has been tested with experiments on various types of images, and significant improvement has been achieved in the convergence for approaching boundary concavities compared with the existing GVF snake method.

Keywords Image segmentation · GVF snake · Active contour method · MAC

46.1 Introduction

Active contour model (snake) [1], widely utilized in digital image processing and computer vision for the recent years, is particularly suitable to solve the problems such as extracting the boundaries of target objects, which is one of the major methods for image segmentation. Image segmentation is one of the basic problems in digital image processing and computer vision, and its purpose is to separate an image into several different regions. The original snake method was introduced by

B. Chen · J. Zhao (✉) · E. Dong · J. Chen · Y. Zhao · Z. Yuan
Computer School, Wuhan University, Wuhan, 430072 Hubei,
People's Republic of China
e-mail: jianhuizhao@whu.edu.cn

Kass et al. in 1987. In the next two decades, snake and its improved models have been widely studied and applied in image segmentation [2–10].

Snake defines a closed curve, it moves through the spatial domain of an image under the influence of the internal force coming from the curve itself and the external force computed from the image data by minimizing the energy function. During the deformation of active contour, the internal force calculated from the curvature of active contour provides smoothness to the contour, while the external force calculated from image data pushes the active contour toward the object boundary. The deformation finally stops when the active contour reaches the minimum of energy function. Generally there are two types of active contour models in the literature: parametric active contour models [2–4, 6] and level set active contour models [7–9]. Due to the availability of the efficient numeric methods, parametric active contour models are often faster than level set ones [8–11], and thus received long-term attentions.

GVF snake, proposed by Xu et al. [6], can help solve the problems of traditional active contour models. The GVF force field, based on the diffuse operation, is obtained by getting the minimum of the energy function. The minimum of the energy function is calculated through solving a pair of linear partial differential equations and the partial differential equations can diffuse the vector flow of a gray-level image or binary edge map. External force field of the traditional active contour model is a static irrotational field while that of the GVF snake can be decomposed into an irrotational component and a solenoidal component. But the GVF snake force field is still a static force field because that the force field is calculated before the deformation of active contour and will not be updated during the deformation procedure. During the diffusion process of GVF force field, the image noise is weakened so that the robustness of GVF snake is better than traditional snake. GVF snake's advantages over the traditional one are its insensitivity to the initialization and ability to move into the boundary concavities. Besides, GVF snake has a large capture range, which means that, barring interference from the other objects, it can be initialized far away from the boundary. However, the force field of GVF snake is stationary and this makes it impossible for GVF snake to move the active contour into very deep boundary concavities.

Magnetostatic active contour (MAC) model, proposed by Xie and Mirmehdi [9], hypothesizes electric currents flowing through both object boundary and active contour, so that the active contour will move toward object boundary under the influence of the magnetic fields generated by each current. The external force field of MAC is a dynamic force field and is able to capture complex boundary and multiple objects even if there is only one initial contour at the beginning of evolution of MAC. MAC force field is updated dynamically during the deformation of the active contour. However, force calculation in the magnetic field is very complex, and the external force field computation becomes time-consuming, which limits the application of MAC.

Comparatively, the deformation of level set active contour models needs a higher dimensional function, which makes their calculation slower than parametric active contour models. But for many applications, only one object needs to be

extracted at last [12, 13], thus parametric active contour modeling becomes a better choice when applying on these applications. Therefore, this paper will just focus on the parametric active contour models as well.

To solve the above problem of GVF method, an improved GVF snake is presented, i.e. we apply the external force of GVF snake to seize the boundary of the active contour combining with the external force of MAC model for the accurate convergence to the deep concave regions. The new external force field of the improved GVF snake is updated dynamically according to the position of active contour so that the active contour can move into deep boundary concavities.

46.2 Parametric Snake Model

Original snake model is a closed curve $X(s) = [x(s), y(s)]$, $s \in [0, 1]$ and can move through the spatial domain of an image to minimize the energy function. The initial snake curve is set by human interaction, and the energy function is defined as follows:

$$E(X(s)) = \int_0^1 \frac{1}{2} (\alpha |X'(s)|^2 + \beta |X''(s)|^2) + E_{ext}(X(s)) ds \quad (46.1)$$

where α and β are weighting parameters of contour elasticity and rigidity, $X'(s)$ and $X''(s)$ denote the first and the second derivatives of $X(s)$ with respect to s . $E_{ext}(X(s))$ is the external energy derived from image so that it has its smaller values at the object boundary. The minimization of (46.1) must satisfy the Euler equation:

$$\alpha X''(s) - \beta X''''(s) - \nabla E_{ext} = 0 \quad (46.2)$$

where $X''''(s)$ is the fourth derivative of $X(s)$. This can be viewed as a force balance equation:

$$F_{int} + F_{ext} = 0 \quad (46.3)$$

where $F_{int} = \alpha X'' - \beta X''''(s)$, $F_{ext} = -\nabla E_{ext}$. The internal force F_{int} discourages both stretching and bending while the external force F_{ext} pulls the active contour towards the desired image edges. Finally the curve will stop at the position with force balance.

Treat active contour X as a function of time t and curve parameter s , the derivative of X with respect to t is defined as follows:

$$X_t(s, t) = \alpha X'' - \beta X''''(s) - \nabla E_{ext} \quad (46.4)$$

When $X(s, t)$ is stable, $X_t(s, t)$ will be zero, and we get a solution of (46.2), so does the minimum of energy function and the final segmentation result.

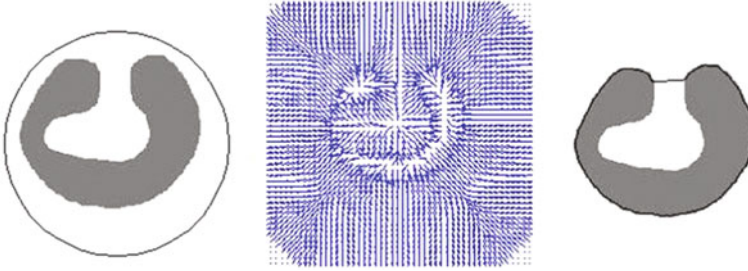


Fig. 46.1 Object and initial curve (*left*), basic external force field (*middle*), final deformation result using basic external force field (*right*)

46.3 GVF and MAC Based External Force

Our approach first defines an edge map $f(x, y)$ derived from the image $I(x, y)$. We use (46.5) to calculate the edge map $f(x, y)$, and the value of edge map is larger at the position near the image edges.

$$f(x, y) = |\nabla(G_\sigma(x, y) * I(x, y))|^2 \quad (46.5)$$

where $G_\sigma(x, y)$ is a two-dimensional Gaussian function with standard deviation σ , and ∇ is the gradient operator.

Based on the edge map, the basic external force field $V(x, y) = (u(x, y), v(x, y))$ will be obtained with the diffusion method of GVF snake. The energy function used in force field diffusion is defined as follows:

$$E = \iint \mu \left(u_x^2 + u_y^2 + v_x^2 + v_y^2 \right) + |\nabla f|^2 |V - \nabla f|^2 dx dy \quad (46.6)$$

The first term of (46.6) is a smooth term, while u and v are determined by Laplace's equation within the homogeneous region. When ∇f is large, the second term dominates the integrand and the minimum of the integrand will be obtained by setting $V = \nabla f$. The energy function makes V nearly equal to the gradient of edge map when it is large and diffuses V from object boundary to homogeneous regions.

As illustrated in Fig. 46.1, the object has very deep boundary concavities and the capture range of basic external force field is the whole image. The result shows that the basic external force field is incapable of helping progress into the deep boundary concavities. The basic force field provides no downward force at entrance of boundary concavities so that the deformation of active contour stops there. While the key idea of our approach is to add a downward force component to the basic force field so that the active contour will continue to move toward the deep concavities.

The active contour at entrance of boundary concavities is called fake contour. The fake contour has a characteristic that the angle between its tangent and basic external force field has a small range, while the angle between the tangent of object boundary and basic external force field has a bigger one. In our method, we calculate MAC external force of fake contour and combine it with the basic external force to obtain the new external force field. The new external force field has downward force, and thus the active contour will move toward the deep boundary concavities under the influence of the new external force field.

Magnetostatic theory is applied to by charging both the object boundary and fake contour with electric currents. The electric currents of object boundary produce the magnetic field and the active contour will move toward object boundary with the influence of the magnetic field. The electric current of object boundary can be obtained using:

$$O(x, y) = (-1)^\lambda (-I_y(x, y), I_x(x, y)) \quad (46.7)$$

where $I_x(x, y)$ and $I_y(x, y)$ are the partial derivatives in x and y for $I(x, y)$. When $\lambda = 1$, the electric current is counterclockwise, while when $\lambda = 2$, the electric current is clockwise. The magnetic force introduced by object boundary is defined as:

$$B(x, y) = \frac{\mu_0}{4\pi} \sum_{c \in f(c)} I_{f(c)} \Gamma(c) \times \frac{\hat{R}_{(x,y)c}}{R_{(x,y)c}^2} \quad (46.8)$$

where c denotes pixels on object boundary, $f(c)$ is object boundary, $\Gamma(c)$ is the electric current vector at c , $\hat{R}_{(x,y)c}$ is the unit vector from (x, y) to c and $R_{(x,y)c}$ is the distance between them. The current applied to the fake contour is I_s , and the force exerted by magnetic force on I_s is computed by:

$$F(x, y) = I_s \Upsilon(s) \times B(x, y) \quad (46.9)$$

where $\Upsilon(s)$ is the electric current vector on the fake contour position s . The force $F(x, y)$ is always imposed on the contour normal along inward or outward direction.

We combine both $V(x, y)$ and $kF(x, y)$ to obtain the final external force field, and k is the weight parameter of magnetic force. The final external force field shown in Fig. 46.2 has forces that point into boundary concavities.

46.4 Experimental Results

The experimental results using the improved external force field are shown in Fig. 46.3, which indicates that the snake finds the deep concave-shaped boundary. Compared with the GVF snake result in Fig. 46.1, the improved GVF snake is

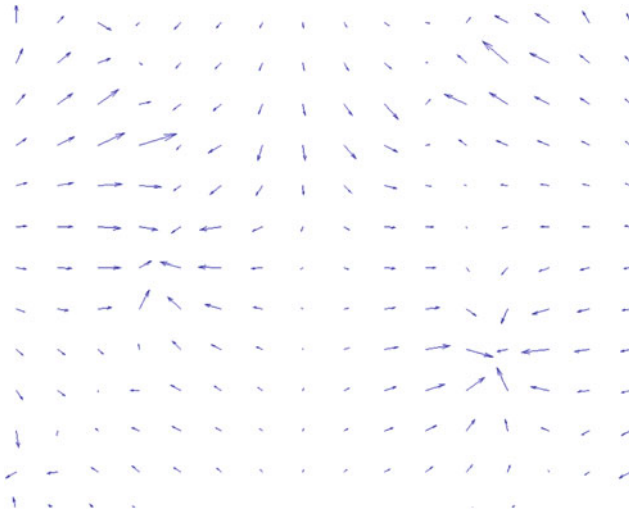


Fig. 46.2 The improved external force field



Fig. 46.3 Improved GVF snake on object with deep concave-shaped boundary

capable of moving into the deep boundary concavities. As displayed in Fig. 46.4, the improved snake method is also able to find the correct boundary of object with several deep concavities. The result on the ultrasound medical image is shown in Fig. 46.5, from which we can find the improved GVF snake is less sensitive to noise. Besides, experiment result on MRI brain image is shown in Fig. 46.6, result on fire region is shown in Fig. 46.7, result on broken area of road is shown in Fig. 46.8, result on baseball area in the satellite image is shown in Fig. 46.9, and result on mouth region of human face is shown in Fig. 46.10.

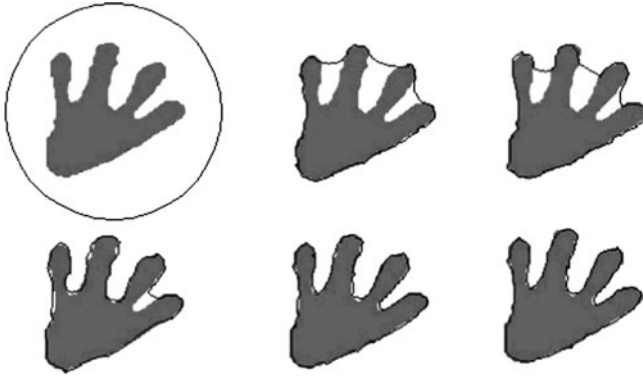


Fig. 46.4 Improved GVF snake on object with several deep concavities

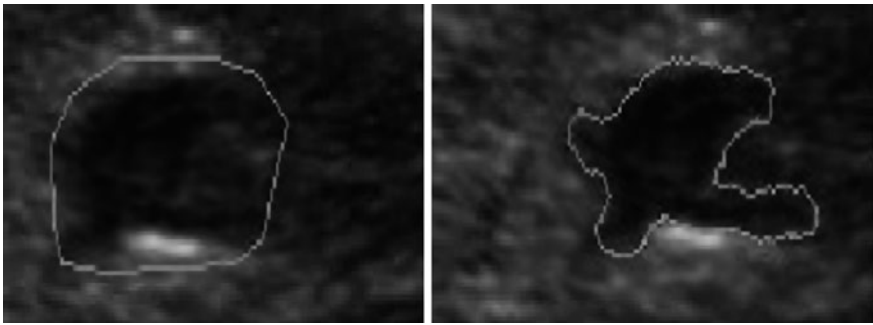


Fig. 46.5 Improved GVF snake on ultrasound medical image

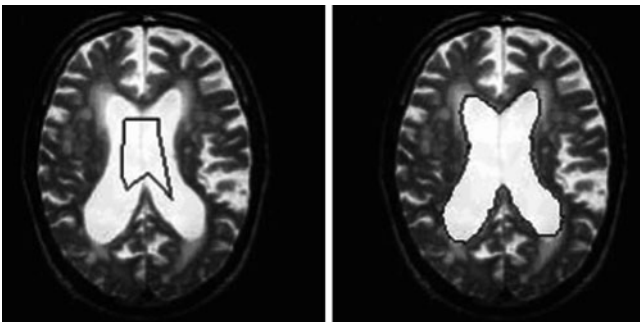


Fig. 46.6 Improved GVF snake on MRI brain image



Fig. 46.7 Improved GVF snake on fire region

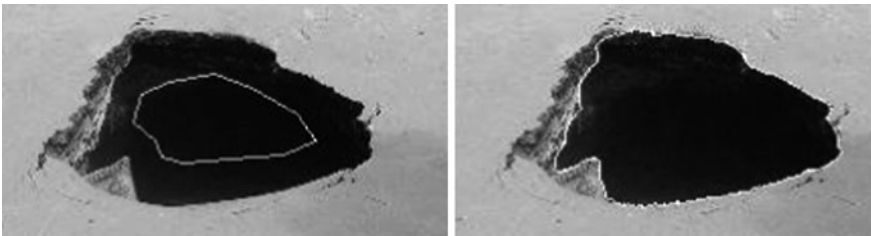


Fig. 46.8 Improved GVF snake on broken area of road

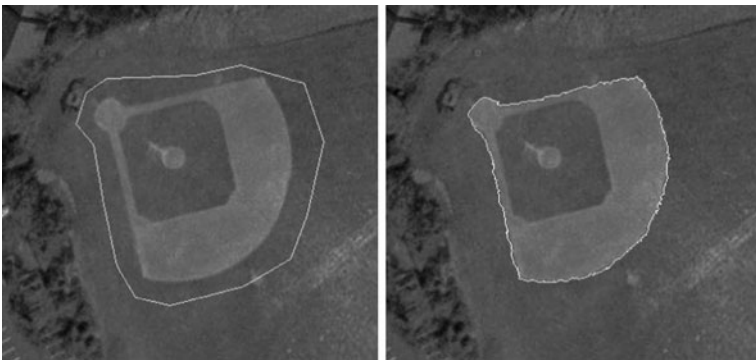


Fig. 46.9 Improved GVF snake on baseball area of satellite image

46.5 Conclusion

GVF snake has a larger capture range than the original snake, but its incapability of progressing into deep boundary concavities also exists. We propose an improved GVF snake to help address the issue of inability to extract the deep boundary concavities. The improved GVF snake in our method combines the

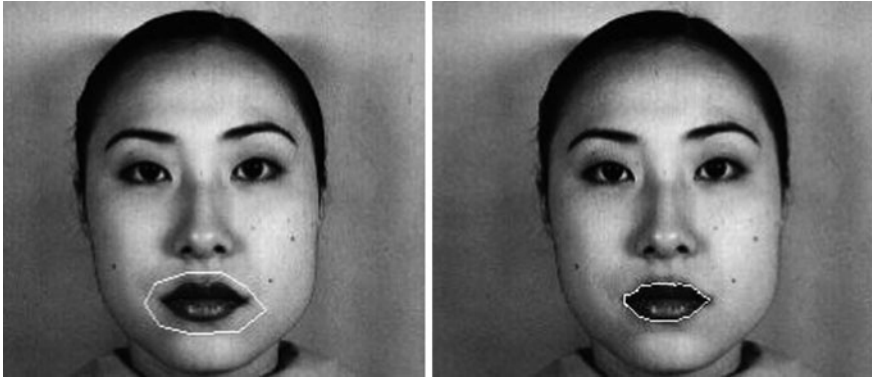


Fig. 46.10 Improved GVF snake on mouth region of human face

advantages of both GVF snake model and MAC model, and its new external force field can help move the snake contour into the deep boundary concavities. Experiments on different types of images show that the improved GVF snake produces better results than GVF snake.

Acknowledgments This work was supported by National Basic Research Program of China (973 Program, No. 2011CB707904), Hubei Provincial Natural Science Foundation of China, Research Foundation (No. AISTC2008_16) from State Key Laboratory of Aerospace Information Security and Trusted Computing Ministry of Education, and Fundamental Research Funds for the Central Universities.

References

1. Kass M, Witkin A, Terzopoulos D (1988) Snakes: active contour model. *Int J Comput Vis* 1(4):321–331
2. Cohen L (1991) On active contour models and balloons. *CVGIP Image Understand* 53(2):211–218
3. McInerney T, Terzopoulos D (2000) T-snakes: topologically adaptive snakes. *Med Image Anal* 4(2):73–91
4. Cohen L, Cohen I (1993) Finite-Element methods for active contour models and balloons for 2-D and 3-D images. *IEEE Trans. Pattern Analysis and Machine Intelligence* 15(11):1131–1147
5. Ostlad B (1996) A. Tonp: encoding of a priori information in active contour models. *IEEE Trans. Pattern Analysis and Machine Intelligence* 18(9):863–872
6. Xu C, Prince J (1998) Snakes, shapes, and gradient vector flow. *IEEE Trans. Image Processing* 7(3):359–369
7. Malladi R, Sethian J, Vemuri B (1995) Shape modeling with front propagation: a level set approach. *IEEE Trans. Pattern Anal and Machine Intelligence* 17(2):158–171
8. Caselles V, Kimmel R, Sapiro G (1997) Geodesic active contours. *Int J Comput Vis* 22(1):61–79

9. Xie X, Mirmehdi M (2008) Magnetostatic active contour (MAC). *IEEE Trans. Pattern Anal and Machine Intelligence* 30(4):632–646
10. Sum KW, Cheung PYS (2007) Boundary vector field for parametric active contours. *Pattern Recognit* 40(6):1635–1645
11. Juan O, Keriven R, Postelnicu G (2006) Stochastic motion and the level set method in computer vision: Stochastic active contours. *Int J Comput Vis* 69(1):7–25
12. He L, Peng Z, Everding B, Wang X, Han CY, Weiss KL, Wee WG (2008) A comparative study of deformable contour methods on medical image segmentation. *Image Vis Comput* 26:141–163
13. Wang T, Cheng I, Basu A (2009) Fluid vector flow and applications in brain tumor segmentation. *IEEE Transactions on Biomedical Engineering* 56(3):781–789

Chapter 47

Block Effect Reduction via Model Based Compressive Sensing

Jin Jianqiu, Zhang Zhiyong and Jiang Zhaoyi

Abstract In this chapter, we propose a novel block effect reduction algorithm based on Model based compressive sensing (MCS). Block effect reduction can be considered as image recovery from a degraded image. It is exactly what compressive sensing does. According to MCS, our approach can catch the tree structured sparseness of natural images in wavelet domain and the discontinuity between adjacent blocks in JPEG images. Hence, our approach has a good performance in visual quality and PSNR as shown in our intensive experiments.

Keywords Compressive sensing · Image deblocking · Wavelet transform · Tree-structured sparsity

47.1 Introduction

Currently, block-based compression algorithms are used in many popular image/video compression standards, including JPEG, MPEG, H.264, and so on. However, it is well known that they present annoying artifacts referred to block effect at low

This work is supported by No. 2009C11034, Zhejiang key scientific and technological project.

J. Jianqiu (✉) · Z. Zhiyong · J. Zhaoyi
College of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou, 310018, People's Republic of China
e-mail: jqjin@mail.zjgsu.edu.cn

Z. Zhiyong
e-mail: zzy@mail.zjgsu.edu.cn

J. Zhaoyi
e-mail: zyjiang@mail.zjgsu.edu.cn

bit rate. For instance, in JPEG standard an image is partitioned into 8×8 blocks before a DCT transform is applied to each block separately, and then each block is quantized and coded separately. At reconstruction, images present discontinuities between adjacent blocks which are especially visible in smooth regions.

In a few of years, many block effect reduction algorithms have been proposed to remove these artifacts. These algorithms can be classified into two classes—pre-processing ones and post processing ones. In preprocessing methods the coding procedure should be modified to improve the quality of compressed images. They are not compatible to current compression standards because both the codec and the decoder need to be modified. However, post processing methods reduce the block artifacts through modifying the decoding procedure instead of the coding one. Hence, the post processing methods have been gotten more attention in this research field.

In this chapter, we shall focus on block effect reduction for JPEG compression standard, but we point out that our idea can be natural extended to other relevant block-based image/video compression algorithms. We will propose a novel post processing method of block effect reduction for JPEG images based on MCS (short for Model based compressive sensing [1]). It can run fast and robust, and improve the visible quality and the PSNR of the compressed images significantly. We organize the rest of this chapter as follows: We review the related works of block effect reduction algorithms in next section. In Sect. 47.3 compressive sensing [2] and MCS theories are introduced in brief. Our method is proposed in Sect. 47.4, and then its experimental results are given in Sect. 47.5. In Sect. 47.6, we discuss the relationship between our method and other methods, and future works.

47.2 Related Works

There are two classes of the post processing methods for block effect reduction. One is based on image enhancement, and the other is based on image restoration. The former usually smooth the discontinuities between adjacent blocks by low pass filter, and the latter can be considered as restoring image from degraded image. Our method to be proposed belongs to the latter. For this reason, we emphasize reviewing this class of algorithms.

Projection on Convex Set (POCS [3]) and Maximum a posteriori probability (MAP) are the popular two categories of image restoration algorithms. They both take advantage of prior information of image to restore image as best as possible. To our knowledge, POCS for block effect reduction is firstly proposed by Zakhor et al. [4]. The literature argued that natural image is band limited, and the discontinuities between adjacent blocks violate this property. Therefore, the constraints both band limitation and the coefficients of block DCT falling into the feasible interval were imposed on images. The set resulting from each constraint is convex. According to the theory of projection on convex set, if their intersection is not empty, the iteration, which projecting alternately image on the two convex

sets, must converge to a solution. If having many kinds of prior information about ground truth image, we can construct many convex sets respectively. And POCS can guarantee the convergence of the iteration. After the literature [4], many researchers extended POCS method by developing various kinds of prior information of natural images and describing them in form of convex sets [5–7]. It should be pointed out that the more convex sets may make the convergence of the iteration slower.

MAP is also called Bayesian estimation. It supposes the ground truth signal is sampled from a prior probability distribution firstly, and then the post probability distribution is computed from noisy/degraded signal. At final, an optimization problem is gotten which maximizes the post probability distribution. The literature [8] applied the theory of Markov Random Field to select Gibbs distribution, which energy function is the sum of square of gradient image, as prior probability distribution. This selection can reduce the block effect of JPEG image at low bit rate, but at the same time it may blur the image so that some detail of image is removed. On the basis of this work, various kinds of MRF are developed to reduce block effect, such as transform-domain MRF (TD-MRF) [9], and so on. Since more prior information is considered, these algorithms have better performance in block effect reduction, but make solving the maximization of post probability distribution more difficult and slower.

Aforementioned algorithms for block effect reduction are time-consuming tasks because they need to solve a large scale nonlinear programming problem. Furthermore, they reduce the block effect with blurring image unavoidably. In this chapter, we will propose a fast and effective block effect reduction algorithm based on Model based compressive sensing. It belongs to image restoration methods, but it can smooth the discontinuity between adjacent blocks accurately, at the same time preserve other image details as far as possible. In next section we will introduce the theory of model based compressive sensing before presenting our block effect reduction algorithm.

47.3 Model Based Compressive Sensing Theory

The theory of compressive sensing, which is a new signal sampling and reconstruction theory proposed by Candes and Tao [10] and Donoho in 2006 [11], allows us to reconstruct a signal from only a few samples if it is sparse in a transform domain. To see how, suppose that we have an n -dimensional signal $\mathbf{x} \in \mathbb{R}^n$ we are trying to estimate with k random samples, where $k \ll n$. We can write $\mathbf{y} = \mathbf{S}\mathbf{x}$, where \mathbf{S} is a sampling matrix which performs the linear measurements on \mathbf{x} . Initially, it seems that perfect reconstruction of \mathbf{x} from \mathbf{y} is impossible, given that the k samples yield a $(n-k)$ -dimensional subspace of possible solution for the original \mathbf{x} that would match our given observations. This is where we use the key assumption of compressed sensing: we assume that the signal \mathbf{x} is sparse in some transform domain Ψ , i.e., $\Psi\mathbf{x}$ is sparse. Mathematically, an n -dimensional

signal is m -sparse if it has at most m nonzero coefficients (where $m \ll n$), which can be written in term of the ℓ_0 norm. This is not an unreasonable assumption for real-world signals such as images, since this fact is exploited in transform-coding compression standard such as JPEG and JPEG2000. Though estimating \mathbf{x} from \mathbf{y} is severely undetermined, one of the key results in compressive sensing shows that we can solve for \mathbf{x} uniquely by searching for the sparsest $\Psi\mathbf{x}$ if $k > 2\|\Psi\mathbf{x}\|_0$ [10]. That is the following equation:

$$\arg \min_{\mathbf{x}} \|\Psi\mathbf{x}\|_0 \quad s.t. \quad \mathbf{y} = \mathbf{S}\mathbf{x} \quad (47.1)$$

The solution to this problem, however, involves a combinatorial algorithm which is known to be NP-complete and is intractable for reasonably-sized signal. However, recent results [12] show that Eq. 47.1 can be solved approximately by replacing the ℓ_0 with an ℓ_1 -norm, or by greedy reconstruction algorithms such as OMP [13], ROMP [12], et al. As long as the coherence between \mathbf{S} and Ψ is small and the number of samples $k = O(m \log n)$, these algorithms have the guarantee of exact recovery [12].

Model based compressive sensing (MCS) extends CS theory. MCS not only exploits the sparseness of a signal, but also the structure of the signal. In this way, less samples are required to reconstruct the original signal. Considering an extreme example, if the locations of nonzero coefficients of a k -sparse signal are known, k samples are enough for reconstructing the signal. For real-world signal, such as natural images, MCS can also improve the efficiency of signal reconstruction significantly [1], i.e., less samples are required.

Many works [14, 15] have show the wavelet coefficients of natural image can be naturally organized into a tree structure, and for many kinds of natural and manmade signals the largest coefficients cluster along the branches of this tree. According to this, La and Do proposed an effective signal reconstruction method using sparse tree representation [16]. In next section, we will propose an image reconstruction algorithm from low-bit-rate JPEG image which can catch the tree structure in wavelet domain and the discontinuities between adjacent blocks.

47.4 Our Approach

Compressive sensing can be considered as recovering a signal from inaccurately and/or partially measured samples. It is exactly what block effect reduction does. According to this, we have the following block effect reduction algorithm for JPEG image:

$$\arg \min_{\mathbf{x}} \|\Psi\mathbf{x}\|_0 \quad s.t. \quad \mathbf{Q}_r(\mathbf{B}(\mathbf{x})) = \mathbf{y} \quad (47.2)$$

where \mathbf{B} is the block discrete cosine transform (BDCT) used in JPEG standard, \mathbf{Q}_r is the quantization operator with quantization step r , and \mathbf{y} is the quantized coefficients of BDCT decoding from a JPEG file. We should seek a transform Ψ

that $\Psi\mathbf{x}$ is sparse. We know wavelet transform and wavelet frame transform may be good candidates. How to choose suitable Ψ will be discussed below.

Equation 47.2 does not catch the tree structure and the annoying block effect. Thus it is likely to smooth the image excessively with removing some details of the image. In order to overcome it, we apply the theory of MCS to integrate tree constraint by replacing Eq. 47.2 with

$$\arg \min_{\mathbf{x}} \|\Psi\mathbf{x}\|_{\mathcal{T}_K} \quad s.t. \quad \mathbf{Q}_r(\mathbf{B}(\mathbf{x})) = \mathbf{y} \quad (47.3)$$

where $\|\cdot\|_{\mathcal{T}_K}$ is defined by

$$\|\mathbf{x}\|_{\mathcal{T}_K} = \inf_{\bar{\mathbf{x}} \in \mathcal{T}_K} \|\mathbf{x} - \bar{\mathbf{x}}\|_1 \quad (47.4)$$

where \mathcal{T}_K is the K-tree sparse subspace of \mathbb{R}^N (N is the dimension of \mathbf{x}) [1]. The problem Eq. 47.3 is usually solved by iterative algorithms. However, to avoid the discontinuity between adjacent blocks being treated as edge, we should choose suitable initial value of \mathbf{x} which contain no the discontinuity between adjacent blocks. Furthermore, it should be pointed out that \mathbf{Q}_r is not a linear transform, so Eq. 47.3 cannot be solved by general numerical methods for MCS. We need to tailor a numerical algorithm to it.

We firstly consider how to choose suitable \mathbf{x}_0 as initial value of \mathbf{x} . \mathbf{x}_0 should contain no block effect. Hence, we choose the resulting image of soft thresholding as \mathbf{x}_0 :

$$\mathbf{x}_0 = \Psi^* \rho_\lambda(\Psi\mathbf{B}^{-1}(r\mathbf{y})) \quad (47.5)$$

where ρ_λ is a soft thresholding operator:

$$\rho_\lambda(x) = \text{sgn}(x) \cdot \max(0, |x| - \lambda) \quad (47.6)$$

and Ψ^* is the dual of Ψ which is set by the translation invariant wavelet transform with the CDF 9/7 basis [17]. This is to avoid the ripple effect resulting from soft thresholding in orthogonal/biorthogonal wavelet transform domain.

Now, it is time to present a numerical method to Eq. 47.3. Motivated by POCS methods and Model-based CospMP method [18], we propose a numerical approach to solve Eq. 47.3 by alternatively performing CSSA algorithm [19] and projecting the BDCT coefficients of the image on the feasible set to satisfy the constraint in Eq. 47.3. Concretely, from Eq. 47.3 we shall seek a recovered image $\hat{\mathbf{x}}$ such that it meets the following requirements:

- a. $\Psi\hat{\mathbf{x}}$ should be tree-structured sparse. This can be done by condensing sort and select algorithm (CSSA) [19].
- b. $\mathbf{Q}_r(\mathbf{B}(\hat{\mathbf{x}})) = \mathbf{y}$. This can be done by projecting $\mathbf{B}(\mathbf{x})$ on $\mathbf{Q}_r^{-1}(\mathbf{y})$:

$$\mathbf{P}_{r,\mathbf{y}}(\mathbf{z}) = \arg \min_{\mathbf{z} \in \mathbf{Q}_r^{-1}(\mathbf{y})} \|\hat{\mathbf{z}} - \mathbf{z}\|_2 \quad (47.7)$$

Taking all in together, we obtain the **Algorithm 1** as follows.

Input: The quantized coefficients \mathbf{y} , quantization step r , Ψ , Ψ^* and λ in Eq. 47.5, K in Eq. 47.4

Output: A recovered image $\hat{\mathbf{x}}$

Initialization: initializing \mathbf{x}_0 from Eq. 47.5, $i = 0$

Iteration:

While halting criterion false do

1. $i = i + 1$
2. Performing CSSA algorithm on \mathbf{x}_{i-1} to get \mathbf{x}_i using Ψ , Ψ^* and K .
3. $\mathbf{x}_i = \mathbf{B}^{-1} \mathbf{P}_{r,y}(\mathbf{B}(\mathbf{x}_i))$

End while

Return: $\hat{\mathbf{x}} = \mathbf{x}_i$

The above algorithm mainly consists of two procedures. Each of them can be considered as projection on a set. Fortunately, the two sets are all convex. Hence, our algorithm is convergent according to POCS theory. Furthermore, the two procedures run fast because CSSA algorithm can be performed with $O(n \log n)$ cost and $\mathbf{P}_{r,y}$ is an $O(n)$ -cost operator.

47.5 Experimental Results

Our intensive experiments show our method has good performance in PSNR and visual quality. Here, limited to space, we present three test images, e.g., Lena, Boat and Barbara. As a comparison, we also give out the experimental results of Cai et al. [20].

As shown in Fig. 47.1, the images in the second row, which are compressed by JPEG with global step $r = 5$, have visible block artifacts. In third row, the images are results of Cai et al. In them no block artifact can be found, but over smoothness is presented. The fourth row is our results which have good visual quality and higher PSNR than the third row.

47.6 Summary

In this chapter, we propose a novel block effect reduction algorithm based on MCS for JPEG compression standard. But this idea can be naturally extended to other block-based image/video compression methods by replacing in Eq. 47.3 with corresponding transform. From the final numerical algorithm (Algorithm 1), we find there are some similarities between our method and POCS methods. Algorithm 1 is deduced from MCS theory, but it also can be considered as projecting on multiple convex sets alternately.



Fig. 47.1 Experimental results. From *top to bottom*: original images, JPEG images with global quantization step $r = 5$, Cai et al.'s results, our results

References

1. Baraniuk RG, Cevher V, Duarte MF, Hegde C (2010) Model-based compressive sensing. *IEEE Trans Inf Theory* 56(4):1982–2001
2. Candes E, Wakin M (2008) An introduction to compressive sampling. *IEEE Sig Process Mag* 25(2):21–30

3. Zhao Y, Cheng G, Yu S (2006) A POCS-based algorithm for blocking artifacts reduction. *J Shanghai Jiaotong Univ (Sci)* 11(3):87–91
4. Zakhor A (1992) Iterative procedures for reduction of blocking effects in transform image coding. *IEEE Trans Circuits Syst Video Tech* 2(1):91–95
5. Alter F, Durand S, Froment J (2004) Deblocking DCT-based compressed images with weighted total variation. In: *Proceedings of IEEE international conference on acoustics, speech, and signal processing (ICASSP '04)*, vol 3, pp iii-221–iii-224
6. Patrick L, Combettes VERW (2005) Signal recovery by proximal forward-backward splitting. *Multiscale Model Simul* 4(4):1168–1200
7. Durand S (2003) Reconstruction of wavelet coefficients using total variation minimization. *SIAM J Sci Comput* 24:1754–1767
8. O'Rourke TP, Stevenson RL (1995) Improved image decompression for reduced transform coding artifacts. *IEEE Trans on Circuit Syst Video Tech* 49(4):490–499
9. Li Z, Delp EJ (2005) Block artifact reduction using a transform-domain Markov random field model. *IEEE Trans Circ Syst Video Tech* 15(12):1583–1593
10. Candes EJ, Tao T (2006) Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans Inf Theory* 52(12):5406–5425
11. Donoho D (2006) Compressed sensing. *IEEE Trans Inf Theo* 52(4):1289–1306
12. Needell D, Vershynin R (2010) Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit. *IEEE J Sel Top Sig Process* 4(2):310–316
13. Tropp JA (2004) Greed is good: algorithmic results for sparse approximation. *IEEE Trans Inf Theory* 50(10):2231–2242
14. Shapiro J (1993) Embedded image coding using zerotrees of wavelet coefficients. *Sig Process IEEE Trans* 41(12):3445–3462
15. Taubman D (2000) High performance scalable image compression with ebcot. *Image Process IEEE Trans* 9(7):1158–1170
16. La C, Do M (2005) Signal reconstruction using sparse tree representations. *SPIE Wavelets XI, San Diego*
17. Daubechies I (1992) Ten lectures on wavelets, ser. CBMS-NSF regional conference series in applied mathematics, vol 6. Society for Industrial and Applied Mathematics (SIAM), Philadelphia
18. Needell D, Tropp J (2009) CoSaMP: iterative signal recovery from incomplete and inaccurate samples. *Appl Comput Harmon Anal* 26(3):301–321
19. Baraniuk RG, DeVore RA, Kyriazis G, Yu XM (2002) Near best tree approximation. *Adv Comput Math* 16(4):357–373 May
20. Cai J-F, Ji H, Shang F, Shen Z (2010) Inpainting for compressed images. *Appl Comput Harmon Anal* 29(3):368–381

Chapter 48

The Design of Logistics Information Platform for the Yangtze River Delta

Peihua Fu and Xiaoli Gu

Abstract Logistics Information Platform for the Yangtze River Delta is to play an important function of regional logistics infrastructure system and it will benefit the economic development of Yangtze River Delta. Research Development of Logistics Information Platform at home and abroad and status of the Yangtze River Delta Logistics, then explore the various departments of the requirements of logistics information platform and network structure of this platform, the overall framework of the functions to be achieved on key technologies.

Keywords Yangtze river delta · Logistics information platform · Logistics Information System

48.1 Introduction

80 years since the 20th century, the United States, Japan, Europe and other developed countries have started a variety of logistics functions, to integrate elements of the “logistics revolution.” The Japanese Government has introduced Integrated Logistics Policy in Apr 1997 and Jul 2001 to promote logistics information platform to grow from there. Hong Kong Government to promote the construction of a Digital Trade and Transportation Network, a regional-level effort to build an open, stable and secure logistics information service platform. Singapore portnet port logistics

P. Fu (✉) · X. Gu
College of Computer and Information Engineering, Zhejiang Gongshang University,
Hangzhou, Zhejiang, People’s Republic of China
e-mail: fph666@hotmail.com

X. Gu
e-mail: ttcumt@163.com

information service platform, Hamburg, Germany dakosy Logistics Information Platform, Rotterdam, the Netherlands wave service information platform, New York, New Jersey FIRST platforms are fully invested by the government. China Logistics Information Platform is in its infancy. Although some cities and regions are to proceed with the planning and construction of logistics information platform, but the construction of the popularity of logistics information platform range is relatively small, the overall level is relatively backward. And it can not meet the requirements of the development of modern logistics, so to adopt vigorous measures to actively promote the construction of logistics information platform [1].

Currently, there are still many areas of logistics structure, management and technical defects, reflected in the regional logistics platform architecture, and difficult to effectively build a platform or platform level is not high, difficult to support the efficient operation of modern logistics system.

Research on domestic and international logistics platform that will present scientific integration of modern information technology to logistics to the development of intelligent logistics system is the requirement of the times and the inevitable social progress.

48.2 Status of the Yangtze River Delta Logistics

Yangtze River Delta region is located in the industrial and densely populated coastal areas of China T-region, a collection of waterways of the Gold Coast and Gold. Currently, the Yangtze River Delta region economy has become China's economic development leader [2, 3]. According to new statistics show that the Yangtze River Delta region in 2008 twenty-five cities reached 6.518507 trillion yuan of GDP for the Country's gross domestic product, one-fifth.

Statistics show that only in the Yangtze River Delta by the State Council approval of a class of 35 in Hong Kong, including the coastline of the coastal port 19, inland waterway port 10, port 5 for the international, and international temporary railway crossings 1. Port cargo throughput in the total trade volume ranks with the country.

Now not only Shanghai, Hangzhou and Ningbo, the first-tier cities want to build a logistics center, and even third-tier cities are also a lot of the logistics center construction project. Preliminary statistics, in the Yangtze River Delta has plans to build up a logistics center, more than sixty, but in fact the Yangtze River Delta region is not large, so many of the logistics center logistics center in such a concentration sufficient to show a serious duplication of investment [4].

48.3 The requirements of the Logistics Information Platform

Yangtze River Delta logistics information platform of logistics management for the government to provide the following functions: Yangtze River Delta logistics

operation of basic data processing and integration of the Yangtze River Delta logistics resources support functions and Yangtze River Delta logistics analysis and planning support [5].

Logistics information platform for industrial and commercial enterprises the main function of the following requirements: Information logistics providers, Logistics business transaction management, and special and other value-added services.

Information needs from the perspective of third-party logistics companies Analysis, in addition to sharing information than it needs logistics information service platform capabilities to provide the following information.

Third-party logistics business logistics information platform requirements: Public logistics infrastructure resources, including roads, railways, waterways, ocean and air transport networks, storage networks, freight station, transportation, handling equipment, such as traffic flow conditions the use of logistics resources across regions; Logistics market information resources; Logistics business resources; Other logistics consulting services resources. Logistics Information Platform Requirements Analysis Chart showed in Fig. 48.1 [6].

48.4 Design logistics platform

There is network transmission equipment, access equipment, networking equipment, switching equipment, cabling systems, network operating systems, servers, network test equipment. Logistics Information Platform Network structure showed in Fig. 48.2.

Logistics Information Platform Yangtze River Delta region consists of the data layer, service layer, user layer showed in Fig. 48.3.

Data layer: Logistics Information Platform is a data layer for all kinds of logistics related data collection, organization, storage and maintenance.

Service layer: Logistics Information Platform services layer is based on the data layer reading of information on the logistics application platform, these platforms according to customer's different requirements at different times in different application systems development. Basic platform has: basic information service platform, logistics operations management information platform, logistics enterprise information management platform, supply and demand, trading information platform, etc.

User Level: Logistics information platform to the user level within the logistics base, the base of external logistics enterprises, social sectors and other relevant customer identity authentication system is the way to level in the system by providing different levels of standard programming interfaces and mechanisms for communication services, support and other systems of data exchange and function calls, the platform's scalability with a certain provision of the relevant security measures the system to provide different information services, and ensure the network, database and application security.

The following is the logistics information platform applications and functions of public showed in Table 48.1.

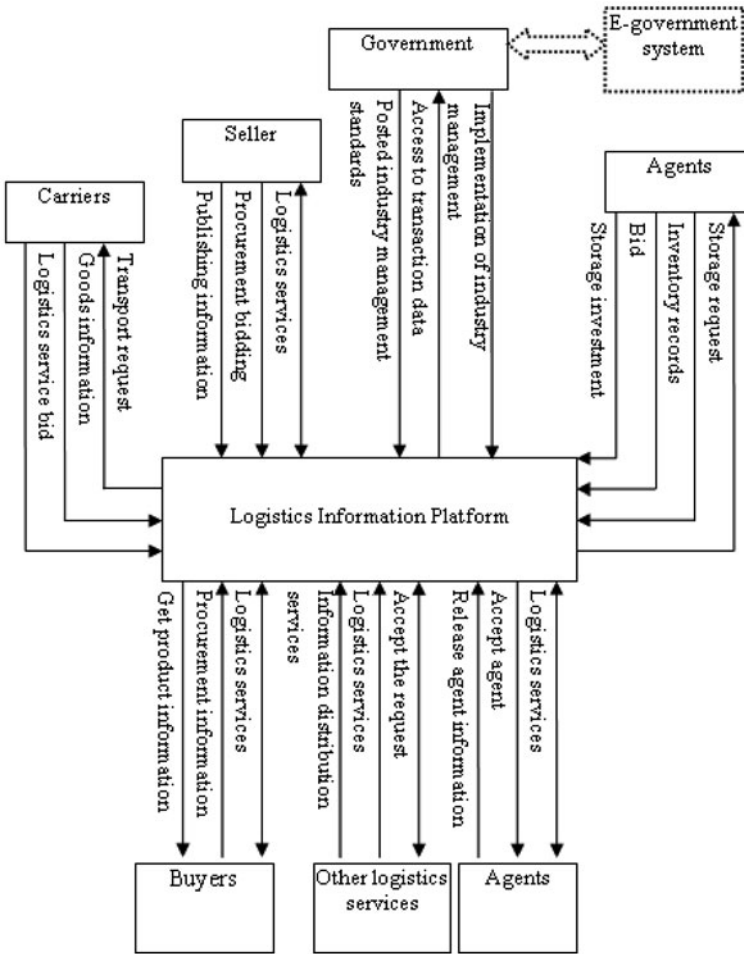


Fig. 48.1 Logistics Information Platform Requirements Analysis Chart

48.5 Yangtze River Delta Key Technology Logistics Information Platform

Yangtze River Delta region for the realization of the logistics information platform design features can be the following IT technology and logistics information management technologies.

Automatic data acquisition and storage: organize and store all types of information will apply computer database technology, data mining and mass data storage and management technology.

Maintenance of data and system security: Yangtze River Delta logistics information platform of the program interface will use the password encryption,

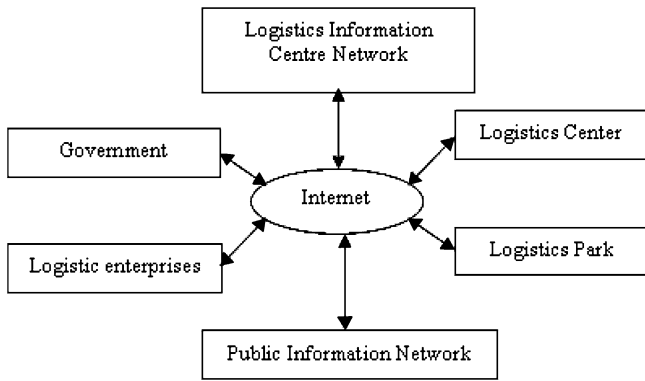


Fig. 48.2 Logistics Information Platform Requirements Analysis Chart

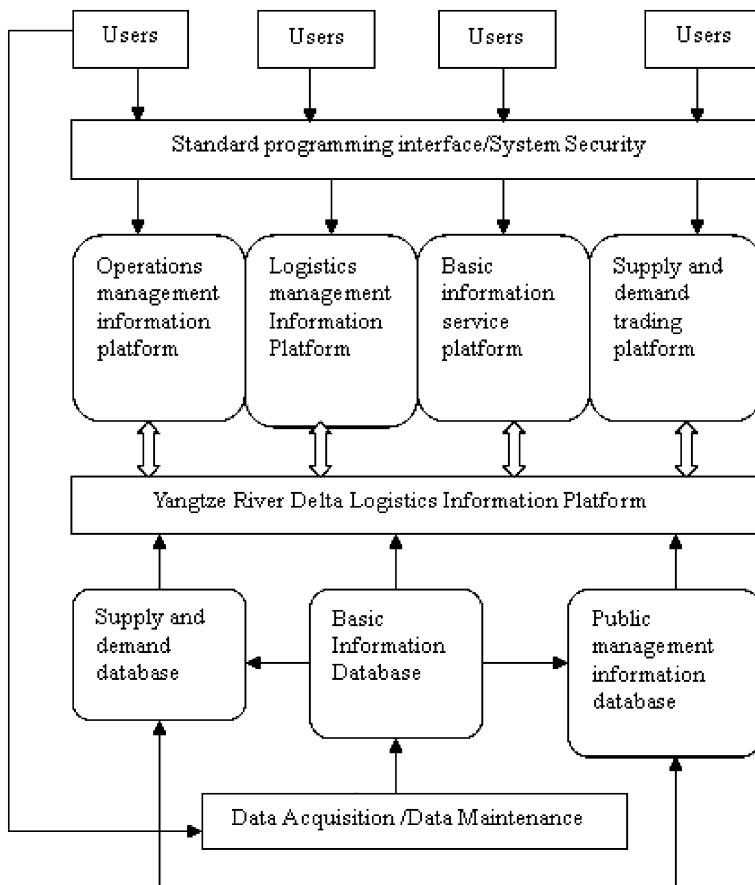


Fig. 48.3 Triangle diagram of the overall framework of Logistics Information Platform

Table 1 Logistics information platform applications and functions of public

Application system name	Function
Logistics transaction management system	Qualification examination and verification, notarization, and breach of contract logistics processing
Logistics EDI, etc. Information interface system	Logistics organization EDI system with other departments, automatic translation into EDI systems, electronic instruments such as transit Send
Inter-regional logistics Resource Integration System	Integrated logistics, regional economics, business entities and other resource database
Logistics market management system	Review of demand and supply logistics and other information
Public logistics information System	Logistics demand and supply information distribution, logistics, handling of complaints.
Decision Support System	Logistics development trend forecast.
Maintenance management system platform	Firewall, the system daily operation and maintenance, network detection.

key management, digital signature technology, electronic watermarking technology, firewall technology, platform for real-time data layer will use the database backup technology and the hot backup technologies.

Data communication and interaction: Yangtze River Delta logistics information platform needs a variety of communications technology and network support, such as the PSDN, DDN, ISDN, and WAN, LAN and VAN and so on. Through these networks to complete the EDI communication, application of CORBA technology, open EDI technology and Internet EDI technologies to meet the logistics industry in Yangtze River Delta, information sharing and information exchange requirements.

Information Standardization: all information is subject to Logistics Information Classification and Coding System and EDI-related code standards.

Online trading: the technology is still at an exploratory stage in our country. The platforms of online transactions were to be phased implementation.

Decision-making and management of logistics: to be integrated application of management information systems technology, and business process reengineering techniques, enterprise resource planning technology, computer decision support system technology, business intelligence technology, optimized management technology.

48.6 Summary

China's economic opening-up pace and the acceleration of global economic integration will also further accelerate the integration process of the Yangtze River Delta, Yangtze River Delta to promote the economic structure and industrial structure adjustment and upgrading. I believe the construction of the Yangtze River Delta logistics information platform will greatly benefit the economic development of Yangtze River Delta region.

References

1. Mingrong T (2008) City Logistics Public Information Platform of government-led development model [J]. *Logistics Technology* (9):0132–0137
2. Wang N (2007) Regional intelligent logistics platform [D]. Shanghai Marit University, China
3. Zhang H (2008) Regional logistics information platform construction [D]. Tongji University, Shanghai
4. Liliang, Cao (2010), Yangtze River Delta region and the status of the logistics industry development [J]. *China Bus* (8):43–48
5. Shuang, Zhao. (2010), Regional logistics information platform planning and construction [J]. *Intelligence J* (8):58–64
6. Yi,Xu., Youwang, Sun,(2007), Yangtze river delta logistics information platform for [J]. *Technol Econ* (1):011–016

Chapter 49

The Case Study for Three Kinds of Mobile Games

Quanyin Zhu, Suqun Cao, Rui Geng and Chuanchun Yu

Abstract Happy farm, Plants versus Zombies, and Gallant fighter with Double Blade—three games structures, the file hierarchies, the main game classes design and methods to develop the mobile game online are illustrated. All of them are widely welcomed recently. Game API in the Mobile Information Device Profile (MIDP) 2.0 is used to build the game engine for Happy farm online, the Android OS 2.1 and 2.2 are used to build the games engine for Plants vs Zombies and Gallant fighter with Double Blade respectively. The components designs and implementation steps of the games are introduced in detail. Various techniques, such as object pool, multi-threaded, socket connection, Maps etc., are applied in game's development. Experiments demonstrate its performance and proves that these cases are meaningful and useful to develop other online mobile games.

Keywords Case study · Happy farm · Plants versus Zombies · Gallant fighter with double blade · Gravity sensing · Mobile game online

Q. Zhu (✉) · R. Geng · C. Yu
Faculty of Computer Engineering, Huaiyin Institute of Technology,
Huaiyin, China
e-mail: hyitzqy@gmail.com

R. Geng
e-mail: bsnh521@qq.com

C. Yu
e-mail: yuchun1986@163.com

S. Cao
Faculty of Mechanical Engineering, Huaiyin Institute of Technology,
Huaiyin, China
e-mail: caosun@126.com

49.1 Introduction

Internet on mobile terminals is developing at an astonishing speed, which indicates the mobile service in China possesses a tremendous market value in the future [1]. Online mobile games have been progressed very quickly recently, and some researchers focused on transforming the internet game to the online mobile game [2–4]. Mobile games of graphics class which are suitable for different people of all ages are widespread now. An Activity Theory (AT) model using modern digital technology, AT components and example are given in the Ref. [5]. A method for customization of all aspects of game (contents, graphics, game style, rules, victory conditions) with a wide possibility is researched in the Ref. [6]. Some interesting and useful ideas are experimented such as rating the descriptiveness of images in a system as well as some methods to improve position [7], estimating and detecting various hand gestures and postures of a user with a two-axis accelerometer [8], and a plausible physically based model for animating and rendering fire and smoke on a mobile platform [9]. Some applications of mobile phone games for the future experience are researched. For example, applying the human ability to control a video game on a mobile device, electroencephalographic (EEG) Mu rhythms are used which depends on the signals obtained using a specially designed electrode cap and equipment, and be sent through a Bluetooth connection to a PC which processes it in a real time [10, 11], using an input data from slide on–off and microphone interface on the mobile phone which is based on pop-up function of wipi platform and uses traditional materials [12]. A new methodology for location-based games that aims at capturing the players’ emotional reactions to the activities in a game whilst in certain locations in order to test the methodology, a location-based game that can be played on any Bluetooth-enabled mobile phone is designed in Ref. [11]. A study of continuous use of mobile services in different use contexts as defined by task and consumption place is research in the Ref. [13]. The Ref. [14] researched the relationships between mobile consumers’ value tendency and their perceptions of mobile Internet service quality in terms of three different mobile quality dimensions.

Online games of mobile phone are welcomed by everyone because it can be played every time and everywhere. The game Happy farm which is widely welcomed in China recently can be played by children and senior citizen. However, it only can be played on internet. The case study is a main approach which are very useful to develop the mobile phone games and was reported on some papers [2, 3].

We focused on the case study of mobile phone games. Three games of classes designing for Happy farm, Plants vs Zombies, and Gallant fighter with Double Blade are studied in this paper.

49.2 Game Class for Happy Farm Online

49.2.1 Class Design

The process of the game development is based on the idea of object-oriented, so that the function of each module should be as independently as possible. There are a total of 8 classes designed for the game depend on Java game class:

- (1) **GameMIDlet.** GameMIDlet Class inherits from javax.microedition.midlet.MIDlet, it should implements three means at least that are startApp(), pauseApp() and destroyApp(). The MIDlet will be on the start which is transferred from startApp() through Java Application Manager (JAM), that means the game can be start. The MIDlet will be paused when JAM implements pauseApp(), and make it to wind up by using the destroyApp().

GameMIDlet is a kind of entry class for game development. It can build a new object employed hereafter for each game class. Some new methods commonly used to facilitate other classes for called. For example, public Image loadImage (String path), by referring to the method, and others can easily import images.

- (2) **GameMenuCanvas.** GameMenuCanvas class inherits from javax.microedition.lcdui.Canva, and it can achieve the interface of java.lang.Runnable this is mainly responsible for the handling of the menu interface and buttons. It affirms that a variable in which GameStatus used to determine the state of the game, when the game is set up state, the display settings screen; and when the game is to help state, display help screen.
- (3) **GameRegister.** GameRegister class carry out the interface of javax.microedition.lcdui.CommandListener and javax.microedition.lcdui.Runnable, they login and registration for user interface processing by getting user input relevant information to the user's login or register.
- (4) **Connection.** Connection class accomplishes the communication between the server and the client. The server has two for each client connection object to complete the data to send and receive. Each client will be automatically activated after start-up connection and establish the conversation with the server.
- (5) **GameRMS.** GameRMS class which encapsulates the method to complete the relevant function will be implement when each user logs need to be storage and retrieval information in the user's mobile phone.
- (6) **MainMenu.** MainCanvas class inherits from javax.microedition.lcdui.Canvas. It is responsible for the main game interface display.
- (7) **ThreadCrop.** ThreadCrop class realize the interface of javax.microedition.lcdui.Runnable. It is primarily responsible for the growth of vegetables.
- (8) **TransformStr.** TransformStr class encapsulates methods to achieve a number of uncertainties related to the length of characters in vertical display which

Fig. 49.1 The methods used in the class of Main Menu

```

● C MainMenu(Display, GameMIDlet)
● initMainMenu(Graphics)
● clean()
● drawGameBack(Graphics)
● drawShop(Graphics)
● drawStorage(Graphics)
● drawWarning(Graphics)
● drawAddone(Graphics)
● drawMyseed(Graphics)
● paint(Graphics)
● keyPressed(int)

```

does this by passing a string and brush; you can return a string with a vertical transparent image.

49.2.2 Game Frameworks

The framework of the game is divided into three kinds of states: Wait for the status, operation status (running state is divided into multiple sub-states), the end of the state, they run the game in the whole world. The game structure is based on the state machine to run the game a variety of different forms of run-time is divided into a one state, at any time there will be a state to be executed. Taking into account the game state nested problems, all of the state of each area separated from the main thread in the game constantly judgments in order to achieve the transition between states.

49.2.3 Game Implementations

(1) The Methods of the MainMenu Class. The game main menu is processed by Main Menu Class. The methods are used in the class, as shown in Fig. 49.1.

MainMenu (Display, GameMIDlet) is a kind of constructor which is used to pass the entrance classes and brushes come in convenient for later on, and some variables are also initialized in the constructor. In it MainMenu (Graphics) is used to initialize the main interface, through which the method call paint (Graphics) method to draw the game main menu, and draw GameBack (Graphics) method draw the game's background using paint (Graphics) method. DrawShop (Graphics) method is used to painting shop. DrawStorage (Graphics) is used to the painting warehouse. DrawAddone (Graphics) for the painting stores to buy goods +1

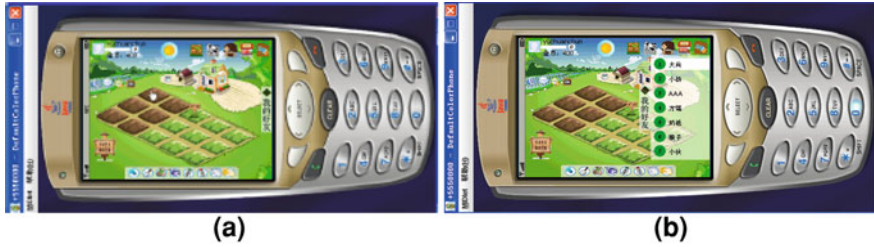


Fig. 49.2 Land and friends widows. **a** Land. **b** Friends Widow



Fig. 49.3 Warehouse, planting and harvest; **a** Warehouse, **b** Planting **c** Harvest

dynamic effects. DrawMyseed (Grahpics) method is used to painting whom purchased the seeds. KeyPressed (int) method is used to key processing.

(2) Running Game. Five static constants (LAND, MESSAGE, MENU, TOOL, and FRIEND) are defined in the MainMenu class which is used to identify the selected state. When the players enter the game from the login window of the main interface, the initial state (SelectStatus) is land (LAND) which can also be selected by press ‘3’ key. Press ‘1’ key for information widow, press ‘9’ key for menus widow, press ‘7’ key for tools widow, and press ‘0’ key for friends widow. The land and friends widows are shown as Fig. 49.2.

In the menu bar, players can buy in a store seeds and the start is 400 gold coins. Once purchased, you can buy in the tool bar to select the seed planted in the land. The vegetables can be a single removal or remove all by the player after mature. Players can use the tool bar, shovel to eradicate leaves when removal is completed. In the warehouse will be shown the fruits after removal is completed, you can sell the fruits of your own. Further more, player’s experience and gold has also increased. Warehouse, planting and harvest are shown as Fig. 49.3.

All manuscripts must be in English. Please keep a second copy of your manuscript in your office (just in case anything gets lost in the mail). When receiving the manuscript, we assume that the corresponding authors grant us the copyright to use the manuscript for the book or journal in question. Should authors use tables or figures from other Publications, they must ask the corresponding publishers to grant them the right to publish this material in their paper.

Use italic for emphasizing a word or phrase. Do not use boldface typing or capital letters except for section headings (cf. remarks on section headings, below). Use a laser printer, not a matrix dot printer.

49.3 Game Class for Plants vs Zombies

49.3.1 Class Design

The classes of the Plants vs Zombies have a total of 10 classes designed for the game depend on Android 2.1.

- (1) **MainGameView**. **MainGameView** class inherits from **SurfaceView**. it includes 33 instantiation object methods.

`initList()` initial the List or arrayies; `initImages()` initial the pictures; `initData()` initial the data; `initObject()` initial the objects which include the sun value of capturing plant cropping and the backfilling instance; `onTouchEvent()` is used to deal with the touch screen events; `judgeSelectWhere()` judges the location of planting selection; `GameRunThread()` is used to startup the thread class external; `Update()` is used to update the data; `doDraw()` is used to draw the object; `updateBg()` is used to update the background; `updatePlantFence()` is used to update the plant fence; `drawPlantFence()` is used to draw the plant fence; `updateSun()` is used to update the information of current sun; `drawWordWorn()` is used to show the attention words; `drawSun()` is used to draw the information of current sun; `drawSunCount()` is used to show the total number of current sun; `updateFlower()` is used to update the flower; `drawFlower()` is used to draw the flower; `updatePea()` is used to update the pea gunmen; `updateBullet()` is used to update the bullet; `drawBullet()` is used to draw the bullet; `updateNut()` is used to update the nut; `drawNut()` is used to draw the nut; `updateZombies()` is used to update the live Zombies; `drawZombiesOnPreview()` is used to show the preview Zombies; `drawZombiesOnGame()` is used to draw the Zombies in the gaming; `updateZombiesDie()` is used to update the processing of Zombies; `drawZombiesDie()` is used to draw the processing of Zombies; `isClogInRow()` is used to judge plants bar for next step of Zombies current line, it will to fight the plants bar when the next step is blocked off, and the plants which hold the fighting skills will to retaliate upon; `hitZombies()` is used to judge the bullet whether or not strike down the antagonist; `updateGameOver()` is used to flash the game over windows; `drawGameOver()` is used to show the game over interface.

- (2) **WelcomeView** and **YouFail**. **WelcomeView** and **YouFail** class inherit from **SurfaceView**. **WelcomeView** class is used to show the welcome interface, and the **YouFail** class is used to show the fail interface. Two classes are shown as Fig. 49.4.

Fig. 49.4 The WelcomeView and YouFail class. **a** WelcomeView class. **b** YouFail class

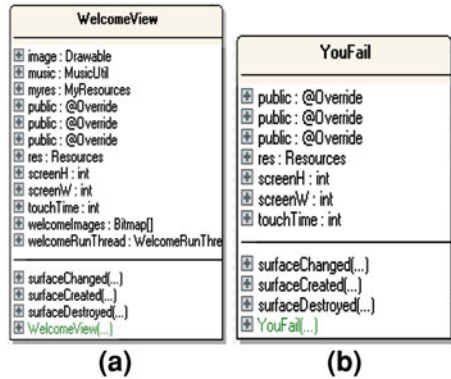
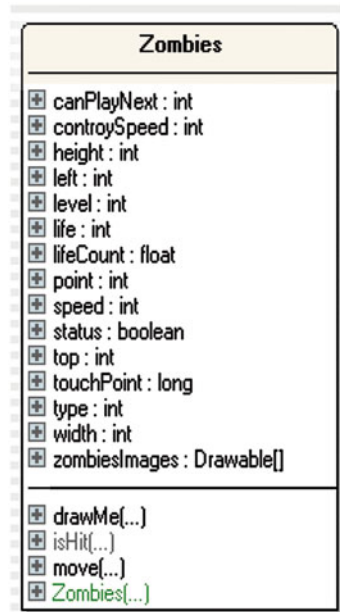


Fig. 49.5 The Zombies class



- (3) Plant. Plant classes include PlantFlower, PlantNut, and PlantPea class. PlantFlower class is used to draw the plants and produce the suns. PlantNut class is used to draw the nuts. PlantPea class is used to draw the peas.
- 4) Zombies. Zombies class is mainly used to define the wide, high, position (x, y), being, and type etc. which is shown as Fig. 49.5.
- (5) MyResources. MyResources class is a kind of index class. It is used to draw and setup the property pictures of plants property list, start flash and attention words. The MyResources class is shown as Fig. 49.6.

Fig. 49.6 The MyResources class

```

MyResources
+ { : <Integer>
+ { : getBgImagesIndex.add
+ { : getBgImagesIndex.add
+ { : getBgImagesIndex
+ { : R.drawable.bg1
+ { : R.drawable.bg2
+ { : }
+ <Integer> : List
+ <Integer> : List
+ = : getBgImagesIndex
+ ArrayList : new
+ getBgImagesIndex : return
+ { : }

MyResources(...)

```

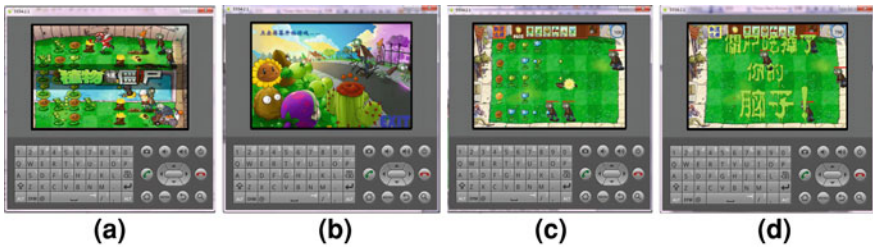


Fig. 49.7 The game windows. a Start flash window. b Welcome window. c Fighting window. d Fail window

(6) Const. Const class is used to define the all the constants of the games. It includes more then 45 constants which like int BGRECTSPANX, BGRECTSPANY, BLUEBULLET, GAMESTART, ZOMBIES_NET, ZOMBIES_BUCKET, ZOMBIES_STICK etc.

49.3.2 Game Implementations

sFive static constants (LAND, MESSAGE, MENU, TOOL, PLANTS and ZOMBIES) are defined in the MainMenu class which is used to identify the selected state. When the players enter the game from the login window of the main interface, the initial state (SelectStatus) is land (LAND) which can also be selected by press '3' key. Press '1' key for information widow, press '9' key for menus widow, press '7' key for tools widow, and press '0' key for plants widow. The start flash, welcome, fighting and fail widows are shown as Fig. 49.7a–d respectively.

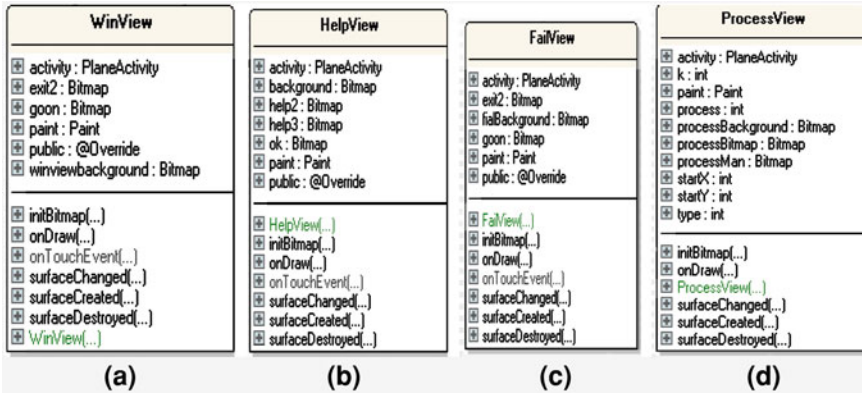


Fig. 49.8 The window classes. a WinView class. b HelpView class. c FailView class. d ProcessView class

49.4 Game Class for Gallant Fighter with Double Blade

49.4.1 Class Design

The classes of the Gallant fighter with Double Blade have a total of 15 classes designed for the game depend on Android 2.2.

- (1) PlaneActivity. PlaneActivity class inherits from Activity. It is the entrance class of the game, the other classes are managed when the game running.
- (2) WelcomeView, WinView, HelpView, FailView, and ProcessView. All the classes above are inherited from SurfaceView. The WelcomeView class is used to show the welcome window, WinView class is used to show the win window, HelpView class is used to show the help window, FailView class is used to show the fail window, and ProcessView class is used to show the processing bar window. The WinView, HelpView, FailView, and ProcessView classes is shown as Fig. 49.8a–d respectively.
- (3) KeyThread, MoveThread, ExplodeThread, and GameViewBackGroundThread. KeyThread class is the listen thread class of keyboard. It listens the keyboard state on timing and gives corresponding process depend on the state listened. MoveThread class is the moving thread class of all moving contents except player airplane. ExplodeThread class is changing frame thread for explosion, it changes the explosion frame depend on the explodeList of GameView as well as possible. GameViewBackGroundThread class is background thread class for the rolling background and appearance contents. KeyThread, GameViewBackGroundThread, MoveThread, and ExplodeThread classes is shown as Fig. 49.9a–d respectively.

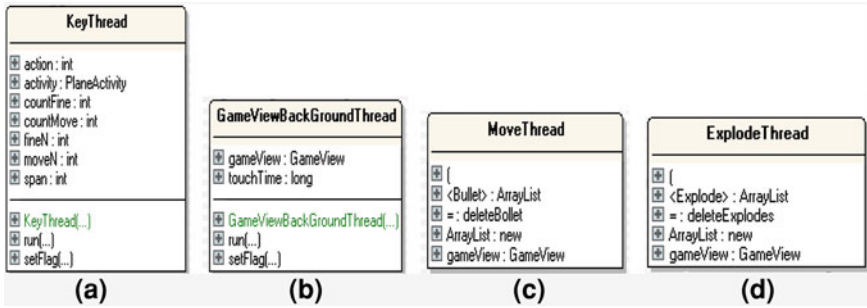
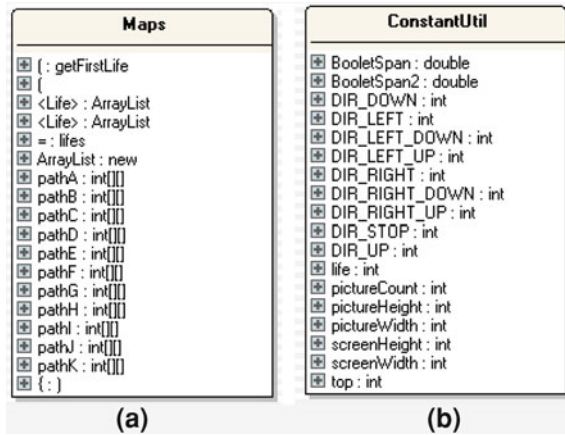


Fig. 49.9 The thread classes. **a** KeyThread class. **b** GameViewBackGroundThread class. **c** MoveThread class **d** ExplodeThread class

Fig. 49.10 Setting classes. **a** Maps class. **b** ConstantUtil class



- (4) Maps and ConstantUtil. Maps class sets the start point, goal point, the step order of current segment, path array, state, trigger time, type, life etc. ConstantUtil class sets all the constants for GameView class, GameViewBackGroundThread class and the direction constants of the game such as 0 represents rest, 1–8 represents up, upper-right, right, right-down, down, left-down, left, upper-left respectively, etc. Maps and ConstantUtil class is shown as Fig. 49.10a and b respectively.
- (5) Explode, Bullet and Life. Explode class draws the explosion on the define the position, other threads transfer the nextFrame method to change the frame. Bullet class record the parameters and produce the bullet encapsulation, bullet moving can be achieved by transferring the move method used parameters. Life class define which the blood can be gotten.

Fig. 49.11 The coordinate system



49.4.2 Gravity Sensing

Gravity sensing: process whereby a cell, bodily structure, or organism (animal or plant) receives or detects a gravity stimulus. Gravity sensing plays an important role in the directional growth and development of an organism [15]. Gravity sensing is used widely on the intelligence mobile phone recently. Therefore, it can be used in the many mobile games and support very interesting for the players.

The Coordinate system of gravity sensing in the Android OS takes the upper-left of the mobile screen as the origin. The direct define is shown as Fig. 49.11.

Using the x, y and z values to get the trigonometric function value, and then the mobile phone moving state can be accurate measurement.

- (1) `SensorManager sensorMgr`. `SensorManager sensorMgr = (SensorManager) getSystemService (SENSOR_SERVICE)` can get a hardware controller. For example, `LocationManager` can determine the location, `AudioManager` can use the audio player.
- (2) `sensorMgr.getDefaultSensor(Sensor.TYPE_ALL)`. `Sensor sensor = sensorMgr.getDefaultSensor (Sensor. TYPE_ALL)` can get the x, y and z values of offset.
- (3) `SensorEventListener()`. `SensorEventListener lsn = new SensorEventListener()` is used to get the value changing. When a `SensorEvent` has been declared to listen, the changing value can be gotten from `TextView`, and these values are `float[]` array. That is x, y and z value respectively.
- (4) `sensorMgr.registerListener (lsn,sensor,SensorManager.SENSOR_DELAY_GAME)`. Three parameters are listen, sensing device, and delicacy respectively. `SENSOR_DELAY_FASTEST`, `SENSOR_DELAY_NORMAL` and `SENSOR_DELAY_UI` represents the fast, normal and slowly respectively.

49.4.3 Game Implementations

Four static constants (`START`, `VOID`, `HELP`, and `EXIT`) are defined in the `MainMenu` class which is used to identify the selected state. Six sub-static constants (`SKY`, `MESSAGE`, `MENU`, `TOOL`, `FLIGHT` and `CLOUD`) are defined in

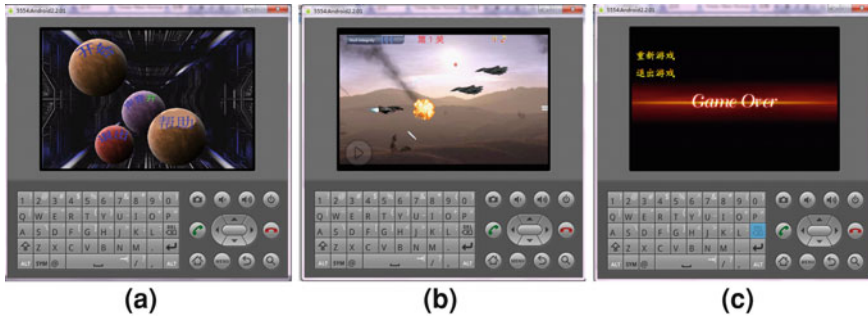


Fig. 49.12 The game windows. **a** Main menu window. **b** Fighting window. **c** Fail window

the ConstantUtil class which is used to identify the selected sub-state. When the players enter the game from the start window of the main interface, the initial state is Fighting window. The Main menu, fighting and fail widows are shown as Fig. 49.12a–c respectively.

49.5 Conclusions and Future Works

Along with the progress of 3 and 4G technology, the market for mobile phone games will have a huge prospect. Mobile phone games are welcoming for numerous mobile ISP. Happy farm, Plants vs Zombies and Gallant fighter with Double Blade are the currently popular games for mobile phone all has been attempted in this paper. JAVA-based games can be implemented on all intelligent mobile phones, and own a better portability.

This paper describes the Happy farm development which is based on MIDP2.0 platform while the Plants vs Zombies is implemented on the Android 2.1 and Gallant fighter With Double Blade is implemented on the Android 2.2. The last two games have been simulated on the Android SDK and AVD Manager of Eclipse 3.5 for Java. The game classes design and methods are given as well. The gravity sensing is adopted for the player implementing Gallant fighter with Double Blade convenient expressly. Due to the restrictions on the mobile processor performance, many technical issues such as game performance and network performance need to be optimized and researched.

References

1. Li J, Alexandra P, Sanxing C, Yajing C (2009) Trends on interactive platforms for social media through web2.0. MASS '09. International conference on 20–22 Sept 2009, pp 1–2
2. Zhu Q, Zhao L, Cao S, Shen J, Zhang S (2009) A BnB mobile game online based on J2ME and J2EE, software engineering research, management and applications. SERA '09. 7th ACIS international conference on 2–4 Dec 2009, pp 19–24

3. Zhu Q, Zhang H, Sun W (2008) Research of key technologies of mobile network games based on J2ME and J2EE. *Comput Eng Des* 29(20):5218–5221
4. Xu C (2008) A software framework for online mobile games. *International conference on computer science and software engineering, CSSE*. 1412, pp 558–561
5. Sedano CI, Botha A, Pawlowski JM (2008) A conceptual framework for ubiquitous mobile environments, Portable information devices. *7th IEEE conference on polymers and adhesives in microelectronics and photonics, 2nd IEEE international interdisciplinary conference on 17–20 Aug 2008*, pp 1–6
6. Mininel S, Vatta F, Gaion S, Ukovich W, Fanti MP (2009) A customizable game engine for mobile game-based learning, systems, man and cybernetics. *SMC 2009. IEEE international conference on 11–14 Oct 2009*, pp 2445–2450
7. Xiaolei H, Stalnacke M, Minde TB, Carlsson R, Larsson S (2009) A mobile game to collect and improve position of images, next generation mobile applications, services and technologies. *NGMAST '09. third international conference on 15–18 Sept 2009*, pp 70–73. Doi: [10.1109/NGMAST.2009.64](https://doi.org/10.1109/NGMAST.2009.64)
8. Baek J, Yun B (2008) A sequence-action recognition applying state machine for user interface, consumer electronics. *IEEE Trans* 54(2):719–726
9. Park D, Woo S, Jo M, Lee D (2008) An interactive fire animation on a mobile environment, multimedia and ubiquitous engineering. *MUE 2008, international conference on 4–26 April 2008*, pp 170–175
10. Pour PA, Gulrez T, AlZoubi O, Gargiulo G, Calvo RA, Brain-computer interface: next generation thought controlled distributed video game development platform, computational intelligence and games. *CIG '09, IEEE symposium on 15–18 Dec 2008*, pp 251–257
11. Baillie L, Morton L, David C, Moffat S, Uzor S (2011) Capturing the response of players to a location-based game. *Pers Ubiquit Comput* 15(1):13–24
12. Kim M, Park S, Song S (2009) Mobile interface for physical interactive games applied, Consumer electronics. *ISCE '09, IEEE 13th international symposium on 5–28 May 2009*, pp 1024–1028
13. Liang T-P, Yeh Y-H (2011) Effect of use contexts on the continuous use of mobile services: the case of mobile games. *Pers Ubiquit Comput* 15(2):187–196
14. Kim DJ, Hwang Y (2010) A study of mobile internet user's service quality perceptions from a user's utilitarian and hedonic value tendency perspectives. *Information systems frontiers, online first*TM, 8 Sept 2010
15. U.S. National Library of Medicine. *Definitions.net*. STANDS4 LLC, 26 March 2011. <http://www.definitions.net/definition/gravitysensing>

Chapter 50

New Methods of Specifying and Modeling Stochastic Concurrent Systems

Jingjing Liao, Mingzhe Wang and Fabin Guo

Abstract For the quantitative analysis of stochastic concurrent systems, specifying methods for the level of symbolic semantics and the level of concrete semantics with general distributions are introduced. First of all, Probabilistic Stochastic Automata is brought forward to achieve the symbolic semantics specification. Based on that, two concrete semantics specification methods are developed—Extended Colored Stochastic Petri Nets and Generalized Stochastic Extended Bundle Event Structures. Then after comparing modeling and performance evaluation characteristics between them, a new idea of analyzing stochastic concurrent systems combining the two methods on the basis of system logic structures and event sequences is formed.

Keywords Stochastic concurrent systems · Probabilistic stochastic automata · Extended colored stochastic petri nets · Generalized stochastic extended bundle event structures

J. Liao (✉) · M. Wang · F. Guo

Department of Control Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, People's Republic of China

e-mail: ljj0331@163.com

M. Wang

e-mail: mzwang@netease.com

F. Guo

e-mail: guofabin@163.com

50.1 Introduction

The importance of time and probability aspects in the modeling and analysis of stochastic concurrent systems has been widely researched [1]. The random duration of an activity is represented by probability distribution functions, and an important special duration distribution is the exponential distribution. Due to its nice mathematical properties, the problem of specifying systems which only make use of exponential distribution is easier than the general case and has been successfully studied [2–4]. And the behavior of these systems can be thought to be Continuous Timed Markov Chains (CTMCs), which is restricted to Markovian approach. In this way, we can find that Markovian approach limits the time characteristic of systems. So it is important to study the modeling and analyzing systems with general distributions.

Most of the specification approaches has a two level structure [5]: the level of symbolic semantics and the level of concrete semantics. For the first level, we introduce a new method to represent timed systems with general distributions—Probabilistic Stochastic Automata (PSA), which includes clocks and probabilities. And for the second level, we show two new modeling methods—Extended Colored Stochastic Petri Nets (ECSPNs) and Generalized Stochastic Extended Bundle Event Structures (GSEBESs). The latter two methods both have the advantages of intuitionist graphs and expressing true concurrent semantics. Besides that, ECSPNs have strong modeling abilities and many good performance evaluation results, while GSEBESs describe systems by event/action sequences instead of time specification, which can be easily used to analyze system event structures and behavior tracks. At the end of this chapter, we compare their advantages and disadvantages in aspects of modeling and analysis.

50.2 Probabilistic Stochastic Automata

In this section, we introduce how to specify concurrent timed systems with general distributions formally. When the restriction to exponential distributions is abandoned, it becomes more difficult to model the behavior of the global system. In that case, the behavior execution will depend on the elapsed time of the corresponding activities. A simple way of representing the execution time of an activity is to adopt a model with clocks in a similar manner as timed automata [6]. A timed automation represents the behavior of a system in terms of a fixed set of clocks c_1, c_2, c_3, \dots . More precisely, labels of timed automata transitions are of three kinds: (1) actions, used to express the occurrence of events and to synchronize system components; (2) guards, expressing a condition on clocks; (3) clock setting events.

Definition 1 A probabilistic stochastic automata (PSA) is defined to be a structure $(St, Ck, Distr, Act, \rightarrow, s_0, C_0)$ where St is a countable set of control states with $s_0 \in St$ being the initial state; Ck is the set of clock names with $C_0 \subseteq Ck$ being the

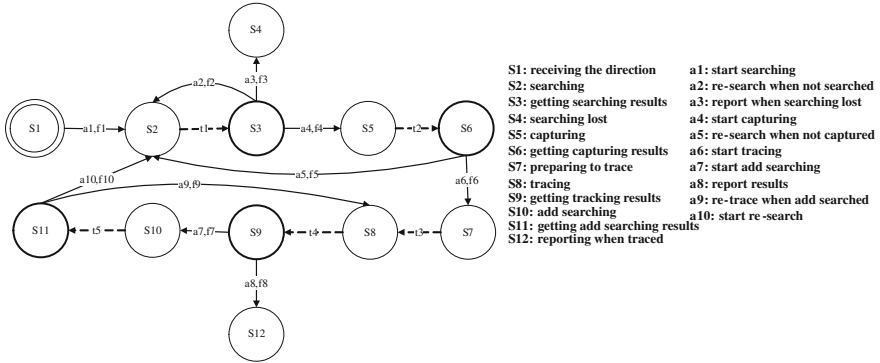


Fig. 50.1 The PSA model of the Warning Radar System

clocks to be started at the initialization; $Distr : Ck \rightarrow PDF$ assigns a probability distribution function to each clock. If $f \in PDF$, we usually name a clock $c_f \in Ck$ to indicate that $Distr(c_f) = f$; Act is the set of actions partitioned in the following sets: Act_d is the set of delayable actions and Act_u is the set of urgent actions, where Act_u is ordered according to the probabilities; $\rightarrow_{\subseteq} St \times 2^{Ck} \times Act \times Prob_d(2^{Ck} \times St)$ is the control transition relation.

We use clocks $c_i(i = 1, 2, \dots)$ to represent the non-negative time variables with associated probabilistic distributions of $f_i(i = 1, 2, \dots)$, as in Fig. 50.1 expressing a PSA. When the transition probability is 1 or the time variable is 0, we can omit them.

PSA is a kind of symbolic semantic specification method of concurrent systems. There have been a lot of research results about its verification, but little on its performance evaluation has been studied. Therefore, we need to develop concrete semantic specification methods, two of which we will introduce in the next two sections-Extended Colored Stochastic Petri Nets (ECSPNs) and Generalized Stochastic Extended Bundle Event Structures (GSEBESs).

50.3 Extended Colored Stochastic Petri Nets

50.3.1 Basic Concepts of ECSPNs

The limitation for Ordinary Petri Nets (OPNs) to projects is the huge and complex modeling scale, which is improved by the Colored Petri Nets (CPNs) [7] created by Professor Kurt Jensen. CPNs enfold the graphs according to classifications and add definitions of data types based on OPNs. For the aim of hierarchical modeling and dynamic evaluation of systems, Hierarchical Colored Petri Nets (HCPNs) and Timed Colored Petri Nets (TCPNs) have been developed.

Stochastic Petri Nets (SPNs) add a time value associated with a probabilistic distribution for every transition between enabled and firing, where the probabilistic

distribution is the exponential distribution. Although the restriction of the exponential distribution makes the SPNs isomorphic to Markov Chains (MCs), it is not up to the real work projects where the time delays are not always associated to the exponential distribution. Then Non-Markov Stochastic Petri Nets are brought forward, where the random time delays of transitions can be with general probabilistic distribution or some certain values. But they can not resolve the problem of the Choice Free (CF) structures. So the thought of adding transitions priorities is developed. In this chapter, we take probabilities instead of priorities to represent the conflict structures. And there still exists state space explosion for Non-Markov Stochastic Petri Nets, which makes us try to add the unfolding method based on them. In the following we will introduce some new concepts of ECSPNs.

Definition 2 An ECSPNs is defined to be a structure of $(\Sigma, P, T, A, N, C, G, E, I, \text{TIME})$, where Σ is a countable set of colors; P is a countable set of places; T is a countable set of transitions; A is a countable set of arcs such that $P \cap A = P \cap T = T \cap A = \Phi$; N is the node function defined as $A \rightarrow P \times T \cup T \times P$; C is the color function defined as $P \rightarrow \Sigma$; G is the guard function defined as $T \rightarrow \text{exp } r$, such that $\forall t \in T, [\text{Type}(G(t)) = B \wedge \text{Type}(\text{Var } G(t) \subseteq \Sigma)]$; E is the arc expression function defined as $A \rightarrow \text{exp } r$ such that $\forall a \in A, [\text{Type}(E(a)) = C(p(a))ms \wedge \text{Type}(\text{Var}(E(a))) \subseteq \Sigma]$, where $p(a)$ is the place of $N(a)$; I is the initialized function defined as $P \rightarrow \text{exp } r, \forall p \in P, [\text{Type}(I(p)) = C(p)ms]$; Time is the time function defined as $T \rightarrow \text{exp } r$ such that $\forall t \in T, [\text{TIME}(t) = \text{Stochastic Distributed Variable} \wedge \text{Const}]$, where *Stochastic Distributed Variable* expresses a random variable, *Const* expresses a nonnegative real number.

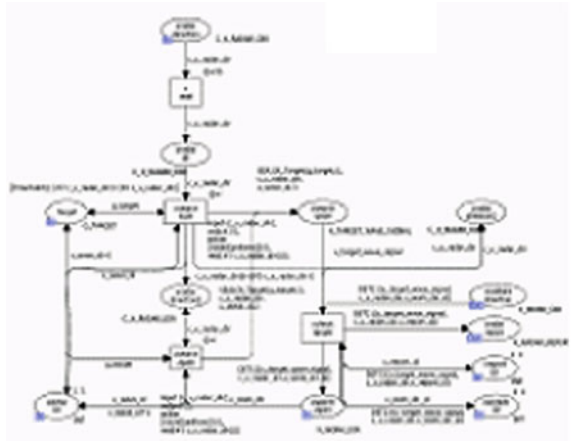
50.3.2 Modeling Methods of ECSPNs

Like CPNs, the modeling process of ECSPNs includes the definitions of colors, timed transitions, arcs and guard functions. The places represent states of the systems and the transitions represented the events or activities. Fig. 50.2 is an example of ECSPNs, where the ellipses represent places with the associated color sets labelled on the side, the rectangles represent transitions, and the lines with arrowheads connecting places and transitions represent arcs, the expressions on which are arc functions. A probabilistic choice of transitions could be expressed as arc functions, and the time functions which could be constants or variables associated with general probabilistic distributions are written behind the label @.

50.3.3 Analysis Methods of ECSPNs

For ECSPNs, we can analyze them using the same way of CPNs, which includes three aspects: structural analysis, state space analysis and simulation-based analysis.

Fig. 50.2 The ECSPNs models of the Warning Radar System searching



Structural analysis has the notable advantage of independence of the system's initial states, from which we can get the whole system logic verification results from the essential structure and capture the behavior characteristics. We can use the equivalent OPNs structural analysis methods for references to validate the systems' behavior. By computing S/T invariants, we can validate the information conservation structures and cycle event structures.

The state space based analysis pays attention to the behavior sequences and the constraint relations. The first ones can be represented the event flows of the execute models and the latter ones can be represented the relations of events (e.g. causality, conflict and so on). We can get the event sequences through computing the process nets of ECSPNs models. On the other side, the state transition graphs of ECSPNs models can help us specify all states for a certain initial marking. Besides that, we can use ASK-CTL to validate the behaviors of ECSPNs models.

We can not use Markovian methods to analyze and evaluate the systems' performance for abandoning the exponential distributions in ECSPNs models. As a result, we can only simulate the models to get some results based on Monte-Carlo. We can consider the following performance metrics: (1) the average time to arrive a certain state; (2) the average time for the first time to arrive a certain state; (3) the probability in a certain state before some time t .

50.4 Generalized Stochastic Extended Bundle Event Structures

50.4.1 Basic Concepts of GSEBEs

Event structures are a prominent noninterleaving model for concurrency. Bundle event structures consist of events labeled with actions—an event modeling the occurrence of its action—together with relations of causality and conflict between

events. It is difficult to simulate a system's behavior only depending on the sequences of activities. For concurrent systems, the performance and security analysis is a more important problem. So it is necessary to consider some performance evaluation parameters (e.g. time metrics and probability metrics). In the following, we will obtain Stochastic Extended Bundle Event Structures by decorating events and bundles with random time and probabilities.

In Stochastic Extended Bundle Event Structures, we bind probabilities to some events, which we denote probabilistic events. And we associate a probability distribution time value with some events, which are random timed events. Then we can obtain Generalized Stochastic Extended Bundle Event Structures.

Definition 3 A Generalized Stochastic Extended Bundle Event Structure ε is defined to be a structure of $\varepsilon = (E, \#, \mapsto, l, \pi, T)$ with E a set of events, $\# \in E \times E$ the (irreflexive and symmetric) conflict relation, $\mapsto \subseteq P(E) \times E$ the causality relation, and $l: E \rightarrow L$ the action-labelling function, $\pi: E \rightarrow_p (0, 1)$ the probability function, \rightarrow_p the partial order function, and $T: E \rightarrow PDF$ the time function. For each event e , there exists $e \in dom(\pi)$, $\exists Q \subseteq dom(\pi) : e \in Q \wedge Q$ such that $\bigwedge \sum_{e' \in Q} \pi(e') = 1$.

50.4.2 Modeling Methods of GSEBESs

We use GSEBESs extended with time and probabilities to simulate the behavior of systems. Stochastic Process Algebra is developed on the basis of Time Process Algebra and Probabilistic Process Algebra. In Stochastic Process Algebra, we add a probabilistic operator and a delay operator, where the delay is associated with a probabilistic distribution. $+_p$ denotes the probabilistic operator, and $P_1 +_p P_2$ expresses the probability of P_1 is p and the probability of P_2 is $1 - p$. $C \mapsto P$ represents the delay operator expressing the process P is enabled when the clocks in C are all terminated, where the clock variables in C are constants or associated with probabilistic distributions.

In GSEBESs, the semantics of the probabilistic operator $E_1 +_p E_2$ is equivalent to $E_1 + E_2$. Figure 50.3 is a GSEBESs example, where p_1, \dots, p_5 are probabilities of activities, $[0.378*4, Tpr + 0.378*4]$ represents the delay of the activity P_Search is associated with a uniform distribution from $0.378*4$ to $(pr + 0.378*4)$, and the delay time of other activities are constants or 0 (omitted) labelled beside the event names.

50.4.3 Analysis Methods of GSEBESs

There exist a lot of methods for performance analysis of EBESs with time and probability parameters, such as PEPA [8], TIPP [9] and MPA [10] etc. But most of

Fig. 50.3 The GSEBES model of the Warning Radar System



these are based on the stochastic concurrent systems with exponential distributions. And for systems with general distributions, few analysis methods have been developed.

In fact, event tracks of GSEBESs are consistent with ones of EBESs, which makes stochastic π -calculus [11] is available. On the other way, we can construct the Event-Attribute Matrices of GSEBESs to decide the event ability ranks to analyze the logic structures of the whole system, which needs us to study in the future.

50.5 Comparing Between the Two Methods

States and events are two dimensions of depicting systems’ behavior. The formal methods based on states (e.g. Petri nets etc.) model systems’ behavior through transitions of states. On the other side, the formal methods based on events (e.g. kinds of Process Algebra) specify sequences of events to depict systems’ behavior. ECSPNs is on the basis of states and GSEBESs events, which two model systems’ behavior form two different dimensions and can be synthesized to be a new analysis method.

ECSPNs model systems based on states, which results in state space explosion. And GSEBESs can avoid this problem based due to the basis of events, but we can not get the delays and probabilities of events through simulations of GSEBESs. So from simulation tools of ECSPNs (e.g. CPN Tools) we could obtain the delays and probabilities and then do some analysis based on performance and logic structures.

50.6 Conclusions

To represent the stochastic systems with general distributions and conflict activities, we introduce a kind of symbol semantics level method—PSA, and two concrete semantics level methods—ECSPNs and GSEBESs. Through compare the characters of the latter two methods; we brought forward the method of combining

them to analyze the systems' structures. In the future, we will pay more attention to the studying of their semantics mapping and event structure analysis.

Acknowledgment This work was supported by a grant from the National Natural Science Foundation of China (No. 60874068).

References

1. Pozsgai P, Bertsche B (2007) A method and its application for the modeling and simulation of the operational reliability characteristics and costs of technical systems. *Qual Technol Quant Manag* 4:279–299
2. Hermanns H, Herzog U, Katoen J (2002) Process algebra for performance evaluation. *Theor Comput Sci* 274:43–87
3. Hermanns H, Herzog U, Mertsiotakis V (1998) Stochastic process algebras-between LOTOS and Markov chains. *Comput Netw ISDN Syst* 30:901–924
4. Gharbi N, Dutheillet C, Ioualalen M (2009) Colored stochastic Petri nets for modeling and analysis of multiclass retrial systems. *Math Comput Model* 49:1436–1448
5. Bravetti M, Pedro D'Argenio R (2004) Concepts, discussions and relations of stochastic process algebras with general distribution, validation of stochastic systems. *LNCS* 2925:44–88
6. Pedro R (1999) Algebras and automata for timed and stochastic systems. Ph.D thesis, Department of Computer Science, University of Twente
7. Jensen K (1992) Colored Petri nets basic concepts, analysis methods and practical use volume I: Basic concepts, 2nd edn. Springer, Berlin
8. Glabbeek R, Plotkin G (2009) Configuration structures, event structures and Petri nets. *Theor Comput Sci* 410:4111–4158
9. Gotz N, Herzog U, Rettelbach M (1993) TIPP-Introduction and application to protocol performance analysis, Technical Report, University of Erlangen
10. Buchholz P (1994) Markovian process algebra: composition and equivalence. In: Proceedings of process algebra and performance modeling workshop
11. Varacca D, Yoshida N (2007) Probabilistic Π -calculus and event structure. *Electron Notes Theor Comput Sci*

Part V
Digital Image Processing

Chapter 51

Digital Image Completion Techniques

Chao Huang, Huadong Hu, Chunxiao Liu and Caiyan Xie

Abstract Image completion is a hot research topic in the multi-disciplinary area of computer graphics, image and video processing, and computer vision. It provides a strong tool for the reuse of captured images and photos, and shows its extensive applications in cultural heritage protection, special visual effects, image and video editing, and virtual reality. As the existing survey papers are out of date, its recent developments are summarized in three parts. First, its technical background is described for readers who are not familiar with it. Then, a comprehensive survey of the state-of-the-art methods is made to guide readers that are interested. Finally, a vision for future work is sketched to help motivate its further progress.

Keywords Image completion · Inpainting · Repairing · Retouching · Object removal

C. Huang · H. Hu · C. Liu (✉) · C. Xie
College of Computer Science and Information Engineering, Zhejiang Gongshang
University, Hangzhou 310018, People's Republic of China
e-mail: cxliu@mail.zjgsu.edu.cn

C. Huang
e-mail: yymm008@qq.com

H. Hu
e-mail: 916141945@qq.com

C. Xie
e-mail: nobodyelse@163.com

51.1 Introduction

Most of us treasure some old images and videos, such as photos of some significant moment or videos for some important activities. With time lapses, they become the only one that can remind us of those past things. However, when we examine them closely, we can see such disagreeable defects as the unwanted people or the left scratches. It's better to eliminate these ugly defects and recover the occluded content, which is called image completion [1]. It means removing the unwanted scene or restoring the damaged regions on the images to reveal its original appearance.

As early as the Renaissance, some of the medieval pictures are brought "up to date" by filling in the appeared gaps, which is called retouching or inpainting [2]. It is mainly executed by the professional artists, which is time-consuming, exhausting and subjective. Inspired by above manual work, the researchers from the computer graphics community try to solve this problem with graphics algorithms to recover the small scratches on the digital images of old paintings or photos [3]. Afterwards, this idea is quickly extended to remove the large disagreeable objects on the images [4], and shows more extensive future applications. Early traditional image completion methods are restricted to utilizing the known pixels in the target image to fill in the missing pixels, while some of the recent methods are extended to make use of other images or the large-scale image databases.

51.1.1 Problem Description

Image completion concerns the problem of removing the unwanted objects or filling in the missing regions on an image with the available information from the same image or another to generate visually plausible result [2–4], which preserves the continuity and consistency of the structure, texture and intensity information.

A typical image completion task includes following three procedures. First, the user provides and inputs the target image to be repaired and other images with useful information. The latter are also called as the source or reference images. Then, the user interactively specifies the damaged regions or the unwanted objects on the target image. Finally, the useful information from the known regions on the target image or other images is detected and copied to complete the missing regions, under the constraint of the known pixels along the boundary of the missing area.

51.1.2 Evaluation Criterion

As for image completion, quality other than speed is the major concern for algorithm evaluation [3, 4], because (1) most of its applications do not demand a

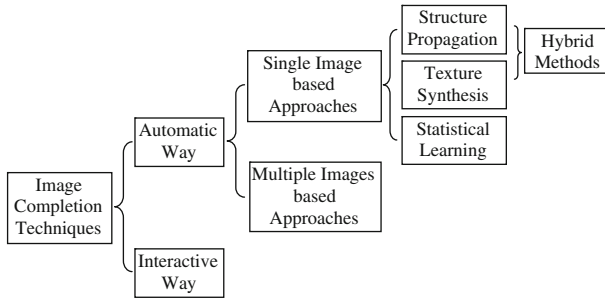


Fig. 51.1 Categories of image completion techniques

real-time response time; and (2) their quality expectations are far from being satisfied by the state-of-the-art methods. Most of the recent work focuses on the improvement of image completion quality, and only little of them touch upon its accelerating schemes.

There are many factors which influence its quality evaluation, and the evaluation criteria may include such visual characteristics as the continuity in texture, structure and intensity as well as the consistency in geometrical relationship (especially the perspective projection relationship) between the recovered visual information and the whole target image. However, they are too complex to be quantified, the current evaluation approaches for image repairing quality are as follows: (1) User Studies: to provide abundant image completion results generated from state-of-the-art methods for users, record the subjective perception from the users with careful investigation, treat most of users' opinions as the quality measure with statistics. (2) Ground Truth: to treat the images captured or generated by the users as the standard solution, damage them manually in various ways, compare the ground truth with the image completion results of the damaged images to get the repairing errors.

51.2 Categories of Image Completion Techniques

Since Bertalmio et al. [5] presented their work on image inpainting in SIGGRAPH' 2000, many methods have been brought forward and some of them become the classical ones. Because its rapid developments in recent years bring the existing survey papers out of date [3, 4], we reclassify them with new categories as Fig. 51.1 and introduce the most advanced research work as comprehensive as possible.

According to whether require the users' help except the standard repairing procedures, they can be divided into two categories, i.e. automatic or interactive way. According to whether need more images other than the target image in the repairing procedures, they can be classified as single image or multiple images

based approaches. Nevertheless, most of them are single image based automatic approaches and can roughly be classified as structure propagation, texture synthesis, and statistical learning based methods as well as hybrid ones. We give a detailed description for each approach as well as their characteristics in the following sections.

51.3 Single Image Based Approaches

Single image based approaches utilize only the known information in the target image to fill in its missing regions. Although most of them run in an automatic way, the interactive algorithms are also proposed mainly for them, we review them all in the following subsections.

51.3.1 Structure Propagation Based Methods

Structure propagation based methods regard image completion as solving partial differential equations (PDEs) [5–7] or variation problems [8–10] by specifying the known pixels around the damaged regions as boundary conditions. The repairing process is therefore the diffusion of the color information from the surrounding known pixels into the missing regions along the important image structures.

Bertalmio et al. propose to estimate the local color smoothness for each color channels with 2D Laplacian method, and propagate the color information outside the isophote for the boundary of the missing region to the inner missing pixels along the normal directions of the contour in an anisotropic way [5], which guarantees the edge continuity in the target image. Curvature-driven diffusion [6] and AM-FM reaction diffusion [7] approaches are also put forward to reconstruct the damaged directional texture image. The anisotropic diffusion process adaptively preserves the distinct edge features and restores the missing texture reliably.

Total Variation model based algorithm [8, 9] can satisfy the seamless transition between the repaired target regions and their surrounding original known pixels, which is based on the constraints of Euler–Lagrange equations and curvature-driven diffusion process. And, it is applied to recover the missing wavelet coefficients during the image transmission, i.e. image completion in the wavelet domain. Dobrosotskaya et al. [10] construct a new variation model for the blind restoration and completion of damaged images based on the local characteristics of the wavelet filter.

In brief, such methods retrieve the color information in the missing regions by re-sampling the color distribution function fitted and approximated under the assumption of color smoothness and continuity. They only work well for small

scale scratches, but produce blurring artifacts for large missing regions and fail for large and highly textured damaged area.

51.3.2 Texture Synthesis Based Methods

Texture synthesis based methods regard the known regions on the target image as texture swatches, then perform example based texture synthesis to generate new image fragments for the missing regions [11–14].

Fragment based algorithm [11] samples the known image fragments, and chooses the best matched one with the known neighborhood of the missing pixels to patch it, iteratively approximate the entire damaged area. It is very time-consuming because of the large search space of the best fragment match in the whole known image region. Criminisi et al. [12] sort the boundary pixels of the missing region according to their repairing priorities which embody the importance of the structural information, and repair the missing pixel with the highest priority at first. Good image completion quality makes it a milestone, and some improvements on it are proposed. Tang et al. [13] decrease the search scope of the best matched image fragment to accelerate the algorithm, and improve the match criteria to avoid the error accumulation. Inspired by global optimization based texture synthesis, Komodakis et al. [14] put forward a new image completion framework based on global discrete optimization, which treats the problem as a labeling problem on the Markov random field.

In a word, such methods implicitly assume distinct texture characteristics on the whole image and similar texture patterns between the known and the missing regions; therefore they can produce satisfactory repairing results for relatively large textured region by copying small source fragments from known regions. Nevertheless, they can hardly recover the precise structural information in the large missing regions.

51.3.3 Hybrid Methods Based on Structure Propagation and Texture Synthesis

Considering that structure propagation methods are sensitive to structure information but hard to handle texture information, on the contrary texture synthesis methods are good at completing missing textures but incapable of controlling complex structures, hybrid methods aim at combining them together to form a strong one [15–18].

Image decomposition based algorithm [15] first decomposes the input target image into two subimages, i.e. structure and texture images; then restores them with structure propagation and texture synthesis methods respectively; finally returns the

reconstructed target image by adding back the two repaired subimages. In addition, the authors propose to analyze the surrounding known information to partition the damaged area into structural or textured regions and automatically fill in them with the most favorable algorithm, which is tested on the images with packet loss during image compression and wireless transmission [16]. After dividing the target image into geometry image and texture image, Grossauer's algorithm [17] reconstructs the geometry image with structure propagation and enhance its structure continuity with the anisotropic diffusion filter, then segments the reconstructed geometry image into several connected regions by a gradient controlled region growth algorithm, finally restores the texture information in the connected regions with texture synthesis to get the final result. Elad et al. [18] break the original image into linear combined texture and cartoon layers using sparse representation based image decomposition method, i.e. morphological component analysis (MCA).

In conclusion, separately handling the image structures or textures to avoid mutual interference, such methods can achieve good inpainting results for the target images with certain complexity. However, decomposition of the image into pure texture and structure parts is very hard in that there are not a precise concept and mathematical model for texture and structure in the image.

51.3.4 Statistical Learning Based Methods

Statistical learning based methods solve the image completion problem by learning the statistical characteristics of the whole image from the known area, which is then used for the estimation and inference of the blank area [19–23].

Jia et al. [19] adaptively extract N dimensional tensors from the color and texture information through texture segmentation, and vote for the missing pixels with the tensor voting algorithm. Levin et al. [20] obtain the global statistical color distribution based on the available part of the image by statistical learning, and infer the most probable repairing result by loopy belief propagation (LBP) algorithm. Expectation Maximization (EM) based method [21] treats the problem as the estimation of the damaged regions, thus adopts EM algorithm for Maximum Likelihood (ML) estimation based on sparse linear representation. Zhu et al. [22] propose a low-level learning based framework. Learning from natural image databases, they establish the compound correspondences between different scenes and their images as well as the relationship among the image fragments represented by Markov networks. The underlying scene structures are retrieved by solving the Maximum posterior probability estimation problem with BP algorithm under the constraint of image continuity. Field of Experts (FoE) [23] trains a high order Markov model capturing the statistical features of the natural scenes for the inference of the missing regions.

Such methods are usually limited to their extracted sparse statistical representations from the known regions, which cannot stand for the whole image partially damaged.

51.3.5 Interactive Way for Single Image Based Approaches

The above single image based automatic image completion approaches cannot work well for the large missing regions, especially for those with strong structures. The primary reason lies in that in some cases the known information is not sufficient for the image completion task, which makes it an ill-conditioned problem [2]. For the purpose of a good inpainting performance, interactive approaches are proposed to introduce user experienced knowledge into the repairing process [24–26].

Sun et al. [24] first ask the user to specify the important structural curves in the blank regions. Then, they guide the restoration of the structural information with structure propagation strategies, and reconstruct the remainder with texture synthesis algorithms. Interactively introducing the projective transformation information in the scene [25], Pavić et al. rectify the perspective distortions in the image in favor of a better image completion quality than the pure texture synthesis based methods. As a high-level interactive image editing tool, Patch-Match framework [26] proposes a user-guided fast image completion method based on the randomized correspondence algorithm under the constraint of intuitive labeling.

Under the guidance of user supported cues, better repairing results are generated by interactive image completion techniques. However, they demand tiresome user interaction for the natural scene with complex structures, which is unbearable for users.

51.4 Multiple Images Based Approaches

Multiple images based approaches [27–29] exploit the available information from more images than the target image to complete the blank area.

Image database based algorithm in [27] retrieves similar image regions with the known surroundings and patches up holes in the target image to form several visual satisfactory effects, therefore the scale and diversity of the image database have a great influence on the final repairing result.

However, the authenticity of the repairing result is more important than the sense of reality in some cases, which reflects the original scene and appears more vivid. To ensure a faithful restoration of the large missing regions, we have proposed a new image completion framework based on an additional large displacement view (LDV) image of the same scene [28, 29], which can be obtained by moving a consumer camera to keep away from the obstacles and make the previously occluded scene visible. The key challenges are how to convert the visible information on the LDV image to be useful and how to exploit them to repair the target image. Based on interactive segmentation of planar scenes [28] and the minimization of warped perspective distortions for the LDV image [29],

two approaches are brought forward and their experimental results show comparable effects with the ground truth.

In conclusion, multiple images based approaches are close to be a practical solution to image completion problem at the cost of multiple views captured on the scene or numerous pictures retrieved from the internet.

51.5 Future Work

As an ill-posed problem, it is hard for single image based approaches to get a satisfactory result in that the core challenge of structure prediction in the blank regions is far from being solved. In addition, the interactive demand put a heavy burden on the users. On the contrary, multiple images based approaches are close to be a solution, whereas more images are required as auxiliary information and its robustness in complex situations must be further investigated.

We believe that the efforts in the following scientific and technical problems will push the development of image completion techniques forward. The first one is the structure prediction problem in the single image based approaches, without which the algorithm is ambiguous about when and where to stop the repairing process and usually results in broken or illegal structures. The second one is the sufficient utilization of a priori knowledge for the guidance purpose to free the user from tedious and difficult interactive task. The third one is the determination of geometrical relationships and the correction of perspective distortions among the multiple views in multiple images based approaches. The fourth one is the automatic detection and localization of the disharmonious defects in the target images with a priori knowledge or supplied multiple views. At least, the user friendly interactive way must be provided. Look forward to its great advances in the near future.

Acknowledgments This project is supported by the National Natural Science Foundation of China under Grant No. 61003188, and Technology Plan Program of Zhejiang Province under Grant No. 2009C11034, and Zhejiang Provincial Natural Science Foundation of China under Grant No. Y1111159, No. Z1101243 and No. Z1111051.

References

1. Collis B, Kokaram A (2004) Filling in the gaps. *IEE Electron Syst Softw* 2(4):22–28
2. Shen JH (2003) Inpainting and the fundamental problem of image processing. *SIAM News* 36(5):1–4
3. Shih TK, Chang RC (2005) Digital inpainting—survey and multilayer image inpainting algorithms. In: *Proceedings of ICITA*, pp 15–24
4. Tauber Z, Li ZN, Drew MS (2007) Review and preview: disocclusion by inpainting for image-based rendering. *IEEE Trans on Syst Man, Cybernetics–Part C: Appl Rev* 37(4):527–540

5. Bertalmio M, Sapiro G, Caselles V, Ballester C (2000) Image inpainting. In: Proceedings of the ACM SIGGRAPH 2000. New Orleans, pp 417–424
6. Chan TF, Shen J (2001) Nontexture inpainting by curvature-driven diffusions. *J Vis Commun Image Represent* 12:436–449
7. Acton AT, Mukherjee DP, Havlicek JP, Bovik AC (2001) Oriented texture completion by AM-FM reaction-diffusion. *IEEE Trans Image Process* 10(6):885–896
8. Chan TF, Shen JH (2005) Variational image inpainting. *Commun Pure Appl Math* 58(5):579–619
9. Chan TF, Shen JH, Zhou HM (2006) Total variation wavelet inpainting. *J Math Imaging Vis* 25(1):107–125
10. Dobrosotskaya JA, Bertozzi AL (2008) A Wavelet-Laplace variational technique for image deconvolution and inpainting. *IEEE Trans Image Process* 17(5):657–663
11. Drori I, Cohen-Or D, Yeshurum H (2003) Fragment-based image completion. *ACM Trans Graph* 22(3):303–312
12. Criminisi A, Pérez P, Toyama K (2004) Region filling and object removal by exemplar-based image inpainting. *IEEE Trans Image Process* 13(9):1200–1212
13. Tang F, Ying YT, Wang J, Peng QS (2004) A Novel Texture Synthesis Based Algorithm for Object Removal in Photographs. In: Proceedings of the 9th Asian Computing Science Conf. LNCS 3321, Chiang Mai, 248–258
14. Komodakis N, Tziritas G (2007) Image completion using efficient belief propagation via priority scheduling and dynamic pruning. *IEEE Trans Image Process* 16(11):2649–2661
15. Bertalmio M, Vese L, Sapiro G, Osher S (2003) Simultaneous structure and texture image inpainting. *IEEE Trans Image Process* 12(8):882–889
16. Rane SD, Sapiro G, Bertalmio M (2003) Structure and texture filling-in of missing image blocks in wireless transmission and compression applications. *IEEE Trans Image Process* 12(3):296–303
17. Grossauer H (2004) A combined PDE and texture synthesis approach to inpainting. *Lecture Notes in Computer Science, ECCV'2004*, 3022, 214–224
18. Elad M, Starck JL, Querre P, Donoho DL (2005) Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). *Appl Comput Harmon Anal* 19:340–358
19. Jia JY, Tang CK (2003) Image repairing: robust image synthesis by adaptive ND tensor voting. In: Proceedings of the IEEE CVPR 1, pp 643–650
20. Levin A, Zomet A, Weiss Y (2003) Learning how to inpaint from global image statistics. In: Proceedings of the IEEE ICCV 1, pp 305–312
21. Fadili MJ, Starck JL (2005) EM Algorithm for sparse representation-based image inpainting. In: Proceedings of the IEEE ICIP. 2, pp 61–64
22. Zhu B, Li HD (2005) Image Completion from Low-Level Learning. In: Proceedings of the Digital Imaging Computing: Techniques and Applications (DICTA'2005)
23. Roth S, Black MJ (2005) Fields of experts: a framework for learning image priors. *Proceedings of the IEEE CVPR 2*, pp 860–867
24. Sun J, Lu Y, Jia JY, Shum HY (2005) Image completion with structure propagation. *ACM Trans Graph* 24(3):861–868
25. Pavić D, Schonefeld V, Kobbelt L (2006) Interactive image completion with perspective correction. *The Vis Comput PG 2006*, 22(9–11), pp 671–681
26. Barnes C, Shechtman E, Finkelstein A, Goldman DB (2009) Patch-Match: a randomized correspondence algorithm for structural image editing. *ACM Trans Graph* 28(3), Article 24:1–11
27. Hays JH, Efros AA (2007) Scene completion using millions of photographs. *ACM Trans Graph* 26(3):4–1–7
28. Liu CX, Yang YZ, Peng QS, Wang J, Chen W (2008) Distortion optimization based image completion from a large displacement view. *Comput Graph Forum* 27(7):1755–1764
29. Liu CX, Guo YW, Pan L, Peng QS, Zhang FY (2007) Image completion based on the views of large displacement. *The Vis Comput* 23(9):833–841

Chapter 52

Application of 2-D Wavelet Transform in Image Compression Based on Matlab

Shi-Gang Hu, Xiao-Feng Wu and Zai-Feng Xi

Abstract As one of the most remarkable tools, wavelet performs better than Discrete Fourier Transform (DFT) or Discrete Cosine Transform (DCT) in many aspects of image processing because of its good characteristics such as compact supported. From the angle of the experiment in the Matlab, the paper discusses and analyzes the application of image compression based on the 2-D wavelet transformation.

Keywords Wavelet transform · Image reconstruction · Image compression

52.1 Introduction

Image is some important kind of information carrier in our life and one image usually contains more abundant information than other Medias, such as words. But the raw digital images are very big in size. If the image data could not be compressed efficiently, it will not meet the needs of high speed transmission and efficient storage nowadays. So, image compression, as a crucial, promising and vigorous technology in the fields of communication and multimedia, is rapidly expanding now.

Wavelet theory has developed since the beginning of the last century. It was first applied to signal processing in the 1980s, and over the past decade it has been recognized as having great potential in image processing applications. Wavelet

S.-G. Hu (✉) · X.-F. Wu · Z.-F. Xi
School of Information and Electrical Engineering, Hunan University of Science and
Technology, Xiangtan 411201, China
e-mail: hsg99528@126.com

transforms are essentially extensions of the idea of high pass filtering. In visual terms, image detail is a result of high contrast between features, for example a light rooftop and dark ground, and high contrast in the spatial domain correspond to high values in the frequency domain. Frequency information can be extracted by applying Fourier transforms, however it is then no longer associated with any spatial information. Wavelet transforms can therefore be more useful than Fourier transforms, since they are based on functions that are localized in both space and frequency. The detail information that is extracted from one image using wavelet transforms can be injected into another image using one of a number of methods, for example substitution, addition, or a selection method based on either frequency or spatial context. Furthermore, the wavelet function used in the transform can be designed to have specific properties that are useful in the particular application of the transform [1–3].

An important application of wavelet analysis is using the two-dimensional wavelet analysis to do the image compression. It is characterized by high compression ratio and high compression speed. And it can maintain essentially the same characteristics after the compression.

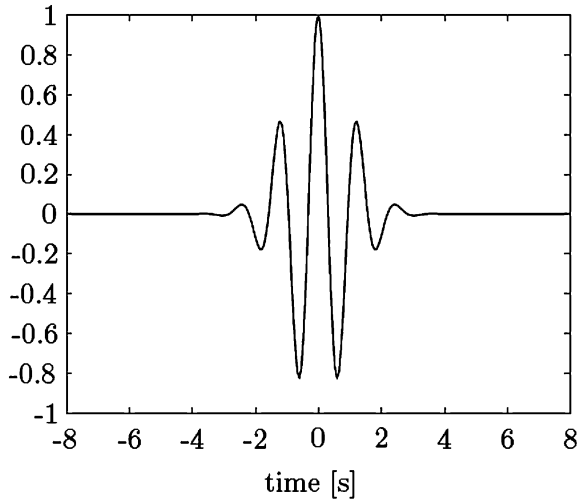
The theory of 2-D wavelet and its decomposition and reconstruction algorithms are illustrated, and the realization of image compression is demonstrated by Matlab. Practice has proved that the 2-D wavelet has a good effect on image compression.

52.2 The Basic Ideal of Wavelet Transform

Wavelets are mathematical functions that cut up data into different components and the study each component with a resolution matched to its scale. Wavelets were developed independently in the fields of mathematics, quantum physics, electrical engineering, and seismic geology.

The wavelet transform is one mathematical analysis method, which is developed in the recent years. It is the development of Fourier transform. All wavelet transforms may be considered forms of time–frequency representation for continuous-time (analog) signals and so are related to harmonic analysis. Almost all practically useful discrete wavelet transforms use discrete-time filterbanks. These filter banks are called the wavelet and scaling coefficients in wavelets nomenclature. These filterbanks may contain either finite impulse response (FIR) or infinite impulse response (IIR) filters. The wavelets forming a continuous wavelet transform (CWT) are subject to the uncertainty principle of Fourier analysis respective sampling theory: Given a signal with some event in it, one cannot assign simultaneously an exact time and frequency response scale to that event. The product of the uncertainties of time and frequency response scale has a lower bound. Thus, in the scaleogram of a continuous wavelet transform of this signal, such an event marks an entire region in the time-scale plane, instead of just one point. Also, discrete wavelet bases may be considered in the context of other forms

Fig. 52.1 Morlet wavelet



of the uncertainty principle. Wavelet transforms are broadly divided into three classes: continuous, discrete and multiresolution-based.

In comparison to the Fourier transform, the analyzing function of the wavelet transform can be chosen with more freedom, without the need of using sine-forms. A wavelet function $\psi(t)$ is a small wave, which must be oscillatory in some way to discriminate between different frequencies. The wavelet contains both the analyzing shape and the window. Figure 52.1 shows an example of a possible wavelet, known as the Morlet wavelet. For the CWT several kind of wavelet functions are developed which all have specific properties.

The continual wavelet transformation is defined as:

$$WT_x(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} \psi\left(\frac{t-b}{a}\right)x(t)dt \tag{52.1}$$

In practice, discrete wavelet transform is used frequently, and it is also required that the signal can be reconstructed. Make $a = a_0^m$ and $b = na_0^m b_0$, then the discrete wavelet transformation is:

$$\varphi_{m,n} = a_0^{-m/2} \varphi(a_0^{-m}t - nb_0) \tag{52.2}$$

Take $a_0 > 1$ and $b_0 \neq 0$, Then the discrete wavelet transform is:

$$WT_x(m, n) = \int_{-\infty}^{+\infty} \varphi_{m,n}(t)x(t)dt \tag{52.3}$$

Take $a_0 = 2$ and $b_0 = 1$, equation (2) can be written as a function systems [4]

$$\varphi_{j,k}(t) = \left\{ 2^{-j/2} \varphi(2^{-j}x - k) \right\}_j \quad k \in z \quad (52.4)$$

The above equation is standard orthogonal basis according to the binary expansion and translation on $L_2(\mathbb{R})$ [5].

The wavelet has satisfied three essential requirements as following:

- Wavelet is a building block of general function. Wavelet can be used as basis function and carries on wavelet series expansion to the general function.
- Wavelet has the characteristic of time–frequency accumulation. Wavelet function is compactly supported in the time domain, fast attenuation in the frequency domain.
- Wavelet has the fast algorithm.

The realization of the above three essential requirements constitutes the central content of wavelet transform, which urges the wavelet to become one kind of useful signal processing tool.

Wavelet transforms have received attention recently due to their suitability for signal and image processing tasks, including image coding. The wavelet transform allows one to obtain information about both the scale size and position of a feature within a profile.

52.3 Application of Wavelet Transformation on Image Compression Code

Image compression is very important for efficient transmission and storage of images. Demand for communication of multimedia data through the telecommunications network and accessing the multimedia data through Internet is growing explosively. With the use of digital cameras, requirements for storage, manipulation, and transfer of digital images, has grown explosively. These image files can be very large and can occupy a lot of memory. A gray scale image that is 256×256 pixels have 65, 536 elements to be stored, and a typical 640×480 color image has nearly a million. Downloading of these files from internet can be very time consuming task. Image data comprise of a significant portion of the multimedia data and they occupy the major portion of the communication bandwidth for multimedia communication. Therefore development of efficient techniques for image compression has become quite necessary. A common characteristic of most images is that the neighbouring pixels are highly correlated and therefore contain highly redundant information. The basic objective of image compression is to find an image representation in which pixels are less correlated. The two fundamental principles used in image compression are redundancy and irrelevancy. Redundancy removes redundancy from the signal source and irrelevancy omits pixel values which are not noticeable by human eye.

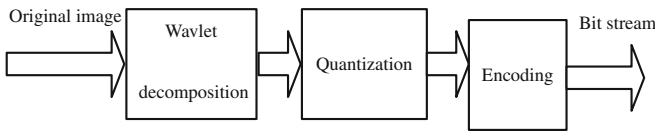


Fig. 52.2 Image compression coding model based on wavelet transform

There are two types of image compression: lossless and lossy. In lossless compression schemes, the reconstructed image, after compression, is numerically identical to the original image. However lossless compression can only achieve a modest amount of compression. Lossless compression is preferred for archival purposes and often medical imaging, technical drawings, clip art or comics. This is because lossy compression methods, especially when used at low bit rates, introduce compression artifacts. An image reconstructed following lossy compression contains degradation relative to the original. Often this is because the compression scheme completely discards redundant information. However, lossy schemes are capable of achieving much higher compression. Lossy methods are especially suitable for natural images such as photos in applications where minor (sometimes imperceptible) loss of fidelity is acceptable to achieve a substantial reduction in bit rate. The lossy compression that produces imperceptible differences can be called visually lossless.

Wavelet Transform has become an important method for image compression. Wavelet based coding provides substantial improvement in picture quality at high compression ratios mainly due to better energy compaction property of wavelet transforms. Wavelet transform partitions a signal into a set of functions called wavelets. Wavelets are obtained from a single prototype wavelet called mother wavelet by dilations and shifting. The wavelet transform is computed separately for different segments of the time-domain signal at different frequencies. Image compression coding model based on the wavelet usually includes three parts: Wavelet decomposition, Quantization, Encoding [6]. The three parts are discussed as shown in Fig. 52.2.

- (1) Wavelet decomposition: An image can be seen as a 2D signal, so an image's wavelet transform can be obtained by the Mallat algorithm. In the wavelet frequency field, an image's edge feature information and detail information are distributed in high-frequency sub-images. But there is still more detailed information in these sub-images. In order to obtain more image detail information, all high-frequency subimages are decomposed with Wavelet transform.

The appropriate wavelet base is selected to decompose the image into two parts, namely low frequency smooth part (LL) and high frequency detail parts (LH, HL, HH), altogether four sub-images. Where LH, HL, HH correspond to the edges and details of the image of the horizontal direction, vertical direction and diagonal

Fig. 52.3 2-D wavelet decomposition of image

LL ₂	HL ₂	HL ₁
LH ₂	HH ₂	
LH ₁		HH ₁

direction. Such decomposition can be repeated for LL part. 2-D wavelet decomposition for image is shown in Fig. 52.3.

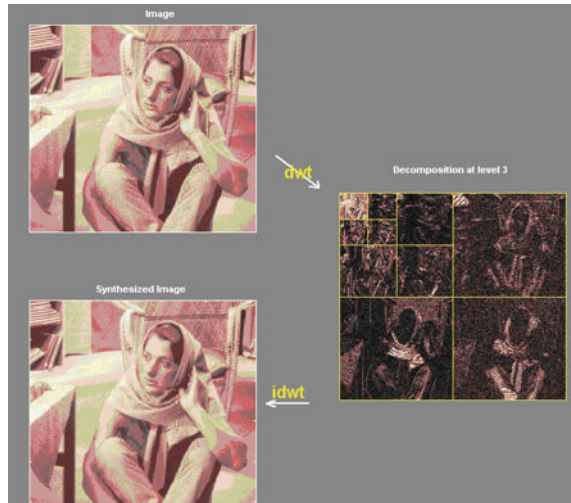
Consider that the separability of wavelet function, the two-dimensional filter can be made of one-dimensional filters compositely. L expresses the low-pass filter and H expresses the high-pass filter. Then LL, LH, HL and HH constitute 4 filters with different frequency characteristics and directions. LL is used to detect the low frequency component of the image, LH the edge and the detail component of the horizontal direction, HL the edge and the detail component of the vertical direction, HH the component of the main diagonal and sub-diagonal directions.

- (2) Quantification: It is the element of lossy systems responsible for reducing the precision of data in order to make them more compressible. Quantization means mapping of a broad range of input values to a limited number of output values, which is an irreversible bins and representing each bin with a value. The value selected to represent a bin is called the reconstruction value. Every item in a bin has the same reconstruction value, which leads to information loss unless the quantization is so fine that every item gets its own bin. The goal of quantization stage is to reduce the values of the transformed coefficients in order to reduce the precision of the subbands and achieve the compression. The transformed wavelet coefficients are quantized to generate character streams which have enough small entropy and have low bit rate.
- (3) Encoding: The results of quantization are encoded. By using Run Length coding or Huffman coding, encoding can be error free. In order to further reduce the bit rate it can be lossy. This technique allows the efficient transmission, storage and display of images.

52.4 Experimental Results and Discussion

We make experiments using mathematical analysis tools Matlab. As shown in Fig. 52.4, three layers wavelet decomposition and reconstruction of the image of wbarb with bior3.7 biorthogonal wavelet, Fig. 52.5 shows the image of low-

Fig. 52.4 Wavelet decomposition and reconstruction of image



frequency approximate image and that of high-frequency horizontal, vertical, diagonal in each layer after decomposition using tree structure.

The decomposition and reconstruction of the image are made simply in the above experiment using biorthogonal wavelet, but without quantified coding. Then the image is quantified by programming in order to realize the compression of the image. First, we get a series of different sub-image resolution after wavelet decomposition, the frequency of sub-image with different resolution is not the same. Since most of the coefficients of high frequency detail in sub-image are close to 0, the higher the frequency, this phenomenon was more obvious. For an image, low-frequency part is the most important part of an image. Removing high-frequency part directly and retaining only the low frequency part by programming so as to realize the compression. Figure 52.6 shows the compression results by reserving low-frequency information in the second layer. Since the low-frequency information of the first layer is extracted after wavelet decomposition of the original image in the first compression, the compression rate is relatively small about $1/3$ at this time, the compression is better. The low frequency obtained after the decomposition of first layer again is extracted in the second compression, the compression rate is higher about $1/12$, the visual effect is also good.

52.5 Conclusions

Image compression is minimizing the size in bytes of a graphics file without degrading the quality of the image to an unacceptable level. The reduction in file size allows more images to be stored in a given amount of disk or memory space. It also reduces the time required for images to be sent over the Internet or

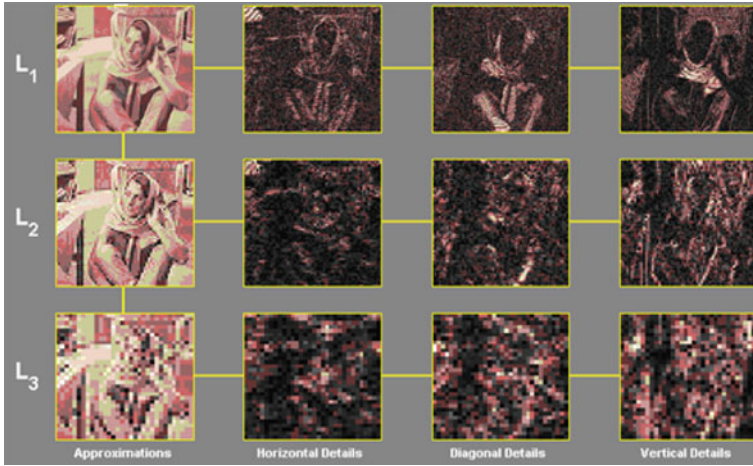


Fig. 52.5 Wavelet decomposition and reconstruction with tree structure

Fig. 52.6 The image after the second compression



downloaded from Web pages. Wavelet transforms are one of the most powerful tools in the field of image compression. Wavelet transforms provide a multiresolution representation of signals. In this paper, the theory of 2-D wavelet and its decomposition and reconstruction algorithms are illustrated, and the realization of image compression is demonstrated by Matlab. Practice has proved that the 2-D wavelet have a good effect on image compression.

References

1. Yu B, Vetterli M (2000) Adaptive wavelet thresholding for image denoising and compression. *IEEE Trans Image Process* 9:1532–1545
2. Averbuch A, Lazar D, Israeli M (1996) Image compression using wavelet transforms and multiresolution decomposition. *IEEE Trans Image Process* 5:4–14
3. Lewis A, Knowles G (1992) Image compression using 2D wavelet transform. *IEEE Trans Image Process* 1:244–250
4. Mallat SG (1989) Multifrequency channel decomposition of images and wavelet models. *IEEE Trans Acoust Speech Signal Process* 37:2091–2110
5. Mallat SG (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans Pattern Recognit Mach Intell* 1:674–693
6. Zhang X, Lu G, Feng J (2004) *Fundamentals of image coding and wavelet compressing: principles, algorithms and standards*. Tsinghua University Press, Beijing

Chapter 53

Fingerprint Image Preprocessing Method Based on the Continuous Spectrum Analysis

Xiaosi Zhan and Zhaocai Sun

Abstract The spectral field analysis is one key means for processing the fingerprint image in automatic fingerprint identification technology. The fingerprint ridge distance can be acquired and the fingerprint image can be enhanced accurately using the spectral field analysis method. However, traditional Fourier transform spectral analysis method had the worse redundancy degree in estimating the ridge distance because it was based on the two-dimension discrete Fourier spectrum. To introduce the sampling theorem into the fingerprint image processing method and transform the discrete spectrum into the continuous spectrum. To acquire the local peak points adopting artificial immune network optimization algorithm and then segment valid fingerprint image region from background region or stressed-noised region, obtain ridge distance and ridge orientation of valid fingerprint regions based on the continuous spectrum analysis in frequency field. The experimental results indicate that the method has higher adaptability and can improve the accuracy of the automatic fingerprint identification system effectively.

Keywords Fingerprint · Fingerprint identification · Ridge distance · Fourier transform · Artificial immune network

X. Zhan (✉)

School of Information Science and Technology, Zhejiang International Studies University, Hangzhou, 310012, People's Republic of China
e-mail: xiaoszhan@263.net

Z. Sun

School of Computer and Information, Fuyang Normal College, Fuyang, 236032, People's Republic of China

53.1 Introduction

Biometrics is formed from the person's selected unique physical attributes which may be applied for the purpose of automated personal identification [1]. Fingerprint identification is one of the popular biometric technology, which has drawn a substantial attention recently [1–6]. In order to improve the quality of fingerprint image effectively, most fingerprint enhancement algorithms regard ridge distance and ridge orientation as two important parameters. So, it is important to estimate accurately the ridge distance and orientation for improving the performance of AFIS.

Lots of literatures focus on the technology of the fingerprint image segmentation. Methre et al. [2–5] used gray variance and the histogram of pixel gradients in a sub-image block for segmentation. Chen et al. [6] used linear classifier to classify the blocks into foreground and background. Bazen and Gerez [7] used the linear classifier to classify the pixels into foreground pixels or background pixels through the analysis of the pixel features. The performance of linear classifier had certain limits, so it was difficult to reach the ideal segmentation results. Yin et al. [8] used the model of quadratic curve surface to carry out the fingerprint image segmentation, which regarded the gray variance, the gray mean and the orientation coherence as the parameters of the model. He et al. [9] regarded the fingerprint image as Markov Random Field and carried out the fingerprint image segmentation successfully. But it can generate the boundary curve only when the edge contrast between fingerprint and background is stronger.

Generally, fingerprint ridge distance can be regarded as the distance from the center of one ridge to the center of the other adjacent ridge or the pieces of ridge width and valley width in close proximity. In the research region, Hong et al. [10] proposed the direction window method to estimate the ridge frequency. This method can estimate the ridge frequency reliably when the contrast of the fingerprint image is good and the ridge orientation in the direction window is consistent. Yin et al. [11] proposed the ridge distance estimation method based on regional level. The method divided the fingerprint image into several regions according to the consistency of the orientation information on the whole fingerprint image and calculates the ridge distance to every region respectively. O'Gorman and Nickerson [12] proposed the ridge distance based on statistics mean value. Kovace-Vajna et al. [13] brought out two kinds of ridge distance estimation methods: the geometry approach and the spectral analysis approach, which were both based on the block-level to estimate the ridge distance. The merit of geometric method was that it calculated ridge direction directly. Spectral method divided a fingerprint image to blocks and converted each block of fingerprint image from spatial field to frequency field with Discrete Fourier Transform (DFT) and estimated ridge distance of a block image according to the distribution of harmonic coefficients. Mario and Maltoni [14] did mathematical characterization of the local frequency of sinusoidal signals and developed a two-dimensional (2-D) model to approximate the ridge-line patterns. Chen et al. [15]

proposed two kinds of methods to estimate the ridge distance: the spectral analysis approach and the statistical window approach.

Castro et al. [16] proposed the evolutionary immune network and worked it into data clustering. Xiao et al. [17] proposed one artificial immune system model and analyzed its principle, configuration, and prospects. Zhan et al. [18] applied the artificial immune network in the fingerprint ridge distance estimation.

This paper puts forward the fingerprint image processing method based on continuous spectrum analysis and artificial immune network optimization algorithm in the frequency field. Firstly, the method transforms the fingerprint image expressed in the spatial field into the spectrum expressed in the frequency field. Secondly, the method transforms the 2-D discrete frequency spectrum into continuous frequency using the sampling theorem. Then, the method calculates two relatively symmetrical local peak points relative to the peak value adopting the artificial immune network optimization algorithm in the continuous frequency spectrum map. Lastly, the paper segments the fingerprint image region into the valid fingerprint region and background region, obtains the ridge distance and the ridge orientation field of the valid fingerprint image region based on the continuous spectrum analysis. The comparing experiment results show that the methods in the paper can improve the performance of the automatic fingerprint identification system effectively.

53.2 Spectrum Analysis Method of Fingerprint Image

Fingerprint image is one kind of regular texture image essentially. So the frequency of fingerprint image must be principle and can be utilized fully in the fingerprint image processing. The following Fig. 53.1b shows the spectrum after carrying the Fourier transform to the regular texture image (Fig. 53.1a) with DFT. It can be seen that there be some “lighter spot” on some principle. It has reflected the frequency information and the orientation information of the ridge in the regular image. The distance among the light spot is in direct proportion to ridge frequency and in inverse proportion to ridge distance. The direction of the line connecting two light spots is vertical with the ridge orientation and parallel to the direction of the normal line of the ridge. The total energy in the spectrum of an image can be defined as the following formula 1 according to Parseval Theorem.

$$\sum_{n=0}^{N-1} |x(n)|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |X(k)|^2 \quad (53.1)$$

So the main energy of texture image can be considered to concentrate on the “lighter spot” in the spectrum. Additionally, the “lighter spot” must appear at the normal orientation of the ridge. Hence, the “lighter spots” include information of

Fig. 53.1 The regular texture image and the corresponding spectrum.
a regular texture image.
b corresponding spectrum

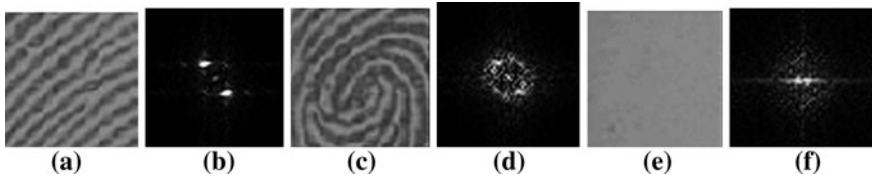
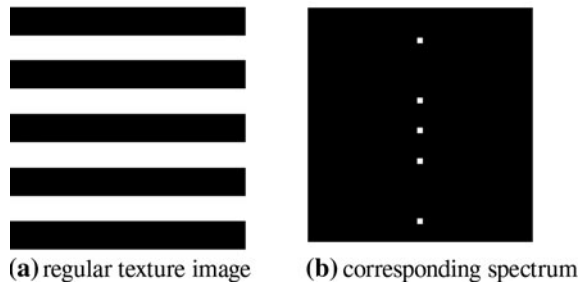


Fig. 53.2 Valid fingerprint region, model region, background region and the corresponding spectrum

the texture. If we acquire the difference in frequency field between different fingerprint image regions, we can utilize it to segment fingerprint image, obtain the ridge distance and orientation information and enhance the fingerprint image.

53.2.1 The Spectrum Analysis of Different Fingerprint Image Regions

In order to illustrate the difference between different fingerprint image regions, we transform the different fingerprint image region from the spatial region into the frequency region firstly and then observe the difference of the spectrum between every different fingerprint image regions.

Now, the key issue is how to acquire the “lighter spots” accurately. But the “lighter spots” are a little dispersed from the Fig. 53.2. In order to observe the spectrum of the fingerprint image accurately, we transform the 2-D into 3-D firstly. The x-axis and the y-axis are respectively the x-coordinate and the y-coordinate of the spectrum image and the z-axis denotes the energy on the corresponding location. The following Fig. 53.3 is the 3-D spectrum image of the valid fingerprint image region and the background region respectively.

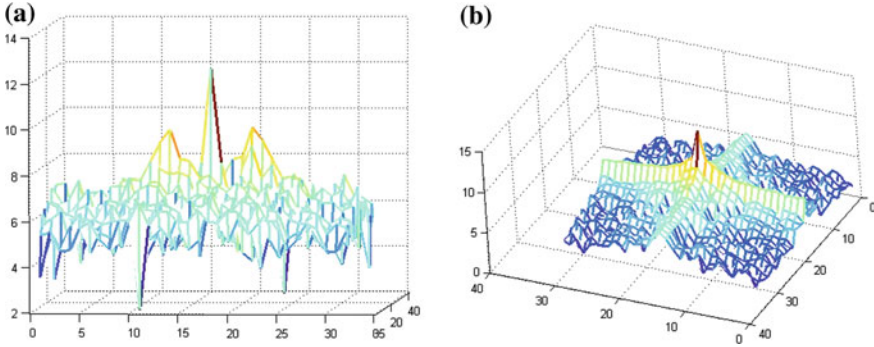


Fig. 53.3 The sketch map about the three-dimension spectrum (a and b are the corresponding spectrum of the Fig. 53.2b and f respectively)

53.2.2 The 3-Dimension Continuous Spectrum of the Fingerprint Image

The sampling theorem plays a great role in the signal processing. It is also suit to the 2-D image processing. Here, we can adopt the 2-D sampling theorem to transform the discrete spectrum into the continuous spectrum.

Suppose that the Fourier transform function $F(s_1, s_2)$ about the function $f(x_1, x_2)$ of $L_2(R_2)$ is tight-supported set (namely that the function F is equal to zero except the boundary region D and the boundary region D can be defined as the rectangle region $\{(s_1, s_2) ||s_1| \leq \Omega \text{ and } |s_2| \leq \Omega\}$ in the Chapter). Here, we firstly assume $\Omega = \pi$ in order to simplify the function. Then the Fourier transform function about the function $f(x_1, x_2)$ can be denoted as follows:

$$F(s_1, s_2) = \sum_{n_1} \sum_{n_2} C_{n_1, n_2} e^{-jn_1s_1 - jn_2s_2} \tag{53.2}$$

Then, we can acquire the following function as:

$$f(x_1, x_2) = \sum_{n_1} \sum_{n_2} C_{n_1, n_2} \frac{\sin \pi(x_1 - n_1)}{\pi(x_1 - n_1)} \frac{\sin \pi(x_2 - n_2)}{\pi(x_2 - n_2)} \tag{53.3}$$

In this way, the discrete signal C_{n_1, n_2} can be recovered to the continuous signal $f(x_1, x_2)$ through the sampling theorem. Then, the discrete frequency spectrum can be recovered to continuous frequency spectrum. The following Fig. 53.4 is the continuous frequency spectrum recovered from the Fig. 53.3 in the step length as 0.1.

From the Fig. 53.4, we can find that the local peak value is obvious in the continuous frequency spectrum of the valid fingerprint image region. Contrarily, there are not the obvious local peak value in the continuous spectrum of the stressed-noised image region or the background region. If we can acquire the local

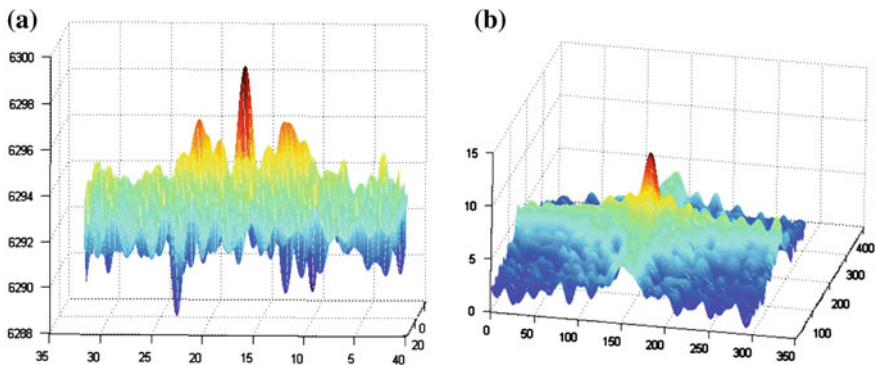


Fig. 53.4 The sketch map about the three-dimension continuous spectrum

peak values in the continuous spectrum of the valid fingerprint image region accurately and quickly, we can segment the valid fingerprint image region from other regions and estimate the ridge orientation and ridge distance effectively. But, how to obtain the local peak value is the key issue. Observe carefully the Fig. 53.4, we find that the continuous spectrum can be regarded as the multimodal function. Here, the paper regards the continuous spectrum as the multimodal function and adopt the artificial immune network optimization algorithm to obtain the local peak value.

53.3 Fingerprint Image Segmentation Based on Continuous Spectrum

As stated previously, the “energy” always concentrate at the normal orientation of the ridge. So, the local maximum at the normal orientation must be the biggest. The method calculating the ridge orientation was proposed by Hong in [10]. It can be summed up as following:

$$V_y(i,j) = \sum_{u=i-\frac{w}{2}}^{i+\frac{w}{2}} \sum_{v=j-\frac{w}{2}}^{j+\frac{w}{2}} 2\partial_x(u,v)\partial_y(u,v) \tag{53.4}$$

$$V_x(i,j) = \sum_{u=i-\frac{w}{2}}^{i+\frac{w}{2}} \sum_{v=j-\frac{w}{2}}^{j+\frac{w}{2}} 2\partial_x(u,v)\partial_x(u,v) \tag{53.5}$$

$$\theta(i,j) = \frac{1}{2} \arctan\left(\frac{V_y(i,j)}{V_x(i,j)}\right) \tag{53.6}$$

Fig. 53.5 Comparison of the local maximum between different slices

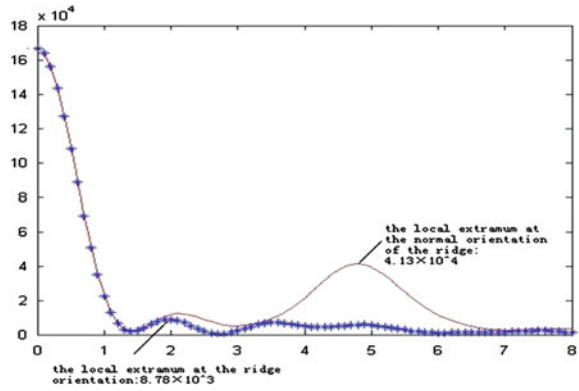
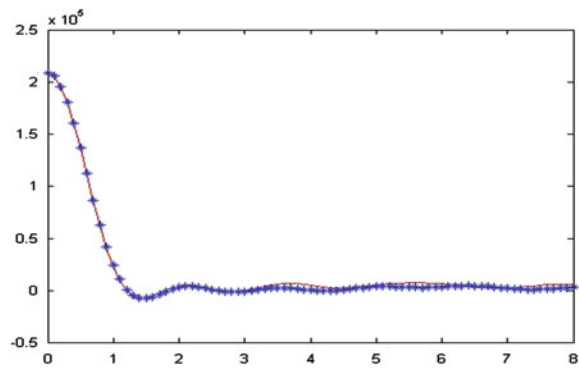


Fig. 53.6 Slices of background region



where, w is the block size, $\partial_x(u, v)$ and $\partial_y(u, v)$ are the differential coefficient of the (u, v) pixel at the x -orientation and the y -orientation respectively.

Supposing the present ridge orientation of the block is θ_0 , then the normal orientation $\theta_1 = \theta_0 + \pi/2$. For example, if $\theta_0 = 1.38$, then $\theta_1 = 1.38 + \pi/2 = 2.95$. Now, we search the local maximum at θ_0 and θ_1 in the continuous spectrum. In the following Fig. 53.5, the red line is the slice searched in the Fig. 53.4 at the normal orientation of the ridge θ_1 , while the “*” line is searched at the ridge orientation θ_0 .

So, it is easy to calculate the local maximum at θ_0 and θ_1 in Fig. 53.5, the local maximum at the normal orientation of the ridge is 4.13×10^4 , while that at the ridge orientation is 8.78×10^3 . Obviously, the comparison is vast. Otherwise, the comparison in the background region is small, as Fig. 53.6 shows.

Since the comparison of the local maximum in fingerprint region is obviously greater than that in background region, we can use the degree of comparison as the texture index. If $peak_0$, $peak_1$ is denoted as the local maximum at the orientation θ_0 and θ_1 , respectively, the texture can be written as,

$$Tex = \text{sign}(peak_1 - peak_0) \times \ln(|peak_1 - peak_0|) \quad (53.7)$$

For example, the Tex of Fig. 53.1 is, $\log(4.13 \times 10^4 - 8.78 \times 10^3) = 10.39$.

53.4 The Realization of Fingerprint Image Processing Method

After obtained the local peak value, the ridge distance and ridge orientation can be obtained from the locality of the two local peak values and the valid fingerprint region can be segmented from the background region or other stressed-noised region. From the analysis about the Fourier frequency spectrum, we can conclude that there are only two local peak values which are the concentrated points of the main spectrum energy. In order to estimate the validity of these local peak values, the paper puts forward the following estimating rules:

- Rule 1: Two local peak points should be in the same plane as possible, that is to say that the energy contained in the two local peak points contain is equal approximately.
- Rule 2: The orientations between the center point and the two local peak points are equal almost, which can be seen the orientation of the tangent of the texture in the fingerprint image.

Based on the above two rules, we can obtain the two local peak points accurately. Then, we segment the fingerprint image into valid fingerprint image region and the background region combined the gray variance and the rate of the main spectrum energy and the total spectrum energy. After segmenting the fingerprint image, we obtain the ridge orientation from the relation between the ridge orientation and the ridge frequency and calculate the ridge distance based on the following function: $r = N/d$ based on the relation between the ridge distance and the ridge frequency in the spectrum.

53.5 Experimental Results

In order to evaluate that our method is suitable for fingerprint image segmentation, we select samples from Database B of BVC2004, which is considered hard to segment commonly. Figure 53.7 shows segmented results of the representative fingerprint images. The results indicate that the method is efficiency, accurately and “robust”.

In order to evaluate that our method is suitable for processing the fingerprint image, we use the fingerprint sample set that includes 30 typical fingerprint images (ten high quality, ten fair quality, ten poor quality) selected from NJUF fingerprint database (1200 live-scan images; ten per individual) used for evaluation. Here, we compare the segmentation results between our method and fingerprint image

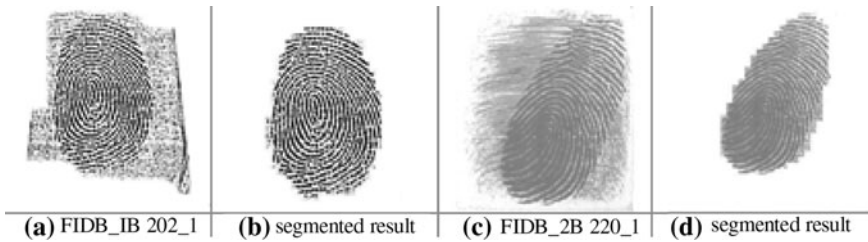


Fig. 53.7 Segmentation results of the representative fingerprint images

Table 53.1 Segmentation results with our method and L. Hong method

Method	SARofH (%)	SARofF (%)	SARofP (%)	PT (s)
L. Hong method	100	94.5	85.4	0.08
Our method	100	96	92.7	0.37

segmentation based on gray variance proposed by Jain and Hong. In addition, we compare our ridge orientation extraction method with Rao method proposed by Jain and Hong. Here, the 4 indexes: SARofH, SARofF, SARofP and PT are the segmentation accuracy rate to high quality fingerprint images set, fair quality fingerprint images, poor quality fingerprint images and the total processing time by our method and Hong method, the four indexes: OARofH, OARofF, OARofP and PT are the ridge orientation extraction accuracy rate to high quality fingerprint images set, fair quality fingerprint images, poor quality fingerprint images and the total processing time by our method and Rao method. The referenced right segmentation results and ridge orientation extraction results are obtained manually. Lastly, we use 30 typical images to estimate ridge distance with traditional spectral analysis method, statistical window method and our method respectively and compare the ridge distance extraction results with the three methods. Sizes of block images are all 32×32 and the value of ridge distance of each block image is manually measured. The three indexes of the performance evaluation method are DER, EA and PT, which were put forward by Yin et al. in [3]. Where DER indicates the robustness of a method for ridge distance estimation in fingerprint images, EA is the degree of deviation between the estimation results and the right ridge distances, PT is the processing time with the method.

From Table 53.1 we can conclude that our method has higher SARofF and SARofP values, which mean that our method is more robust than Hong method. To the poor quality fingerprint images, our method's segmentation result is more accurate.

The results are illustrated in Table 53.2 comparing Rao method for fingerprint ridge orientation extraction proposed by Jain and Hong with our method based on the continuous spectrum proposed in the Chapter.

Table 53.2 Ridge orientation extraction results with two method

Method	OARofH (%)	OARofF (%)	OARofP (%)	PT (s)
L. Hong method	100	97.5	65.4	0.15
Our method	100	95.7	88.7	0.43

Table 53.3 Ridge distance extraction results with three methods

Method	DER (%)	EA (%)	TC(s)
Traditional spectral analysis	44.7	84	0.42
Statistical window	63.8	93	0.31
Our method	92.1	96	0.74

From the Table 53.2 we can consider that our method has the higher performance than Rao method for the poor quality fingerprint images. For the high quality fingerprint images, the performance of the two methods can be considered equal approximately. The most disadvantage for our method is the processing time.

The ridge distance estimation results are illustrated in Table 53.3 comparing the traditional spectral analysis method, the statistical window method and our method.

From the Table 53.3, we can find that the traditional spectral analysis method has the lowest DER value and EA value with the middle TC value, the statistical method has the middle DER value and the EA value with the lowest TC value while our method has the higher DER value and the EA value with the higher TC value. The obvious disadvantage of statistical window method is that it doesn't perform well in regions where there is acute variation of ridge directions although the average performance of statistical window method is superior to that of the spectral analysis method. The experimental results show that our method can obtain the ridge distance of most regions in a fingerprint image except the pattern region and the strong-disturbed region because the sub-peak is not obvious in these regions. It has higher performance although the processing time of our method is more than the other two methods.

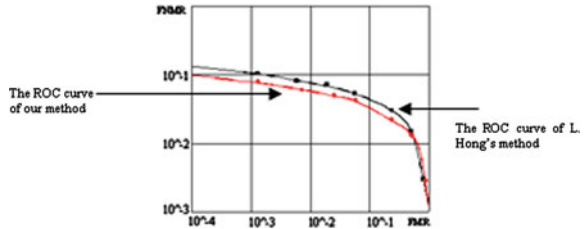
In order to verify that the continuous spectrum analysis works efficiently in fingerprint image processing, we chose the NJUF fingerprint database to evaluate the performance of our method. Here, we chose the same fingerprint match method for evaluating the accuracy of the fingerprint minutiae extraction using our method and Hong method respectively. The following Table 53.4 shows that the performance of our method is better than the performance of Hong method.

The Fig. 53.8 describes the performance ROC curves about our method and Hong method. Performance improvement is obtained by using the continuous spectrum analysis method for acquiring the more accurate fingerprint minutiae set.

Table 53.4 Results of our method and Hong method over NJUF fingerprint database

Method	EER (%)	EER* (%)	ZeroFMR (%)	ZeroFNMR (%)
L.Hong's method	3.86	3.86	11.92	100
Our method	2.94	2.94	6.72	99.52

Fig. 53.8 The ROC curves about the performance with our method and L. Hong's method (where, the FNMR index is the false not match rate and the FMR index is the false match rate)



53.6 Conclusions and Future Works

The continuous spectrum analysis method has some advantages in the signal analysis. Here, we introduce it into fingerprint image processing for improving the performance of the automatic fingerprint identification system and propose the fingerprint image processing methods based on the continuous spectrum analysis. More accurate ridge distance and orientation can be acquired by our method. In addition, we can segment the valid fingerprint image from the background region effectively. The method is more complicated and time-consumed but more effective.

Our experimental results show that fingerprint images can be well processed using the continuous spectrum analysis method. But some fingerprint images of very bad quality as well as unclear ridge structures can not be processed until now. In addition, our future research will focus on looking for other better methods which can transform the spatial fingerprint image into the continuous frequency spectrum and improve the local peak points obtained algorithm for acquiring the two sub-peak points faster and more accurately.

Acknowledgments The authors would like to be grateful to the editors and anonymous reviewers for their works. This work was supported in part by Zhejiang Province Science Foundation of China under of No.Y1101304, Anhui Province Science Foundation of China under of No.20090412072 and Anhui Province Education Department Science Foundation of China under Grant No.2005KJ089. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors.

References

1. Prabhakar S, Pankanti S, Jain AK (2003) Biometrics recognition: security and privacy concerns. *IEEE Secur Priv Mag* 1:33–42
2. Mehtre BM, Murthy NN, Kapoor S (1987) Segmentation of fingerprint images using the directional image. *Pattern Recognit* 20(4):429–435

3. Mehtre BM, Murthy NN (1996) A minutiae-based fingerprint identification system. Second international conference advances in pattern recognition and digital techniques, Calcutta
4. Mehtre BM (1989) Segmentation of fingerprint images-a composite method. *Pattern Recognit* 22(4):381–385
5. Mehtre BM (1993) Fingerprint image analysis for automatic identification. *Mach Vis Appl* 6:124–139
6. Chen X, Tian J, Cheng J et al (2004) Segmentation of fingerprint images using linear classifier. *EURASIP J Appl Signal Process* (4):480–494
7. Bazen AM, Gerez SH (2001) Segmentation of fingerprint images. In: *Proceedings of the 12th annual workshop on circuits, systems and signal processing, Veldhoven*, pp 29–30
8. Yin Y, Yang X, Chen X et al (2004) Method based on quadric surface model for fingerprint image segmentation. In: *Proceedings of SPIE*, pp 417–324
9. He Y, Tian J, Zhang X (2002) Fingerprint segmentation method based on Markov random field. In: *Proceedings of the 4th China graph conference*, pp 149–56
10. Hong L, Wan Y, Jain AK (1998) Fingerprint image enhancement: algorithm and performance evaluation. *IEEE Trans Pattern Anal Mach Intell* 20(8):777–789
11. Yin L, Wang Y, Yu F (2003) A method based on region level for ridge distance estimation. *Chinese Comput Sci* 30(5):201–208
12. O’Gorman L, Neckerson JV (1989) An approach to fingerprint filter design. *Pattern Recognit* 22(1):29–38
13. Kovacs-Vajna ZM, Rovatti R, Frazzoni M (2000) Fingerprint ridge distance computation methodologies. *Pattern Recognit* 33:69–80
14. Maio D, Maltoni D (1998) Ridge-line density estimation in digital images. In: *Proceedings of the 14th international conference on pattern recognition, Brisbane, Australia*, pp 534–538
15. Chen Y, Yin L, Zhan X et al (2003) A method based on statistics window for ridge distance estimation. *J image graph, China*. 8(3): 266–270
16. Castro L, Zuben F (2000) An evolutionary immune network for data clustering. In: *Proceedings of the 6th Brazilian symposium on neural networks, Rio De Janeiro, Brazil*, pp 83–89
17. Xiao B, Wang L (2002) Artificial immune system: principle, models, analysis and prospects. *Chin J Comput* 25(12):1281–1293
18. Zhan X, Yin Y, Sun Z et al (2005) A method based on continuous spectrum analysis and artificial immune network optimization algorithm for fingerprint image ridge distance estimation. In: *Proceedings of the 5th international conference on computer and information technology, Shanghai, China*, pp 728–735

Chapter 54

A Zero-Watermarking Algorithm for Digital Map Based on DWT Domain

Yun Ling, Chuan-Feng Lin and Zhi-yong Zhang

Abstract This paper presents a zero-watermarking algorithm for grid map based on wavelet transform. The algorithm of grid map data block to do a two-dimensional wavelet transform, the histogram information of wavelet coefficients extracted features as watermark information. Experimental results show that the algorithm of grid digital map with the general attack is robust, and grid digital map data does not make any alteration, has good practical value.

Keywords Grid map · Digital watermarking · Discrete wavelet transformation (DWT)

54.1 Introduction

With the development of information technology, especially the wide applications of the geographic information systems (GIS), digital maps are playing an increasingly important role in car navigation systems, web-based services and urban planning. However, like all the other digital products, the grid map can be

Y. Ling (✉) · C.-F. Lin · Z. Zhang
College of Computer Science and Information Engineering, Zhejiang Gongshang
University, Hangzhou 310018, China
e-mail: yling@mail.zjgsu.edu.cn

C.-F. Lin
e-mail: cf_lin@live.cn

Z. Zhang
e-mail: zzy@mail.zjgsu.edu.cn

easily copied and misused without legal permission. Therefore, it has become extremely urgent to effectively protect the copyright of the grid map and prevent its illegal use of the grid map on this highly developed network.

54.1.1 Grid Map Watermarking Technology

As one kind of images, the grid map of the watermarking technology can learn about the general image watermarking technology, but current research grid digital maps specifically for digital watermarking technology was too few. Document [1] proposed a complementary watermarking algorithm for digital grid map. The watermark was embedded respectively into direct current section and high frequency section after map DCT transform. When the watermark was embedded into the DC section, the modulation amplitude was controlled according to the luminance. The watermark was embedded into the high frequency section through the parity in the sum of segmented high frequency coefficients. Document [2] proposed a wavelet-based watermarking algorithm for digital raster map. The features of digital grid map with higher luminance and greater sensual capability than natural image were utilized by the proposed algorithm. The watermark, which was disposed by period spread technique, was adaptively embedded into low frequency section according to the luminance; The same watermark was embedded into high frequency detail section to improve the robustness. When the watermark was embedded into high frequency detail section, the embedded places were determined by the characteristics of human visual systems and the intrinsic relations of wavelet coefficients. In document [3], with the blocking idea, the watermark is embedded into the Fourier semicircle region. When the watermark is extracted, the watermarked image is first preprocessed, and then the spectrum is mapped to standard size with neighbor averaging method, which assures blind extraction, and resolves the problem caused by block-image embedding as well as whole-image extraction. However, without exception, that the above figures for the raster map data watermarking algorithm will do some tampering. And some dull color in the grid map, the original image after embedding the greatest impact.

54.1.2 Zero-Watermarking Algorithms

To address the robustness and watermark hidden contradiction in terms of image watermarking algorithms, zero-watermarking [4] method was proposed. In document [5], a zero-watermark algorithm based on affine transform and wavelet transform is presented, which scrambles the embedded watermarks, then calculate the low-frequency sub-map of the wavelet transformed image with the scrambled watermarks. The result of the operation is shifted in a cycle to generate the cryptography images. Because the low-frequency sub-map of the wavelet

transformed image contains the main features of the image, this watermarking algorithm gains a high robustness against noise, filtering, and JPEG compression. Document [6, 8] proposed a zero-watermark algorithm based on wavelet transform. The algorithms embedded watermark into wavelet domain of images, and the algorithm has a relatively high robustness for cutting, compression and other attacks. Document [9] proposed a new digital image watermarking algorithm based on DWT and SVD, attacks on the geometric transformation has some robustness. As a kind of image, the watermarking algorithm of the grid map can also learn the methods of image zero-watermarking.

54.2 Zero-Watermarking Algorithm for Grid Map

A stable feature must be taken in the zero-watermarking algorithm. Document [11, 12] take the SIFT features of images as the watermark, and respectively proposed a zero-watermarking algorithm and circular spatial improved watermark embedding algorithm. Document [7] utilized MSP method to detect corners from the low frequency components as the stable futures of images. Grid maps can be viewed as a special image. However, comparing from the general image, colors are more obvious and the same color will appear in the piece. Grid map has a sparse histogram. Therefore, changes cannot be seen by eyes when embedding watermark into general images, but for the grid map, after embedding the watermark, the results will be different. So the watermarking algorithms of grid maps should be treated specially.

The method in our paper does not have to alter the data of the original grid map. Instead, we do wavelet transform based on the obvious color differences of the grid map blocks and extract the zero-watermarking information from the wavelet coefficients. On the one hand this watermarking method avoids embedding watermarks into the map data as redundant information, which will change the original data. Thus, with good concealment, it solves the contraction between robustness and imperceptibility of the digital watermark. And it also fills in the security loopholes of overcome the reversible watermarking. As a natural blind watermarking system, it has a great practical meaning. While on the other hand it can resist many common attacks, such as zooming, rotating, cropping and so on, therefore it has a strong resistance and robustness.

54.2.1 Embedding Watermarking

Steps are as follows:

Embedded watermark information as described is extracting zero-watermark information from the original grid maps, then save storage, including the following steps:

Step 1 modulation of the watermark information, binary watermarking key and the use of map and Arnold transform the watermark scrambling to get the binary bit stream w : set the watermark image is $N \times N$ of the binary image, the watermark length is $l = N \times N$, the discrete integration of the Arnold transform and the watermark key is to get the final bit stream $w = (w_1, w_2, \dots, w_i, \dots, w_l)$, where $w_i \in \{0,1\}$, $0 \leq i < l$; simultaneously generate two length l of the binary bit stream $x = (x_1, x_2, \dots, x_i, \dots, x_l)$ and $y = (y_1, y_2, \dots, y_i, \dots, y_l)$, used to preserve features of the data, where $x_i = y_i = 0$, $0 \leq i < l$;

Step 2 Partition grid map data into blocks: According to the watermark length l , the grid map data is segmented into l blocks, and each sub-block image's size is $M \times M$ square pieces; and then performs the following steps:

Step a, Decompose each sub-block image with *sym2* wavelet, and the decomposition layer is 2;

Step b, Get sub-band $LL2_i$ and detailed sub-band $HH2_i$ from the wavelet coefficients of the first i sub-block image: and $LL2_i$ stands for layer 2 approximation coefficients from the first i sub-block, $HH2_i$ stands for layer 2 diagonal detail coefficients from the first i sub-block. Then do histogram statistics for $LL2_i$ and $HH2_i$: respectively assigning $LL2_i$ and $HH2_i$ elements into the 10 equally spaced ranges, and returns the number of elements within each range as a one-dimensional vector, denoted as V_1 and V_2 ;

Step c, Sort the elements of V_1 and V_2 descending respectively, denoted as V'_1 and V'_2 , and respectively get the first two elements of V'_1 and V'_2 , denoted as A'_1, B'_1 and A'_2, B'_2 , and their position in the V'_1 and V'_2 , denoted as H_{A1}, H_{B1} and H_{A2}, H_{B2} , and then do the following treatment:

If $\text{abs}(H_{A1}-H_{B1}) > T$, $x_i = 1$;

If $\text{abs}(H_{A1}-H_{B1}) \leq T$, $x_i = 0$;

If $\text{abs}(H_{A2}-H_{B2}) > T$, $y_i = 1$;

If $\text{abs}(H_{A2}-H_{B2}) \leq T$, $y_i = 0$;

Step 3 Generate zero-watermark: XOR the scrambled watermark w with the binary bit stream x , and do the same treatment to y , the resulting data is zero-watermark information I_1 and I_2 respectively;

$$I_1 = \text{XOR}(w,x);$$

$$I_2 = \text{XOR}(w,y);$$

Step 4 Save watermarking information: save the generated watermark information I_1 and I_2 to the watermark library, and then prepare for testing when needed.

54.2.2 Watermark Detection Algorithm

In the zero-watermark algorithm, most steps in the watermark detecting are the same with the steps in the zero-watermark embedding. Steps described as follows:

Step 1 According to the grid map data to be detected, get I_1 and I_2 from watermark library. Then generate one binary bit stream w with the size of $l = N \times N$, $w = (w_1, w_2, \dots, w_i, \dots, w_l)$, $w_i \in \{0,1\}$, $0 \leq i < l$; and generate two the binary bit streams with the length of l , $x = (x_1, x_2, \dots, x_i, \dots, x_l)$ and $y = (y_1, y_2, \dots, y_i, \dots, y_l)$, which are used to preserve features of the data, set $x_i = y_i = 0$, $0 \leq i < l$;

Step 2 Partition grid map data into blocks: According to the watermark length l , the grid map data is segmented into l blocks, and each sub-block image's size is $M \times M$ square pieces; and then performs the following steps:

Step a, Decompose each sub-block image with *sym2* wavelet, and the decomposition layer is 2;

Step b, Get sub-band $LL2_i$ and detailed sub-band $HH2_i$ from the wavelet coefficients of the first i sub-block image: and $LL2_i$ stands for layer 2 approximation coefficients from the first i sub-block, $HH2_i$ stands for layer 2 diagonal detail coefficients from the first i sub-block. Then do histogram statistics for $LL2_i$ and $HH2_i$: respectively assigning $LL2_i$ and $HH2_i$ elements into the 10 equally spaced ranges, and returns the number of elements within each range as a one-dimensional vector, denoted as V_1 and V_2 ;

Step c, Sort the elements of V_1 and V_2 descending respectively, denoted as V'_1 and V'_2 , and respectively get the first two elements of V'_1 and V'_2 , denoted as A_1, B_1 and A_2, B_2 , and their position in the V'_1 and V'_2 , denoted as H_{A1}, H_{B1} and H_{A2}, H_{B2} , and then do the following treatment:

If $\text{abs}(H_{A1} - H_{B1}) > T$, $x_i = 1$;

If $\text{abs}(H_{A1} - H_{B1}) \leq T$, $x_i = 0$;

If $\text{abs}(H_{A2} - H_{B2}) > T$, $y_i = 1$;

If $\text{abs}(H_{A2} - H_{B2}) \leq T$, $y_i = 0$;

Step 3 Get the detected zero-watermark information, Do XOR operation respectively to x and I_1 , y and I_2 :

$w'_1 = \text{XOR}(x, I_1)$;

$w'_2 = \text{XOR}(y, I_2)$;

Step 4 Anti-scrambling the watermark: Do the anti-Arnold transform to the binary bit stream w'_1, w'_2 , denoted as z_1 and z_2 .

Step 5 Similarity comparison: Comparing the detecting result of the zero-watermark image z_1 and z_2 with original watermark for similarity. Formula is as follows:

$$S(X, Y) = 1 - \frac{\sum_{i=1}^l |X_i - Y_i|}{l} \quad (54.1)$$

X and Y are two binary images, and their length all are l .

$s_1 = S(z_1, w)$;

$s_2 = S(z_2, w)$; where s_1 is the similarity of z_1 and w , s_2 is the similarity of z_2 and w . Larger value would be taken as the final similarity.

Fig. 54.1 Original image and original watermark

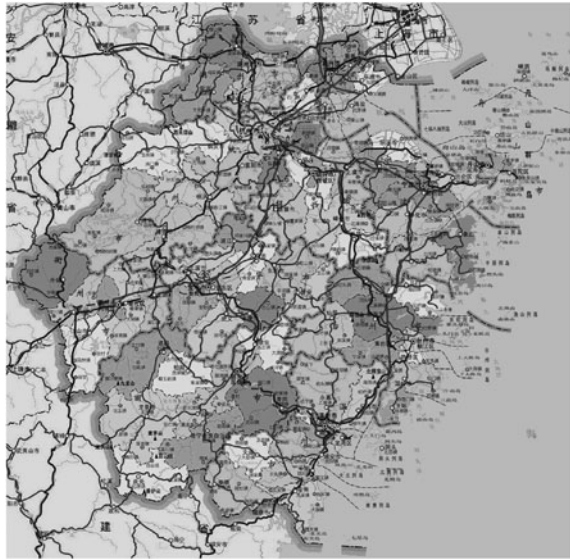
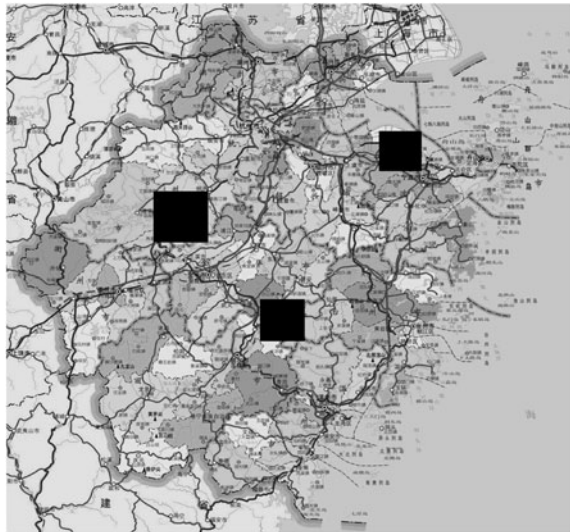


Fig. 54.2 Part of the altered



54.3 Experiment

In this experiment, the threshold T is set to 2, and the experimental picture shows 1024×1024 bmp format maps, photo watermark embedded 32×32 binary image, shown in Fig. 54.1. Figure 54.2 is to be part of the original experiment

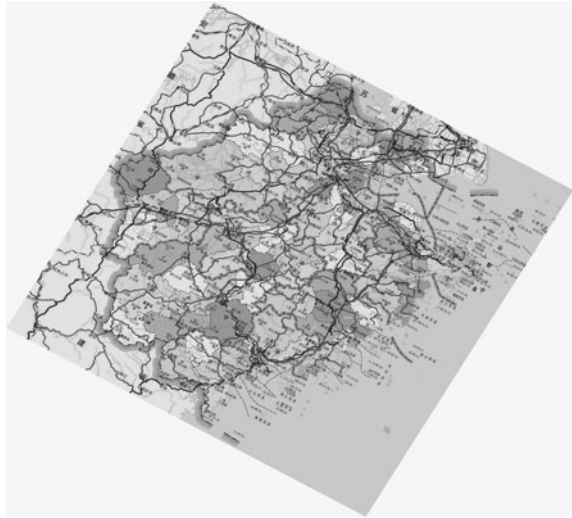
Fig. 54.3 Scale 1/4 recovery



Fig. 54.4 Jpeg compression (1: 10)



altered, Fig. 54.3 is the experimental original scale 1/4 after, and then restored to its original size image size for testing. Figure 54.4 is the original experiment carried out jpeg compression, the compression ratio is 1/10, and its watermark. Fig. 54.5 is a 30° rotation experiment conducted after the original back to the original size, and then tested.

Fig. 54.5 30° rotation**Table. 54.1** Attack and extraction results

Attack	Extraction results (similarity)
Part of the altered	0.9873
Scale 1/4 recovery	0.9590
Scale 1/2 recovery	0.9678
Jpeg compression (1: 10)	0.9746
30° rotation	0.9512
Histogram qualization	1.0000
Fuzzy	0.9678
Gaussian filter	0.9883

The results can be seen, all the extracted watermark image can tell, the watermark extraction success. Watermarking algorithm that this chapter has good imperceptible, can guarantee the accuracy of map data.

Experimental results show that the algorithm of raster map data in the zoom, jpeg compression and cutting the circumstances, would not affect the extracted watermark; for rotation attack, as long as the pre-correction to restore the map data can also be extracted correctly Watermark information. This article watermarking algorithm is robust, indicating that the algorithm is feasible and effective. Table 54.1

54.4 Conclusion

The existing watermark embedding process is too complex, and computationally expensive is too long, so that is not conducive to practical application. In this paper, for the characteristics of the grid map, by extracting the histogram feature from wavelet coefficients of each block, grid map watermarking algorithm based

on DWT transform was proposed. Our algorithm does not make any alteration in the original map, and at the same time, ensure that the watermark can be correctly extracted, but also for contrasting adjustment, rotate, jpeg compression and scaling attacks with strong robustness. This algorithm has good invisibility, can overcome the existing reversible watermarking security vulnerabilities.

Acknowledgments This research was supported by the Science and Technology Planning Project of Zhejiang Province (Grant No.2009C11034 and No.2008C13080) and Graduate Innovative Research Foundation of Zhejiang Gongshang University (GrantNo. 1130XJ1510127) as well as the Xinmiao Project of Zhejiang Province.

References

1. Xun W, Xia-Jun Z, Hu-Jun B (2006) Complementary watermarking algorithm for digital grid map. *J Zhejiang Univ (Eng Sci)*, 40(6)
2. Hao-Jun F, Chang-Qing Z, Hui-Ning X (2009) Watermarking algorithm for digital grid map based on wavelet transformation. *Sci Sur Mapp* 34(3)
3. Jing-Jing Z, Ping Z, Kun X (2008) Fast robust watermark algorithm for digital raster maps. *Comput Eng*, 34(1)
4. Quan W, Tan-Feng S, Shu-Xun W (2003) Concept and application of zero-watermark. *Acta Electron Sinica*, 31(2)
5. Bing H, Xuan W (2009) Zero digital watermark algorithm based on image confusion and wavelet transforms. *Comput Eng Des*, 30(1)
6. Xin-Ke G, Xian-Qing M (2008) Image digital watermarking technology based on image-scrambling and dwt. *Sci Technol Eng* 8(19)
7. Juan Y, Dan Y, Zhang X (2009) Research on second generation watermarking algorithm based on image corner feature. *J Imag Graph*, 14(5)
8. Ma J, He J, (2007) A wavelet-based method of zero-watermark. *J Imag Graph*, 12(4)
9. Yang Z (2009) A digital image zero-watermarking algorithm based on DWT domain. *Comput Security* (7)
10. Zhong W, Xiong Fe Ye (2010) Geometrical zero watermarking based on feature points. *Appl Electr Technique* 36(4)
11. Zhao X, Sun J (2010), Novel zero-watermarking based on sift feature of digital image. *Appli Res Comput* 27(4)
12. Verstrepen L, Meesters T, Dams T, Doooms A, and Bardyn D, (2009) Circular spatial improved watermark embedding using a new global SIFT synchronization scheme, *Proc IEEE DSP*, 1048

Chapter 55

A Fault-Tolerance Shortest Routing Algorithm with PDPE on (n, k) -Star Graph for NoC

Jianguo Xing and Ming Feng

Abstract This chapter presents a fault-tolerance shortest routing algorithm with PDPE which is based on the priority of permutation cycle elements. By using the priority rule of permutation cycle elements, we get a priority table to help to select a deterministic shortest route. We stimulate three groups of Network-on-Chip (NoC), they are based on different scalable (n, k) -star graph, and we set several different numbers of fault nodes for each NoC. By comparing PDPE algorithm with DFS, Floyd–Warshall, PSV algorithms, the experimental data shows that PDPE algorithm is optimal, and is the most suitable fault-tolerance routing algorithm for NoC based on (n, k) -star graph.

Keywords Network-on-chip · (n, k) -star · Routing algorithm · Fault-tolerance

55.1 Introduction

As System-on-Chip (SoC) design complexity increases fast, manufacturing faults, crosstalk interferes of transmission line and random bit-flip errors also increase. So fault-tolerance is not a feature but a requirement for designing systems that operate in the presence of all kinds of failures. Recently, Network-on-Chip (NoC) is proposed as a novel and practical approach to replace traditional interconnect like dedicated point-to-point hardwiring, shared bus or segmented buses with bridges

J. Xing · M. Feng (✉)
College of Computer Science and Information Engineering, Zhejiang Gongshang
University, Hangzhou 310035, China
e-mail: fengming19861986@gmail.com

J. Xing
e-mail: jgxing.hz@gmail.com

[1]. Borrowing lessons learned from large-scale network, the NoC design connects IP blocks by a packet-based on-chip network. With redundant links and routers, NoC provides some kind of fault tolerance capability. However, implementing traditional fault tolerant algorithms in NoC is infeasible due to chip area and energy consumption restrictions. Some stochastic routing algorithms have been proposed to overcome random errors [2]. Those algorithms need not to maintain large lookup table and require very few communication overhead. Flooding is another possible fault tolerant solution. However, under smaller number of faults, such schemes will increase unnecessary packet transmissions. Pirretti et al. [1] compared several probabilistic flooding schemes' fault-tolerance performance, chip area and energy consumption. Those probabilistic flooding algorithms works well for most time but cannot guarantee the required Quality-of-Service (QoS) which is very crucial for reliable on-chip communication.

In this chapter, we present a deterministic fault tolerant routing algorithm for (n, k) -star graph [3]. (n, k) -star graph is a generalized star graph [2], which has many advantages over the mesh and hypercube, such as lower degree and smaller diameter for network interconnection. By using the priority rules of permutation cycle elements, we can construct a priority table to help to select a deterministic shortest route under different link or router failures. If there exists one or more paths to the destination, such algorithm will choose a shortest route. The worst case transmission latency can be chosen as a design parameter of (n, k) -star graph. So the QoS requirement can be guaranteed. The overhead of constructing and look-upping of this priority table is relatively small.

This chapter is organized as follows: the first part is introduction; the second part gives the definition of (n, k) -star graph and several routing algorithms; the third section describes the proposed fault-tolerance shortest routing algorithm with PDPE; Simulation results and comparison are given in section four; the last section is conclusion.

55.2 (n, k) -Star Graph and Routing

55.2.1 Definitions and Properties of (n, k) -Star

(n, k) -star graph is firstly proposed by Chiang and Chen in 1995, which has some excellent topology properties for network interconnection. To describe conveniently, we denoted $\{1, 2, \dots, n\}$ as $\langle n \rangle$.

Definition 1 (n, k) -star graph, denoted by $S_{n,k}$, is specified by two integers n and k , where $1 \leq k < n$. The node set of $S_{n,k}$ is the set of all k -permutations of n , denoted by $\langle V \rangle = \{p_1 p_2 \dots p_i \in \langle 1, 2, \dots, n \rangle \text{ and } p_i \neq p_j\}$ [3].

Definition 2 A node u of $S_{n,k}$ is denoted by label $u_1 u_2 \dots u_k$, where $u_i \in \langle n \rangle$, $u_i \neq u_j$, $1 \leq i, j \leq k$ and $i \neq j$.

In node u , the symbols belonging to u are called internal symbols, and the others are called external symbols.

Definition 3 For node u and v , a cycle is defined as $\lambda = (s_1, s_2, \dots, s_l)$, where $\lambda \subset u$ and l is the length of the cycle.

If all symbols in λ are internal symbols of v , λ is an internal cycle and denote it with a pair of parentheses, (s_1, s_2, \dots, s_l) . If λ has an external symbol of v , λ is an external cycle. We denote it with a pair of angle brackets, $\langle s_1, s_2, \dots, s_l \rangle$ [4].

In $S_{n,k}$, the neighbors of u on the shortest paths to v are called preferred neighbors; the neighbors of u with the same distance to v are called semi-preferred neighbors; the neighbors of u with distance $d(u, v) + 1$ to v are called non-preferred neighbor.

For example, in $S_{4,2}$, for routing from node 14 to 23, node 34 is a preferred neighbor, nodes 41 and 24 are semi-preferred neighbors; for routing from 12 to 43, node 32 is a preferred neighbor, node 42 is a semi-preferred neighbor, node 21 is a non-preferred neighbor.

55.2.2 Routing

A routing algorithm can make the decision based on one of the three types of information, local information, global information and limited global information. There has been three routing algorithm based on different type of information. They are DFS [2], Floyd–Warshall, and PSV [5] algorithms. But these algorithms have drawbacks. DFS algorithm needs local information, but with the network scale and bad nodes increasing, its performance declined rapidly. Floyd–Warshall algorithm needs global information; it spends a lot of communication overhead, so it cannot get high efficiency. PSV algorithm which requires limited global information must update PSV by communicating with other nodes, which occupied bandwidth, increased power consumption and delay of transmission. We present a fault-tolerance shortest routing algorithm based on the priority of destination node’s permutation elements (PDPE); it doesn’t need any information of other nodes. Because it doesn’t need external communication, it can get low delay and high throughput.

55.3 Fault-Tolerance Shortest Routing Algorithm with PDPE

55.3.1 Foundation of PDPE Algorithm

We can get a routing algorithm based on permutation cycles in a faultless $S_{n,k}$. Node $a = a_0a_1\dots a_{k-1}$ to $b = b_0b_1\dots b_{k-1}$, we use three rules as follow:

- (1) If $a_0 = b_i$, $i \neq 0$, then exchange a_0 with the i th element in node a .
- (2) If $a_0 = b_0$, then exchange a_0 with any other element in a node whose position is not same as in node b .
- (3) If an external element of some external cycle is at the first position, then exchange it with the element of external cycle which is not at the correct position.

However, there exists a problem. We may get several routes by above algorithm. In these routes, not all of them are the shortest path. For example, in $S_{4,2}$, node 12 to 43, we can get two routes as follow: ①12→43: 12→42→24→34→43; ②12→43: 12→32→23→43.

Obviously, the algorithm can not get a deterministic route. Therefore, we improve the routing algorithm above. We add a mapping rule table of permutation elements of destination node to get a deterministic route. Precisely, we sort of elements of external cycles of destination node, and establish a mapping table.

55.3.2 The Sorting Rule of Priority of Permutation Elements

Multiple paths and non-shortest path are caused by uncertainty exchange of the external elements. Therefore, we must ensure that the exchange occurs between a node and its preferred neighbor. Generally speaking, preferred neighbor has the highest priority, and then the semi-preferred neighbor, the last is non-preferred neighbor. The priority is divided into k levels.

Sorting rule. If node u has no external element in destination node v , we do nothing; otherwise, according to the position of external element of node u in node v , we define priority value of node's elements. We begin from the first element of node v to the end, if it is an external element of node u , we record the element in priority table, meanwhile, give a value mapping to the element; then continue to the next element, if the next element is an external element of node u , we add it to the priority table, at the same time, give it a value which is the value of previous element in the table plus one; else go to the next element.

For example: for node 1,246 and 2,378, we can obtain a priority table, as shown in Table 55.1.

55.3.3 PDPE Algorithm

In a fault $S_{n,k}$, the optimal route from $a = a_0a_1\dots a_{k-1}$ to $b = b_0b_1\dots b_{k-1}$ can be obtained by repeatedly using the following three rules.

- (1) If $a_0 = b_i$, $i \neq 0$, then exchange a_0 with the i th element in node a ; if the node after exchanging is a bad node, then exchange a_0 with any element of internal cycle which does not has a_0 .

Table 55.1 Priority table of node 1,246 to 2,378 in $S_{8,4}$

Element	3	7	8
Position	1	2	3
Priority	1	2	3

- (2) If $a_0 = b_0$, then exchange a_0 with any other element in a node whose position is not same as in node b .
- (3) If an external element of some external cycle is at the first position, then exchange it with the element of external cycle which has the highest priority; if the exchanged node is a bad node, exchange it with the element of external cycle which has the second highest priority. If there is no external cycle element for exchanging, exchange the first element with the element of internal cycle which is not at the correct position, if it is unsuccessful, the links is bad, and stop.

55.3.4 Fault-Tolerance Analysis of PDPE Algorithm

In order to analyze the fault tolerance of PDPE algorithm, we will use the route of node 12 to 43 as an example.

Firstly, we analyze the routing path of node 12 to 43 in faultless $S_{4,2}$. And then we give the computation of routing path of node 12 to 43 in the situation of existing bad nodes. We analyze 3 situations, namely routing with one bad node, two bad nodes and three bad nodes.

Routing in faultless $S_{4,2}$. As Fig. 55.1a shows, we use the routing algorithm in 3.1 and get two paths, but by using PDPE algorithm, we get a deterministic shortest path: 12→43: 12→32→23→43.

One bad node. When there exists one bad node, as shown in Fig. 55.1b, according to the algorithm, we get the shortest path: 12→43: 12→42→24→34→43.

Two bad node. When there are two bad nodes, node 32 and 42 which is shown in Fig. 55.1c, the route is: 12→43: 12→21→31→13→43 and the length of the path is the optimal path length pulsing one.

Three bad nodes. When there are three bad nodes: node 31, 32 and 42, which is shown in Fig. 55.1d. Node 21 which is non-preferred neighbor is chosen, and the route is: 12→43: 12→21→41→14→34→43.

55.4 Simulation

We take use of SystemC [6] to simulate NoC based on (n, k) -star graph. Four routing algorithms (DFS, Floyd–Warshall, PSV and PDPE) are applied for NoC, we will compare their performance in different scale NoC. We have three groups

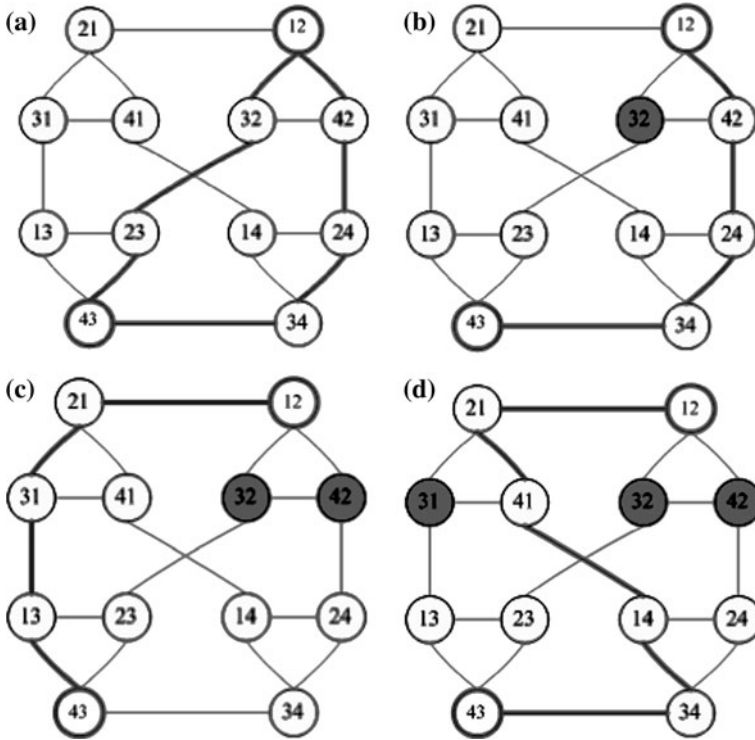


Fig. 55.1 Four kinds of routing path analysis in $S_{4,2}$. **a** is routing in faultless $S_{4,2}$. **b** is routing with one bad node. **c** is routing with two bad nodes. **d** is routing with three bad nodes

of (n, k) -star graph ($S_{5,3}, S_{6,3}, S_{7,4}$) simulated, the sizes of 3 groups network are increasing, so we can get the efficiency and performance comparison of NoC when the network is expanding. The evaluation indicators of algorithm performance are: throughput, delay and energy consumption. The reason of choosing energy consumption is that power consumption is an important requirement in NoC.

In simulation, in order to analyze fault-tolerance of algorithms, we set different numbers of bad nodes for $S_{5,3}, S_{6,3}$ and $S_{7,4}$, as Table 55.2 shows. Each case of NoC with different number of bad nodes will be tested 20 times; at last, we compute and get the average data to record.

As Fig. 55.2 shows, at the same scale of network, PDPE gets the best throughput performance and has minimum energy consumption; Floyd-Warshall algorithm gets the smallest throughput, but has the largest energy consumption. Because it needs status information of the whole network, it will spend some bandwidth, increase latency and energy consumption. PSV and DFS algorithm also need to collect information of neighbors, but they cost less than Floyd-Warshall algorithm and more than PDPE. However, PDPE doesn't need external

Table 55.2 Number of nodes and fault nodes comparison table of $S_{5,3}$, $S_{6,3}$, and $S_{7,4}$

(n, k)-star	$S_{5,3}$	$S_{6,3}$	$S_{7,4}$
Number of nodes	60	120	840
Number of fault nodes	0 5 10 20 30	0 5 10 15 20 25 30 35 40	0 20 40 60 100 150 200

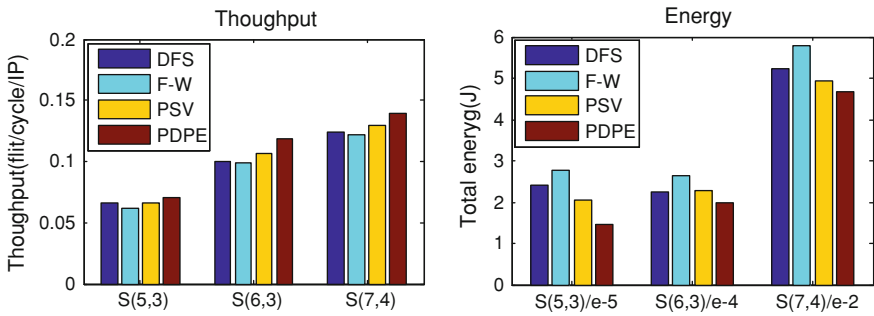


Fig. 55.2 Average throughput and total energy consumption comparison of four algorithms in fault $S_{5,3}$, $S_{6,3}$ and $S_{7,4}$

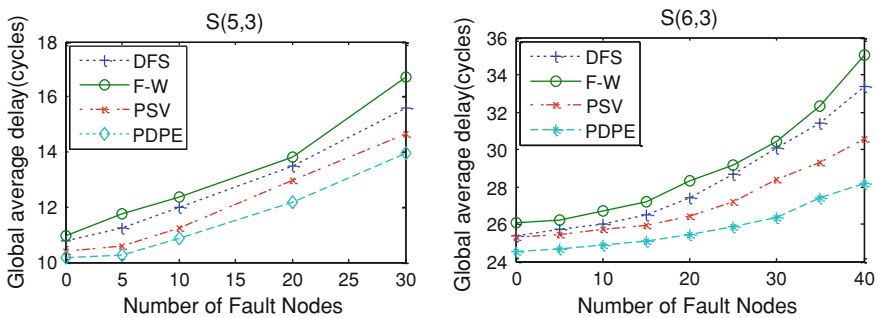
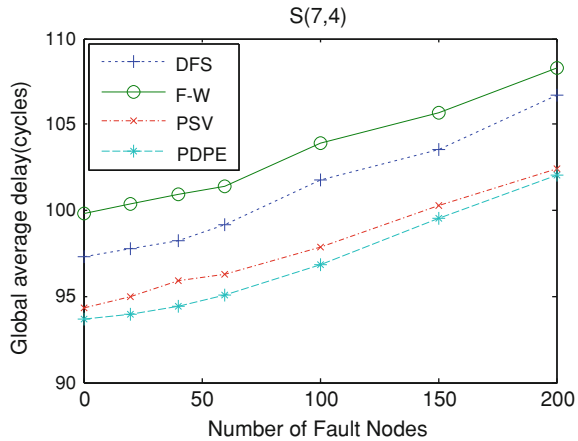


Fig. 55.3 Global average delay comparison of four algorithms in $S_{5,3}$ and $S_{6,3}$

information, it communicates with neighbors directly. So PDPE gets the best throughput, and costs lowest energy consumption.

Comparing Figs. 55.3 and 55.4, we can see PDPE algorithm gets the smallest delay, and the growth trend is smaller than other three algorithms. It doesn't need status information by communicating with other nodes, but other three algorithms need. DFS needs information of neighbors; PSV needs information to create PSV table, and when bad nodes present, the table must be updated, which cost much time. Floyd-Warshall needs information of whole network, the cost is the biggest. Therefore, PDPE spends less time on routing, the delay is smallest. With the

Fig. 55.4 Global average delay comparison of four algorithms in $S_{7,4}$



number of bad nodes increasing, the growth trend of delay of PDPE is also the smallest, it is important in large scale network.

55.5 Conclusion

In this chapter, we proposed a modified fault-tolerance (n, k) -star graph routing algorithm which is based on the permutation cycle by adding a priority table. The priority table is constructed on the priority rule of destination permutation elements; which is used to get a deterministic shortest route under some fault nodes/links. By comparing, we simulated three groups of different scale (n, k) -star, setting different number of bad nodes for each group. By comparing, we get a conclusion that PDPE algorithm is the most appropriate algorithm for NoC based on (n, k) -star. Further research is needed on the power and computing efficiency of the proposed PDPE under different pay loads.

References

1. Pirretti M, Link GM, Brooks RR, Vijaykrishnan N, Kandemir M, Irwin MJ (2004) Fault tolerant algorithms for network-on-chip interconnect. In: VLSI, IEEE computer society annual symposium on, p 46. IEEE computer society annual symposium on VLSI: emerging trends in VLSI systems design (ISVLSI'04)
2. Yeh S-I (2000) Fault-tolerant routing on the star graph with safety vectors. Master's thesis, Department of Applied Mathematics, National Sun Yat-sen University
3. Chiang WK, Chen RJ (1995) The (n, k) -star graph: a generalized star graph. *Inf Process Lett* 56(5):259–264
4. Chiang WK, Chen RJ (1998) Topological properties of the (n, k) -star graph. *Int J Found Comput Sci* 9(2):235–248

5. Hsu HC, Hsieh YL, Tan JM, Hsu LH (2003) Fault hamiltonicity and fault hamiltonian connectivity of the (n, k) -star graphs. *Networks* 42:189–201
6. Wang L, Subrammanian S, Latifi S, Srimani PK (2006) Distance distribution of nodes in star graphs, *Appl Math Lett* 19(8):780–784

Chapter 56

A Robust Zero-Watermarking Algorithm for 2D Vector Digital Maps

W. Xun, H. Ding-jun and Z. Zhi-yong

Abstract This chapter presented a new zero-watermarking algorithm for 2D vector digital mapping. The watermark information is constructed by utilizing the original data's characteristics. We divide the map into rings by using concentric circles and count the number of vertices in each ring, which is the feature information. To construct a watermark image which by using rings data and copyright information. Experiments show that the watermarks are resilient to translation, scaling, vertex deletion and growth, rotation and cropping. This is a robust algorithm.

Keywords Vector map · Zero-watermarking · Vertex · Concentric circles

56.1 Introduction

With the extensive use of Geographic Information System (GIS), vector digital map has been widely adopted in our social life, such as car navigation systems, web-based map services, and geographical information systems for city planning. Just as other digital products, vector digital map is easy to update, duplicate, and

W. Xun (✉) · H. Ding-jun · Z. Zhi-yong

College of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou, China
e-mail: wx@mail.zjgsu.edu.cn

H. Ding-jun
e-mail: huangdingjun@163.com

Z. Zhi-yong
e-mail: zzy@mail.zjgsu.edu.cn

distribute. It is also prone to be illegally duplicated and embezzled [1, 2]. With the prevailing trend of highly developed internet technique, it is extremely urgent to find out how to protect the copyright of vector digital mapping effectively and prevent it from being illegally embezzled.

Watermarking technique can be applied to protect the copyright of vector map. At present, the digital watermarking technique on vector map are obtained by modifying map data, such as spatial watermarking algorithm [3–5] and frequency-domain watermarking [6–13]. This direct or indirect modification of map data will affect the accuracy. And unpredictable attacks may lead it to lose efficacy. What's more, the algorithm complexity used to guarantee the visual effect of watermark will increase explosively. So even if we can extract the watermark correctly, it can't be applied into practice. To meet the high precision and topology invariance requirements of vector data, it seems that the traditional algorithms of digital watermark are more and more incapability. As a new digital watermark algorithm, zero-watermarking [14] can resolve conflict between imperceptibility and robustness of digital watermark because it does not change the data to construct watermark information. It is a natural blind watermark algorithm. However, at present, very few researches have been done on zero-watermarking algorithm [15, 16] in the area of geographical spatial data. Therefore, in this chapter, a new zero-watermarking algorithm for 2D vector digital maps is proposed. First, we divide the map into grids by using concentric circles and count the number of vertices in each grid, which is the feature information. And then we construct a watermark image which using grids data and copyright information. Experimental results show that the proposed algorithm is resilient to random noise, general geometric transformation and cropping. And it can resolve conflict between imperceptibility and robustness of traditional watermarking algorithm.

The remaining part of this chapter is organized as follows: Sect. 56.2 presents the main idea of our zero-watermarking scheme; Sect. 56.3 presents the experiment results to demonstrate our ideas; Sect. 56.4 gives the conclusions and future work.

56.2 The Watermarking Algorithm

Algorithm [15] presented a zero-watermarking scheme for vector data. The watermark signal is constructed by using the vector data's topology characteristics. And algorithm [16] also presented a zero-watermarking algorithm for 2D vector maps. To construct zero-watermarking that uses the vertex distribution statistics of vector digital map which is divided into several grids.

Considering the characteristics of vector map, various data objects are composed of vertex. As we know, vector data might suffer coordinate transformation, projection transformation, data format transformation, data compression, and other operations. Although those operations changed the data information or structure, they do not influence the visual content of the map. Therefore, we use the

distribution of vertex coordinates to construct zero-watermark information. This method will prevent destroying the watermark information when the original vector map is attacked by the general geometric transformation. In addition, in order to make the watermark resilient to rotation, the vector digital map is divided into several parts according to a statistical method of concentric circles. The main idea of the algorithm is summarized as follows.

First, we calculate a mean coordinate point called *AvePoint*, and mean distance value called *AveDistance*, using all vertices of the vector map. The formulas of calculating the *AvePoint* and *AveDistance* are:

$$\bar{x} = \sum_{i=1}^m x_i. \quad (56.1)$$

$$\bar{y} = \sum_{i=1}^m y_i. \quad (56.2)$$

$$AveDistance = \frac{\sum_{i=1}^m \sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2}}{m}. \quad (56.3)$$

Where (\bar{x}, \bar{y}) means the *AvePoint*'s coordinate value and m means the number of vector map vertices.

Second, we design a circle. The center of circle is *AvePoint* and radius of circle is 2 times of *AveDistance*. Then, we divide this circle into a number of rings by using concentric circles method. The length of rings is l , which is the binary image size.

$$l = N \times N. \quad (56.4)$$

Where N is the width or height of binary image.

At last, we count the number of vertices in each ring as the vector map's characteristics, and then we construct a watermark image using characteristics data and copyright information. In watermark extraction, we use the reverse process of watermarking construction to detect the copyright authentication.

In order to increase the secrecy of the watermark information, we do the original watermark image scrambling transformation before with the construction of zero-watermark. A commonly used scrambling method is Arnold transform, which is also called cat face scrambling [17].

The following are the specific steps of the watermark embedding and extraction algorithm.

56.2.1 Zero-Watermark Generation

Zero-watermarking generation algorithm based on the original vector map's features. The specific steps of the algorithm are described as follows:

Step 1. To generate scrambled copyright bits. The copyright bits are from a binary image (Watermark image), using secret key and Arnold transform to modulate the copyright bits, called binary bit stream w : if the binary image size is $N \times N$, then the length of watermark bits is $l = N \times N$. Arnold transform and key are used to generate and integrate into the final watermark bit stream, $w = (w_1, w_2, \dots, w_i, \dots, w_l)$, where $w_i \in \{0, 1\}, 0 \leq i < l$;

Step2. To generate zero-watermark information. According to the length of binary image, the vector map is divided into several rings by using concentric circles method and then count the number of vertices in each ring, which is treated as the feature messages. We do XOR operation between feature messages and scrambled copyright bits to generate zero-watermark information. At last, zero-watermark is registered into IPR depository. The specific steps are:

Using formula (56.1), (56.2), (56.3) to calculate the values of *AvePoint* and *AveDistance*.

To design a circle. The center of circle is *AvePoint* and radius of circle is 2 times of *AveDistance*. Then, we divide this circle into a number of rings by using concentric circles method. The length of rings is l , seen from formula (56.4), so the interval from each ring is $\frac{2 \times AveDistance}{l}$.

To count the number of vertices for each ring. We traverse each vertex in vector map to calculate the distance between the current vertex and the center point-*AvePoint*, and then we make classification of each vertex which belongs to its ring in order to calculate the number of vertices in each ring. The numbers are stored in the array M . M is a one-dimensional array, the length is l .

To generate zero-watermark information using formula (56.5) as follow. I is the zero-watermark information;

$$I = M \oplus W. \quad (56.5)$$

Step 3. Zero-watermark registration. Zero-watermark is registered into IPR depository. The producer of original data takes his data production and zero-watermark generated to authority organ, and if the data and watermark pass the censor, they will be registered and published for notarization efficacy.

56.2.2 Zero-Watermark Extraction

Zero-watermark of the data can be extracted by watermark extraction method. The watermark extracted will be compared with the registered watermark, and the comparability of them can be determined. Therefore the copyright of the data might be affirmed and the evidence could be provided. The specific steps of the algorithm are described as follows:

Step 1. According to the detect vector map, I find the zero-watermark I from IPR depository.

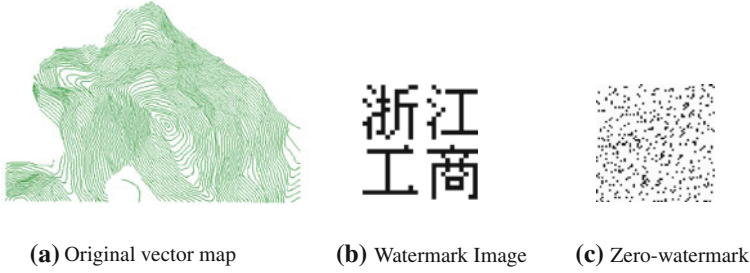


Fig. 56.1 The result of zero-watermark generation. **a** Original vector map. **b** Watermark Image. **c** Zero-watermark

Step2. To generate scrambled copyright bits. According to the length of binary image, the vector map is divided into several rings by using method described in Sect. 56.2.2. The specific steps are:

Using formula (56.1), (56.2) (56.3) to calculate the values of AvePoint and AveDistance.

To divide rings. The center of circle is AvePoint and radius of circle is 2 times of AveDistance. Then, we divide this circle into a number of rings by using concentric circles method. The length of rings is l , see from formula (56.4), the interval from each ring is $\frac{2 \times AveDistance}{l}$.

To count the number of vertices for each ring. The new numbers are stored in the array M' .

To generate watermark image copyright bits using formula (56.6) as follow. W' is the zero-watermark information;

$$W' = M' \oplus I \quad (56.6)$$

Step 3. Copyright bits are generated as a binary image. Before generating watermark binary image, inverse Arnold transform operations are performed and the watermark bits extraction is encrypted with the secret key as described in Sect. 56.2.1.

56.3 Experiments and Results

In this experiment, the test original vector map is shown as Fig. 56.1a; the copyright watermark image is shown as Fig. 56.1b; it is a 32×32 binary image, and the zero-watermark image is shown as Fig. 56.1c, it's generated by test original vector map Fig. 56.1a and watermark image Fig. 56.1b.

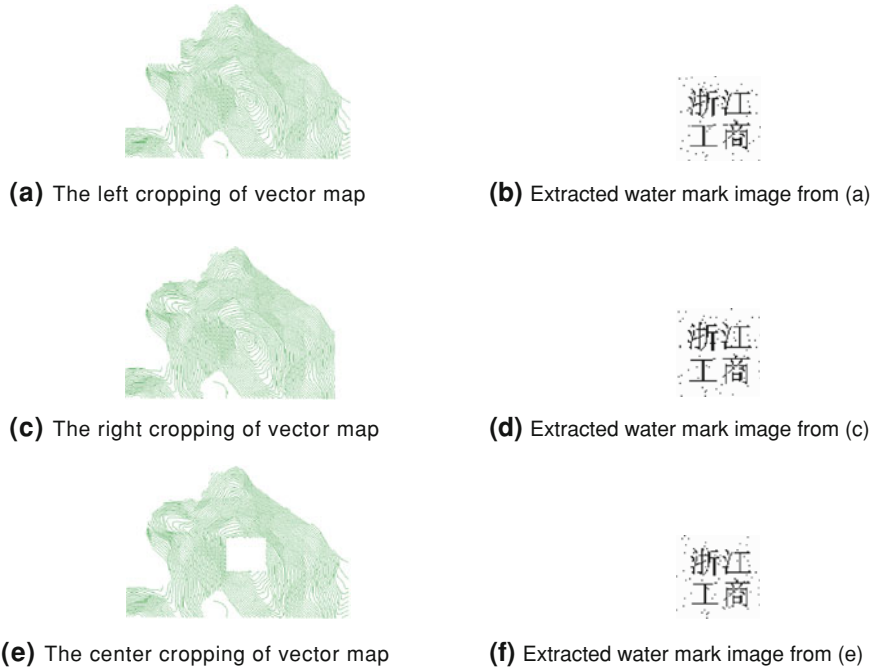


Fig. 56.2 The results of extracted watermark image by cropping attack. **a** The left cropping of vector map. **b** Extracted watermark image from (a) **c** The right cropping of vector map. **d** Extracted watermark image from (c). **e** The center cropping of vector map. **f** Extracted watermark image from (e)

To test the robustness of the proposed algorithm, we make the following five attacks to test vector digital map: (1) translation; (2) scaling; (3) rotation; (4) increase or decrease the vertex; (5) cropping.

For (1) attack, all of the vertices are translated to the direction of X and Y or opposite direction respectively; for (2) attack, the region of vector digital map is modified with the different scale; for (3) attack, vector digital map is rotated according to the different angles. Because of the vector digital maps' own characteristics and our algorithm only using the coordinates of vector digital maps to calculate the number of vertices, the operation of coordinate translation, scaling or rotation will not affect the number of vertices of the rings, so the above three kinds of attacks have no effect on watermark detection, and the watermark image's similarity is 1.

For (4) attack, due to increase or decrease the number of vertices, it will directly change the number of count-matrix M in the vector digital map, thus affecting the XOR operation when extracting the watermark information from the zero-watermarking.

Table 56.1 Similarity analysis of the results of watermark detection

Attack mode	Similarity (NC)		
Translation	1.0		
scaling	1.0		
Rotation	1.0		
Increase random vertices (p)	0.9998 (p = 100)	0.9979 (p = 200)	0.9938 (p = 300)
Delete random vertices (q)	0.9958 (q = 100)	0.9913 (q = 200)	0.9856 (q = 300)
cropping (a)	0.9836		
cropping (c)	0.9826		
cropping (e)	0.9832		

For (5) attack, we are going to cropping the vector digital maps. Figure 56.2 shows the experimental results of watermark detection by cutting off the top-left corner, lower-right corner and center of vector digital map.

To check comparability with the experimental results of extracted watermark image mentioned above, we use similarity (NC) calculation formula (56.7) to calculate similarity between the extracted watermark image and original watermark image; the results are shown as Table 56.1.

$$NC = \frac{\sum_{i,j} f(i,j)f'(i,j)}{\sum_{i,j} f^2(i,j)} \quad (56.7)$$

In the formula (56.7), $f(i,j)$ and $f'(i,j)$ means the gray value of two images at (i,j) .

Experimental results show that our algorithm have no effect on the translation, scaling and rotation of vector digital map data. For the deletion and increase of vertex and cropping attacks, we can also extract the watermark under certain limits. Thus, watermarking algorithm of this chapter have the stronger robustness, it is indicating that the algorithm is feasible and effective.

56.4 Conclusions and Future Work

In this chapter, we proposed a new zero-watermarking algorithm based on characteristics of 2D vector map. First, we divide the map into grids by using concentric circles and count the number of vertices in each grid, which is the feature information. And then we construct a zero-watermark image which using grids data and copyright watermark image. Experimental results show that the proposed algorithm is resilient to translation, scaling, vertex deletion and growth, rotation and cropping. It can resolve conflict between imperceptibility and robustness of traditional watermarking algorithm.

As the original vector map data undergoing some normal operations, the data will be changed and a little different with the original one. Also the latter extracted

zero-watermark will not be same with the original one. Therefore, we want to improve the watermark algorithm in our future work to find a comparability measuring for watermark extraction and enhance the robustness of zero-watermark.

Acknowledgments This research is supported by Zhejiang Natural Science Foundation of China (Grant No. Y1101326). And our research is also supported by the Science and Technology Planning Project of Zhejiang Province (Grant No. 2009C11034 and No.2008C13080).

References

1. Cox GS, deJager G (1993) A survey of point pattern matching techniques and a new approach to point pattern recognition [A]. In : Proceedings of symposium on communication and signal processing, Lesotho, pp 243–248
2. Johnson Neil F, Zoran D, Sushil J (2000) Information hiding: steganography and watermarking-attacks and countermeasures [M]. Kluwer Academic Publishers, Portland
3. Kurihara M, Komatsu N, Arita H Watermarking Vector Digital Maps, Tokyo: Special Interest Group Report Vol. 2000, No. 36, Information Processing Society of Japan
4. Kitamura I, Kanai S, Kishinami T (2000) Watermarking vector digital map using wavelet transformation. In: Proceedings Annual Conference of the Geographical Information Systems Association (GISA), Tokyo, pp 417–421
5. Ohbuchi R, Ueda H, Endoh S (2002) Robust watermarking of vector digital maps. In: Proceedings of the IEEE conference on multimedia and expo 2002, Lausanne, Switzerland, pp 1–4
6. Solachidis V, Nikolaidis N, Pitas I (2000) Watermaking polygonal lines using fourier descriptors, ICASSP2000, Instablul Turkey, pp 5–9
7. Nikolaidis N, Pitas N, Giannoula A (2002) Watermarking of sets of polygonal lines using fusion techniques. In: Proceedings of the IEEE international conference on multimedia and expo, Lauzanne, Switzerland, pp 549–552
8. Ohbuchi R, Ueda H, Endoh S (2003) Watermarking 2D vector maps in the mesh-spectral domain. 2003 International Conference on Shape Modeling and Applications, Seoul, Korea, IEEE Computer Society, 12-16 May
9. Ohbuchi R, Mukaiyama A, Takahashi S (2002) A frequency-domain approach to watermarking 3D shapes. *Comput Graph Forum* 21(3):1–10
10. Voigt M, Yang B, Busch C (2004) Reversible watermarking of 2D-vector data[C]//Proc of the 2004 Multimedia and Security Workshop on Multimedia and Security, Germany, Magdeburg, pp 160–165
11. Zhang HL, Gao MM (2009) A Semi-fragile Digital Watermarking Algorithm for 2D Vector Graphics Tamper Localization. In: Proceedings of the 2009 international conference on multimedia information networking and security, 01: 549–552
12. Agarwal P, Prabhakaran B (2007) Robust blind watermarking mechanism for point sampled geometry. In: Proceedings of the 9th workshop on Multimedia and security, Sept 20-21, Dallas, Texas, USA, 2007, pp 175–186
13. Vallet B, Lévy B (2008) Spectral geometry processing with manifold harmonics. In: Proceedings of Eurographics, vol 27, pp 251–260,
14. Wen Q, Sun TF, Wang SX (2003) Concept and application of zero watermark. [J]ACT A Electronica Sinica 31(2):214–216
15. Li A, Lin B, Chen Y, Lü G (2008) Study on copyright authentication of GIS vector data based on zero-watermarking. The international archives of the photogrammetry, remote sensing and spatial information sciences. vol XXXVII, Part B4, Beijing

16. Zhang ZL, Sun SS, Wang YM, Zheng KB (2009) Zero-watermarking algorithm for 2D vector map. *Comput Eng Des* 30(6):1473–1475
17. Ding W, Yan WQ, Qi DX (2001) Digital image scrambling technology based on arnold transformation [J]. *J Comput-Aided Des Comput Graphics* 3(4):338–341

Chapter 57

An Efficient Non-Local Means for Image Denoising

Gang Yang, Qiegen Liu and Jianhua Luo

Abstract The non-local means algorithm proposed by Buades et al. has been proved that can get better results than conventional local denoising algorithms. However, its computational complexity is very large that makes it too slow in practical use. In this paper, we propose an efficient non-local means algorithm in order to accelerate the denoising process and get better denoised images. Experimental results demonstrate the proposed algorithm can well remove additive white Gaussian noise and preserve details with lower complexity.

Keywords Non-local means · Image denoising · Computational complexity

57.1 Introduction

Digital images are inevitably corrupted by noise. This degrades the qualities of images and influences the post-processing of these images, so image denoising is a critical issue in image processing.

There are a lot of image denoising methods like Gaussian smoothing model [1], anisotropic diffusion (AD) [2] and total variation (TV) [3]. An outstanding review of these methods can be found in [4]. More recently, a promising method called non-local means (NLM) is proposed by Buades et al. [4, 5]. Unlike conventional denoising methods, NLM takes advantage of redundancy property in normal images. It recovers a pixel through weighted average of all pixels in the image.

G. Yang · Q. Liu · J. Luo (✉)
College of Biomedical Engineering, Shanghai Jiaotong University,
Shanghai 200240, China
e-mail: jhluo@sjtu.edu.cn

The weight is based on the similarity of two patches centered on the recovered pixel and the selected pixel. It has been demonstrated that NLM algorithm outperforms other conventional local filters [4].

Although NLM achieves better denoising results, its computational complexity is very large due to the large amount of similarity calculations between patches. Since all pixels, no matter the corresponding weights are large or small, are considered, NLM has not yet attained the best restored image as it can. In Mahmoudi et al. [6], try to accelerate NLM algorithm by pre-classifying patches based on their means and gradients in order to eliminate pixels which have dissimilar neighborhoods with the recovered pixel. However, in Coupe et al. [7], propose an accelerated algorithm named fast non-local means (FNLN), they indicate that gradient is sensitive to noise level and use standard deviation instead of gradient to pre-classify patches. Whereas two patches which have close means and variances may have very different forms, thus, comparing the means of whole patches is not accuracy. In this paper, we propose an efficient non-local means (ENLM) algorithm which pre-classifies patches based on the means of four sub-patches and variances of whole patches. Experimental results demonstrate that the proposed algorithm is more efficient than original NLM and FNLN.

57.2 Method

In this section, we first briefly review non-local means algorithm. Then, the detailed analysis of pre-classification is given in the second subsection.

57.2.1 Non-Local Means

Consider a noisy image $U = \{u(i)|i \in I\}$, N_i is the patch (also called similarity window) centered at pixel i , $u(N_i)$ is a vector contains intensities of all pixels in N_i . NLM recovers a pixel as weighted average of all pixels in the image as (57.1).

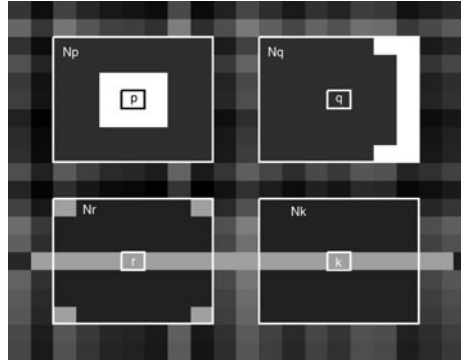
$$\tilde{u}(p) = \sum_{q \in I} w(p, q)u(q) \quad (57.1)$$

The weight $w(p, q)$ depending on the similarity between similarity windows of p and q is calculated as (57.2).

$$w(p, q) = \frac{1}{Z(p)} e^{-\frac{\|u(N_p) - u(N_q)\|_{2,\alpha}^2}{h^2}} \quad (57.2)$$

The similarity windows of pixels p and q are illustrated in Fig. 57.1. $Z(p)$ is a normalizing constant such that $\sum_q w(p, q) = 1$, h is an exponential decay

Fig. 57.1 Similarity windows of pixels



parameter depends on standard deviation of noise. Since the similarities of neighborhoods instead of the similarities of pixels are considered, NLM can preserve more details and edges than conventional local methods. This has been demonstrated in [4]. To reduce computation complexity, only pixels in the square neighborhood (larger than similarity window and also called search window) centered at recovered pixel are taken into account in practical implementation.

Considering an image of size $M \times M$, original NLM takes account of all pixels in the image when recovering a pixel, thus, its computational complexity is as high as $O(M^4)$. In practical implementation, if the size of search window is set to $m \times m$ ($m < M$), the computational complexity is reduced to $O(m^2M^2)$. Whereas this complexity is still too high in practical use that makes NLM cannot be used extensively.

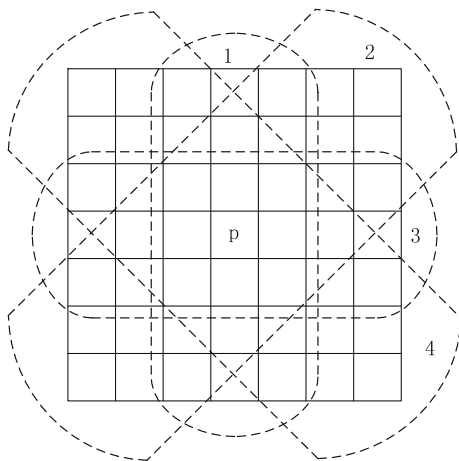
57.2.2 Pre-Classification

Original NLM algorithm calculates all weights $\{w(p, q) | q \in S_p\}$ when recovering pixel p , S_p is the search window of pixel p , thus the computational complexity is very large. To reduce the complexity, Mahmoudi et al. pre-classify the similarity windows base on means and gradients of $u(N_p)$ and $u(N_q)$ [6]. However, Coupe et al. point out that gradient is sensitive to noise level and use variance instead of gradient to pre-classify similarity windows expressed as 57.3 [7].

$$w(p, q) = \begin{cases} \frac{1}{Z(p)} e^{-\frac{\|u(N_p) - u(N_q)\|_{2,\alpha}^2}{n^2}} & \text{if } \mu_1 < \frac{u(N_p)}{u(N_q)} < \mu_2 \text{ and } \sigma_1^2 < \frac{\text{var}(u(N_p))}{\text{var}(u(N_q))} < \sigma_2^2 \\ 0 & \text{otherwise} \end{cases} \tag{57.3}$$

The main purpose of [6, 7] is to accelerate NLM algorithm by reducing the number of pixels taken into account. In other words, only the weights of pixels

Fig. 57.2 Four parts of similarity window of pixel p



which have similar neighborhoods with the recovered pixel p will be calculated. Thus, the computational complexity is reduced and NLM algorithm is speeded up. Since the pixels which are expected to have small weights are eliminated while pixels have large weights are remained, the results is comparable or even superior to that of the original NLM.

In fact, two similarity windows which have close (even equal) means may have very different forms, like N_p and N_q in Fig. 57.1. The means and variances of $u(N_p)$ and $u(N_q)$ are equal while the similarity windows of pixels p and q have very different geometries. However, pixels p and q are considered as they have similar neighborhoods and the weight $w(p, q)$ is calculated in FNLM. This is obviously wrong, since it not only increases the computational time but also introduces some disrelated information and degrades the quality of the denoised image. Meanwhile, two similar squares with one corrupted on the edges may have dissimilar means, like N_r and N_k in Fig. 57.1. The weight $w(r, k)$ is ignored and pixels r and k are considered to have dissimilar neighborhoods when recovering pixel r or k using FNLM, this will discard some useful information although it decreases computational time. In other words, FNLM may select some disrelated pixels while eliminate some similar pixels. This will degrade the quality of the recovered image.

To solve problems existed in FNLM, we divide a similarity window into four sub-patches (see Fig. 57.2). We pre-classify similarity windows base on the means of these four parts and variance of the whole similarity window as (57.4). In fact, we can consider two pixels have similar neighborhoods when one of these four sub-patches rather than the whole patches have close means and the variances of whole similarity windows are close. Since ENLM takes account of different forms of similarity window, it can more accurately eliminate dissimilar pixels while preserve similar pixels than FNLM. In other words, ENLM can discard more unrelated information while preserve more useful information, so ENLM can get

better denoised image than FNLM. At the same time, ENLM also reduces the number of weights to be calculated, therefore, it needs less time than original NLM. In conclusion, ENLM is more efficient than original NLM and FNLM.

Like pixels in Fig. 57.1, patches N_p and N_q have equal means and variances but the means of these four parts are not close. On the contrary, although means of N_r and N_k are not close, means of the first and the third parts of these two patches are close. So $w(r, k)$ will be calculated in contrast $w(p, q)$ is not. Thus, ENLM can get accurate results than FNLM.

$$w(p, q) = \begin{cases} \frac{1}{Z(p)} e^{-\frac{\|u(N_p) - u(N_q)\|_{2,a}^2}{h^2}} & \text{if } (\mu_1 < \frac{u_1(N_p)}{u_1(N_q)} < \mu_2 \text{ or } \mu_1 < \frac{u_2(N_p)}{u_2(N_q)} < \mu_2 \\ & \text{or } \mu_1 < \frac{u_3(N_p)}{u_3(N_q)} < \mu_2 \text{ or } \mu_1 < \frac{u_4(N_p)}{u_4(N_q)} < \mu_2) \\ 0 & \text{and } \sigma_1^2 < \frac{\text{var}(u(N_p))}{\text{var}(u(N_q))} < \sigma_2^2 \\ & \text{otherwise} \end{cases} \quad (57.4)$$

57.3 Experimental Results

In this section, the original NLM, FNLM and ENLM are implemented for comparison. To ensure fair comparison, we take the same parameters in all experiments, and all experiments are performed in Matlab2009a on a standard personal computer with an AMD Phenom 9650 Quad-Core Processor at 2.30 Hz, 3 GB of RAM. We compare results based on four criteria: peak signal to noise ratio (PSNR), speed, visual quality and method noise. Method noise is defined as the difference between original image and its denoised version [4]. The reference images are corrupted by additive white Gaussian noise with standard deviation σ and we set $h = 1.2\sigma$, $\mu_1 = 0.95$, $\mu_2 = 1.05$, $\sigma_1^2 = 0.5$, $\sigma_2^2 = 1.5$, search window size (21×21) and similarity window size (7×7) in all experiments.

The first experiment is designed to evaluate PSNR and the speed of original NLM, FNLM and ENLM. To get general conclusion, tests are performed on five standard testing images: “Lena”, “Barbara”, “Boat”, “House” and “Peppers”. The sizes of first three images are 512×512 and the sizes of the last two images are 256×256 . The standard deviations σ of additive white Gaussian noises are assigned as 10, 15, 25, 35, 50 and 75.

Table 57.1 shows denoised results in PSNR. We can see that ENLM achieves the highest PSNR in all standard deviations. For example, an average 0.7155 dB gain is achieved by using ENLM comparing to NLM in recovering image “Lena”. When standard deviation is too low or too high, FNLM is inferior to original NLM in most cases. Table 57.2 compares the computational speed. FNLM is the fastest one among these three algorithms, and ENLM is also faster than original NLM. For example, the average ratios of time used by FNLM and ENLM to time used by

Table 57.1 Denoised results in PSNR (in [dB])

Image	Methods	Standard deviation of noise					
		10	15	25	35	50	75
Lena	NLM	34.320	32.266	29.758	28.077	26.303	24.362
	FNLM	34.577	32.775	30.173	28.297	26.097	23.043
	ENLM	34.774	33.010	30.604	28.986	27.231	24.774
Barbara	NLM	33.283	30.945	27.654	25.565	23.724	22.079
	FNLM	33.259	31.171	28.043	25.861	23.771	21.201
	ENLM	33.526	31.430	28.348	26.263	24.418	22.451
Boat	NLM	32.020	29.997	27.369	25.734	24.151	22.645
	FNLM	32.006	30.358	27.996	26.297	24.420	21.830
	ENLM	32.412	30.716	28.413	26.868	25.251	23.317
House	NLM	34.993	33.284	30.249	27.990	25.814	23.366
	FNLM	34.994	33.553	30.861	28.729	26.265	23.043
	ENLM	35.252	33.883	31.309	29.390	27.297	24.530
Peppers	NLM	32.710	30.779	27.953	25.958	23.733	21.421
	FNLM	32.005	30.559	28.174	26.124	24.108	21.223
	ENLM	32.937	31.365	28.920	27.056	25.126	22.701

Table 57.2 Denoised results in time (in [s])

Image	Methods	Standard deviation of noise					
		10	15	25	35	50	75
Lena	NLM	676.05	680.23	678.03	676.56	681.27	682.56
	FNLM	325.91	325.32	308.03	281.20	246.22	204.33
	ENLM	422.05	460.37	478.29	477.18	446.52	406.32
Barbara	NLM	684.18	681.28	685.85	678.41	683.82	679.00
	FNLM	263.47	261.69	254.03	235.58	215.32	182.30
	ENLM	387.89	408.64	422.52	420.73	410.94	377.89
Boat	NLM	683.37	682.26	681.52	684.60	685.08	677.46
	FNLM	290.12	299.45	296.35	274.54	247.43	204.80
	ENLM	401.79	433.31	458.63	459.92	448.87	413.87
House	NLM	161.07	161.65	163.17	162.00	162.97	164.09
	FNLM	95.76	95.24	89.30	80.40	68.59	55.45
	ENLM	110.01	114.00	115.23	114.36	108.26	98.62
Peppers	NLM	162.15	161.57	161.13	163.78	161.16	161.33
	FNLM	51.45	51.50	51.76	50.17	46.35	40.83
	ENLM	71.12	76.62	82.98	84.58	86.41	79.49

original NLM are 0.4152 and 0.6604 in recovering image “Lena”. To put it briefly, ENLM can get higher PSNR using less time.

The second experiment is designed to compare visual quality and method noise. Denoised results of image “Peppers” corrupted by additive white Gaussian noise with standard deviation 30 are shown in Fig. 57.3. Figure 57.3d–f show results of

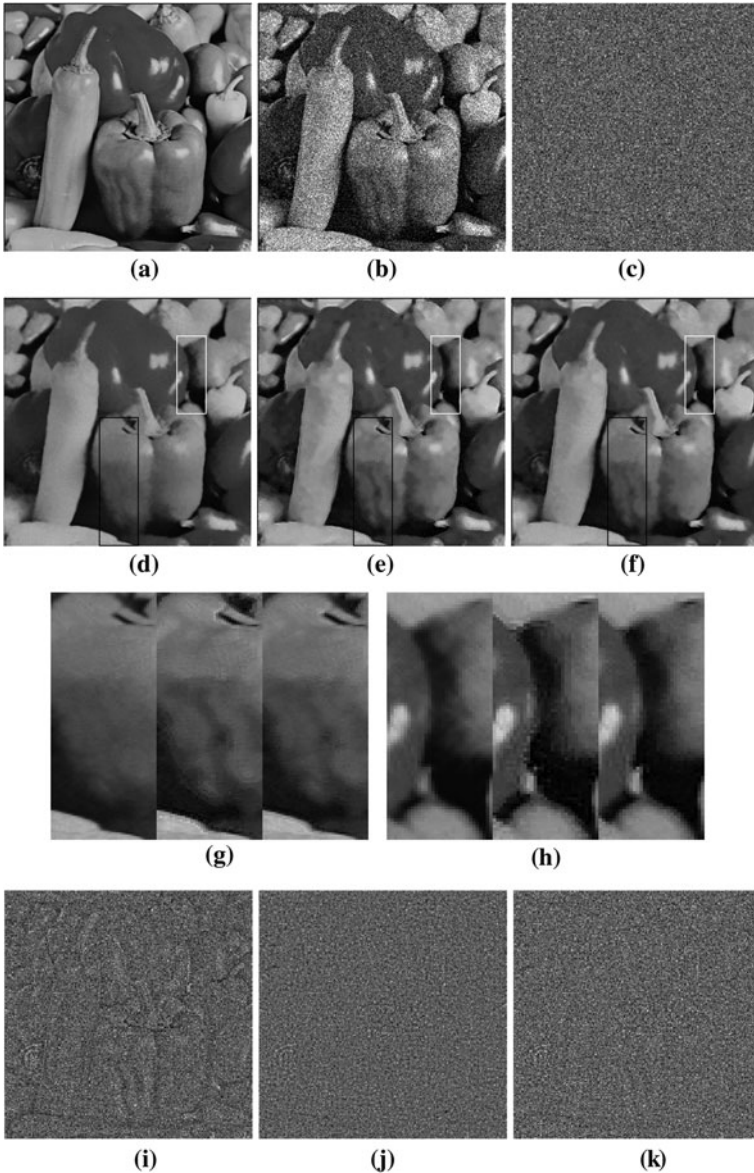


Fig. 57.3 Results of image “Peppers”: **a** original image **b** noisy image (PSNR = 18.5932 dB) **c** true noise, **d** denoised image by NLM (PSNR = 26.8991 dB) **e** denoised image by FNLM (PSNR = 27.2703 dB) **f** denoised image by ENLM (PSNR = 28.0263 dB). **g** sub-images contained in the black rectangles of **d**, **e** and **f** respectively, **h** sub-images contained in the white rectangles of **d**, **e** and **f** respectively. **i**, **j** and **k** are method noises of NLM, FNLM and ENLM respectively

original NLM, FNLM and ENLM respectively. To better assess these results, we extract two parts which are marked by white rectangles and black rectangles in standard denoised images and displayed in Fig. 57.3g and h. By comparing sub-images in Fig. 57.3g, we can see that the sub-image of original NLM is so smooth that it loses details contained in original image. Also, we can see that the result of FNLM still contains some noises by comparing sub-images in Fig. 57.3h. Method noise of original NLM (in Fig. 57.3i) contains many details while method noises of FNLM and ENLM almost not contain information of details. In summary, ENLM can well remove additive white Gaussian noise and preserve details, on the other hand, FNLM preserves some noises and original NLM loses some information of details and edges.

57.4 Conclusion

In this paper, we propose an improved NLM filter that pre-classifies similarity windows by taking advantage of means of four sub-patches and variances of whole similarity windows. Experiments demonstrate this algorithm can well remove additive white Gaussian noise and preserve details. Based on overall consideration of four criteria: PSNR, speed, visual quality and method noise, ENLM outperforms original NLM and FNLM .

Acknowledgments This work was partly supported by Shanghai International Cooperation Grant under 06SR07109, Region Rhône-Alpes of France under the project Mira Recherche 2008, and the joint project of Chinese NSFC (under 30911130364) and French ANR 2009 (under ANR-09-BLAN-037201).

References

1. Lindenbaum M, Fischer M, Bruckstein AM (1994) On Gabor contribution to image enhancement. *Pattern Recognit* 27:1–8
2. Perpona P, Malic J (1990) Scale space and edge detection using anisotropic diffusion. *IEEE Trans Pattern Anal Mach Intell* 12:629–639
3. Rudin L, Osher S (1994) Total variation based image restoration with free local constraints. *Proc IEEE Int Conf Image Process* 1:31–35
4. Buades A, Coll B, Morel JM (2005) A review of image denoising algorithms, with a new one. *Multiscale Model Simul* 4:490–530
5. Buades A, Coll B, Morel JM (2005) A non-local algorithm for image denoising. *IEEE Comput Vis Pattern Recognit* 2:60–65
6. Mahmoudi M, Sapiro G (2005) Fast image and video denoising via non-local means of similar neighborhoods. *IEEE Signal Process Lett* 12:839–842
7. Coupe P, Yger P, Barillot C (2006) Fast non local means denoising for 3D MR images. *Med Image Comput Computer-Assisted Intervention* 4191:33–40

Chapter 58

An Efficient Augmented Lagrangian Method for Impulse Noise Removal via Learned Dictionary

Qiegen Liu, Gang Yang, Jianhua Luo and Shanshan Wang

Abstract This paper presents a novel augmented Lagrangian method for impulse noise removal via learned dictionary. We reformulate the $L1$ – $L1$ minimization into an augmented Lagrangian scheme through adding a new auxiliary variable, additionally the dictionary is updated by simply adding the multiplication of dual and primal variables. Experimental results demonstrate that the new proposed method can obtain very significantly superior performance than the current state-of-the-art variational methods for salt-and-pepper noise removal.

Keywords Salt-and-pepper noise removal · Augmented Lagrangian · Alternating direction method · Dictionary learning

58.1 Introduction

Digital images are often corrupted by impulse noise during image acquisition or transmission. An important type of impulse noise is salt-and-pepper noise, which is usually found in imaging with quick transients such as faulty switching, or transmitting images over noisy digital channels [1]. When images are contaminated by salt-and-pepper noise, only part of pixels is changed, and the noisy pixels like white and black dots sprinkled on the images. Because of its “sparse” distribution nature, impulse noise poses new challenges to traditional image denoising methods.

Q. Liu · G. Yang · J. Luo (✉) · S. Wang
College of Biomedical Engineering, Shanghai Jiaotong University, 200240
Shanghai, China
e-mail: jhluo@sjtu.edu.cn

One of the most popular methods is the standard median (SM) filter [2], the idea of which is to replace gray-level value of every pixel in the image by the median value of its neighbors. Also many other improved methods based on the SM filter are proposed, such as the directional weighted median filter (DWM) [3] and the contrast enhancement-based filter (CEF) [4]. However these filters cannot efficiently distinguish the noisy pixels from the edge and fine texture pixels when dealing with high ratio noisy images. To alleviate this drawback, some variational methods for impulse denoising have been proposed. The earliest method L1-TV model was proposed by Nikolova et al [5], in their work they combined total variation regularization with an L1-data-fitting term for removing impulse noise. Inspired by the robust property of the L1 norm, Takeda proposed a method named L1-SKR [6], which incorporates the L1 norm estimator in the steering kernel regression. In this research, we focus on variational methods and present a dictionary learning approach for impulse noise removal. In most of the current literature, sparse representation has been used successfully for the removal of the Gaussian noise with very good performance [7]. The assumption here is that the natural signals are sparse over some learned over complete dictionary. To the best of our knowledge, no dictionary learning method is applied to non-Gaussian noise directly.

In this paper, we adopt a dictionary learning based L1–L1 minimization strategy for noise removal, the outline of which is to transform the constrained dictionary learning problem into the augmented Lagrangian (AL) framework firstly and then refine the dictionary from low scale to high. AL is a standard and elementary tool in optimization field, it converges fast even super-linearly when forcing its penalty parameter updated to infinity [8]. Usually, a “decouple” strategy i.e. alternating direction method (ADM) is used to solve the subproblem of the AL scheme, it facilitates the AL to be implemented efficiently in many inverse problems [9]. In this paper we inherit the spirit of this strategy to deal with the subproblem.

58.2 AL Based Dictionary Learning Algorithm

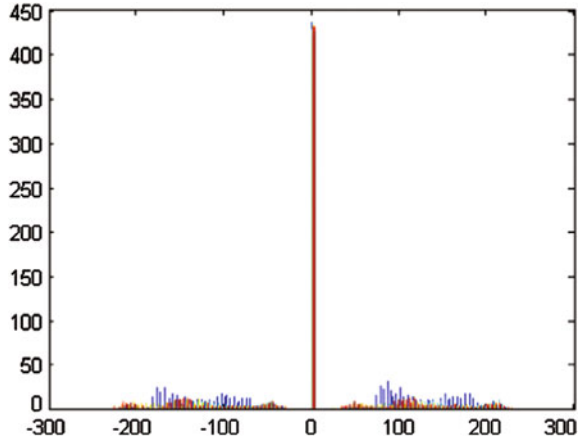
58.2.1 The Sparseness Characteristic of Salt-and-Pepper Noise

Salt-and-pepper noise appears in many images. In an image corrupted by it, only a few pixels are very noisy, the others are noise free. Suppose an image I_0 is corrupted by salt-and pepper noise, and assume that the probability of pixels corrupted is p , then the noisy image \tilde{I} is generated as follow:

$$\tilde{I} = \begin{cases} I_{\min}, & \text{with probability } p/2 \\ I_{\max}, & \text{with probability } p/2 \\ I_0, & \text{otherwise} \end{cases} \quad (58.1)$$

where $[I_{\min}, I_{\max}]$ is the gray-level region. Figure 58.1 shows the histogram of the difference between the original “Lena” image and its noisy version contaminated

Fig. 58.1 The histogram of the difference between the original “Lena” image and its noisy version contaminated by 20% salt-and-pepper noise



by $p = 20\%$. From Fig. 58.1, we can observe that salt-and-pepper noise is evidently one kind sparse distributed noise. Since the impulse type noise tends to have a heavy-tail distribution, we consider the robust L1 norm $\|\tilde{I} - I_0\|_1$ instead of $\|\tilde{I} - I_0\|_2$.

In the following context, we consider the patch-based dictionary learning problem. Suppose L image patches with size $\sqrt{M} \times \sqrt{M}$ are extracted from the image \tilde{I} , after converting every patch into one column, we obtain the data samples $B = [b_1, b_2, \dots, b_l, \dots, b_L]$. It is natural to conclude that the noise in the patch-based formation is also sparsely distributed. Consequently we use the robust L1 norm data-fidelity term in the following L1–L1 minimization problem:

$$\text{Min}_{A, X} \{ \|X\|_1 + \lambda \|B - AX\|_1 \} = \text{Min}_{A, X} \sum_{l=1}^L \{ \|x_l\|_1 + \lambda \|b_l - Ax_l\|_1 \} \quad (58.2)$$

where $X = [x_1, x_2, \dots, x_l, \dots, x_L]$. The first term in the functional favors a sparse solution, whereas the second term is the L1 data-fidelity term. The regularization parameter λ controls the balance between fitness and sparseness. The ultimate goal of the L1–L1 minimization problem (58.2) is to find the sparse representation of solution under the optimal learned dictionary and thereby obtain the recovery patches collection $\{AX\}$.

58.2.2 The General AL Based Dictionary Learning Framework

This section introduces the proposed AL–ADM–DL algorithm for solving the dictionary learning problem, it is achieved by firstly recasting the constrained

dictionary learning problem into an AL scheme, and then updating the dictionary after each inner iteration of the scheme.

Firstly we let the dictionary A fixed and consider the AL scheme for minimizing Eq. 58.2 with regard to X . By adding a new auxiliary variable, Eq. 58.2 is then converted into a different constrained problem, for the sake of clarity the subscript variable l is omitted, namely:

$$\min_{x,z} \{ \|x\|_1 + \lambda \|z\|_1 \} \quad \text{s.t.} \quad b = Ax + z \quad (58.3)$$

We apply the method of AL to solve this constrained problem, which is replaced by solving a sequence of unconstrained subproblems in which the objective function is formed by the original objective of the constrained optimization plus additional “penalty” terms, the “penalty” terms are made up of constrained functions multiplied by a positive coefficient (for more details of AL scheme, see Ref. [8]), i.e.

$$(x^{k+1}, z^{k+1}) = \arg \min_{x,z} L_\mu(x, z) \quad (58.4)$$

where $L_\mu(x, z) \triangleq \|x\|_1 + \lambda \|z\|_1 - \langle y^k, A_k x + z - b \rangle + \mu \|A_k x + z - b\|_2^2 / 2$

$$y^{k+1} = y^k + \mu(-A x^{k+1} - z^{k+1} + b) \quad (58.5)$$

where $\langle \bullet, \bullet \rangle$ denotes the usual duality product.

It would be much promising if we update dictionary A after inner iteration of Eq. 58.4–58.5. If we extend the penalty functional $L_\mu(x, z)$ in Eq. 58.4 to be $L_\mu(x, z, A)$ by introducing a new variable A . Then updating the dictionary will substantially decrease the functional. Taking the derivative of the extended functional $L_\mu(x, z, A)$ with respect to A , we get the following gradient descent update rule:

$$A_{k+1} = A_k - \mu[-Y^k + \mu(A_k X^{k+1} + Z^{k+1} - B)](X^{k+1})^T = A_k + \mu Y^{k+1} (X^{k+1})^T \quad (58.6)$$

An evident merit of the AL methodology is its superior convergence property: $A x^{k+1} \rightarrow A x^* = b^* - z^*$, $k \rightarrow +\infty$ [8], where each iterative variable “ $A x^{k+1}$ ” can be viewed as a low-resolution or smoothed version of the true image patch “ $A x^*$ ”. Suppose that each iterative step is regarded as a scale, then the dictionary updating Eq. 58.6 can be seemed as a refinement process from the low scale to high.

58.2.3 An Efficient Inner Solver

In the proposed AL based dictionary learning framework (58.4)–(58.6), it is obvious that the speed and accuracy of the proposed method depend largely on how the Eq. 58.4 over variables x and z is solved. A natural approach is to use

alternating direction method (ADM) [9], in which Eq. 58.4 is solved by alternating minimization with respect to x and z , while keeping the other variable fixed. In particular, for variable x , there is now extensive literature in optimization and numerical analysis that addresses it. To date, the most widely studied first-order method for solving the Eq. 58.4 with respect to x is Iterated Shrinkage/Thresholding Algorithm (ISTA) [10]. It iteratively performs a gradient descent update followed by a soft thresholding. The advantage of the popular ISTA is in its simplicity. It is worth noting that since the number of training samples is very big for dictionary learning problem, thus a simple and fast iteration formula is essential.

Firstly, with x fixed, the minimization of Eq. 58.4 with respect to z can be computed analytically:

$$\begin{aligned} \min_z \left\{ \lambda \|z\|_1 - \langle y^k, Ax^k + z - b \rangle + \mu \|Ax^k + z - b\|_2^2 / 2 \right\} \\ = \min_z \left\{ \lambda \|z\|_1 + \mu \|z - (-Ax^k + b + y^k / \mu)\|_2^2 / 2 \right\} \end{aligned} \quad (58.7)$$

We obtain that $z^k = \text{Shrink}((-Ax^k + b + y^k / \mu, \lambda / \mu)$.

$$\text{Where } \text{Shrink}(f, \mu) = \begin{cases} f - \mu, & f \geq \mu \\ 0, & -\mu \leq f < \mu \\ f + \mu, & f < -\mu \end{cases}$$

Secondly, with z fixed, the minimization of Eq. 58.4 with respect to x is solved by ISTA [10]. i.e.

$$\begin{aligned} x^{m+1} &= \arg \min_x \left\{ \|x\|_1 + \mu \|Ax - (b + y^k / \mu - z^k)\|_2^2 / 2 \right\} \\ &= \arg \min_x \left\{ 2 \|x\|_1 / \mu + \gamma \|x - [x^m + A^T(b + y^k / \mu - z^k - Ax^m) / \gamma]\|_2^2 \right\} \\ &= \text{Shrink}(x^m + [A^T(b + y^k / \mu - z^k - Ax^m)] / \gamma, 1 / \gamma \mu) \end{aligned} \quad (58.8)$$

where $\gamma \geq \text{eig}(A^T A)$.

To sum up, the proposed AL-ADM-DL algorithm consists of a two-level nested loop; the outer loop updates variables z and A ; meanwhile the inner loop minimizes x . Diagram 1 lists the detailed description of the algorithm, the initial dictionary A_0 in line 1 can be any predefined matrix such as the redundant DCT.

Diagram 1: THE AL-ADM-DL ALGORITHM

1. initiation: $X^0 = 0; A_0$
while stop-criterion not satisfied (loop ink):
2. $z_l^k = \text{Shrink}((-Ax_l^k + b_l + y_l^k / \mu, \lambda / \mu), l = 1, 2, \dots, L)$
while stop-criterion not satisfied (loop inm):
3. $x_l^{k,m+1} = \text{Shrink}(x_l^{k,m} + [A^T(b_l + y_l^k / \mu - z_l^k - Ax_l^{k,m})] / \gamma, 1 / \gamma \mu), l = 1, 2, \dots, L$

4. end while
5. $X^{k+1,0} = X^{k,m+1}$; $Y^{k+1} = Y^k + \mu(-AX^{k+1} - Z^{k+1} + B)$
6. $A_{k+1} = A_k + \mu C^{k+1}(X^{k+1,0})^T$
7. end while

58.3 Results

To demonstrate the effectiveness of our method, we compare our results with two of the most representative state-of-the-art algorithms: L1-SKR [6] and L1-TV¹ [11]. The L1-L1 with fixed dictionary DCT is also implemented to illustrate the strength of the learned dictionary. The L1-SKR² is initialized with L1 classical kernel regression, and the global smoothing parameter in the method is hand tuned for best improvement in PSNR. We set the number of CG iterations in L1-TV to be 8, and the regularization parameter in the method is 0.7, 0.8, 1.0 and 1.1 for $p = 20, 30, 50$ and 70% respectively. For our approach, we set $\lambda = 0.7, 0.6, 0.5$ and 0.4 for $p = 20, 30, 50$ and 70%, respectively. In the experiment, the whole process involves the following steps: (a) After the addition of salt and pepper noise, 62001 examples $\{b_1, b_2, \dots, b_{62001}\}$ of 8×8 pixels extracted from the noisy images \tilde{I} are used with a DCT based over complete dictionary for initialization. (b) In the sparse coding stage of learning procedure, each patch extracted is sparse-coded. We set $m = 2$, $\mu = 0.001$ and the iteration repeats until $k = 30$. After the learning procedure, we obtain approximate patches with reduced noise $\{\tilde{b}_1, \tilde{b}_2, \dots, \tilde{b}_{62001}\}$. (c) The output image I is obtained by adding the patches $\{\tilde{b}_1, \tilde{b}_2, \dots, \tilde{b}_{62001}\}$ in their proper locations and averaging the contributions in each pixel, the implementation is the same as in [7].

Comparisons among different denoising methods are shown in Table 58.1 for the standard test images.³ The best result obtained for a particular noise level using the various techniques is shown in boldface. From the tables, it can be observed that the proposed method outperforms other competitive schemes for the removal of salt-and-pepper noise at various noise levels, especially when the noise level p is low the gap between them is evident. The proposed algorithm over learned dictionary can preserve the details and texture efficiently. Some visual results for “Barbara” are shown on Fig. 58.2, since L1-TV prefers to the “cartoon-like” outputs, the trousers stripe parts of the “Barbara” presented in Fig. 58.2a are poorly recovered, as shown in Fig. 58.2b, this drawback is alleviated somewhat due to the fact that L1-SKR uses quadratic approximation of the image signal combined with an adaptive steering kernel. Finally, very significantly superior performance results are obtained by our AL-ADM-DL in Fig. 58.2c, the texture

¹ A software implementation is available at <http://numipad.sourceforge.net/>.

² Matlab code is available at <http://users.soe.ucsc.edu/~htakeda/KernelToolBox.htm>.

³ All these images are available at http://decsai.ugr.es/~javier/denoise/test_images/index.htm.

Table 58.1 The denoised PSNR results for four test images and four noise powers. Each test setting is divided into four parts: L1-TV (*top left*); L1-SKR (*top right*); the sparse coding with DCT (*bottom left*); and AL-ADM-DL (*bottom right*)

p	Barbara		Lena		Boat		Peppers	
20%	25.30	25.48	33.22	33.84	29.70	30.18	28.45	28.34
	30.15	33.51	33.48	35.60	30.32	32.45	28.17	31.51
30%	24.46	24.42	31.46	29.77	28.11	27.87	26.58	25.43
	27.77	30.54	31.03	33.35	28.20	29.88	25.89	28.15
50%	23.23	22.83	28.64	27.27	25.43	24.04	23.24	22.97
	24.56	25.90	27.76	29.26	25.34	26.29	23.17	24.45
70%	21.71	18.84	25.28	23.31	22.86	19.22	20.83	17.96
	22.68	22.80	25.88	26.35	23.61	23.87	21.55	21.87



Fig. 58.2 Denoised results by **a** L1-TV **b** L1-SKR and **c-d** AL-ADM-DL and the learned dictionary. $p = 20\%$

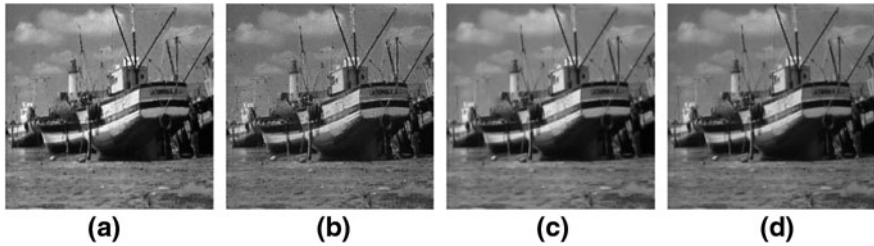


Fig. 58.3 Denoised results by **a** L1-TV **b** L1-SKR **c** the sparse coding with DCT and **d** AL-ADM-DL. $p = 30\%$

and details in “Barbara” are properly sparse represented by our L1 norm based learned dictionary (Fig. 58.2d), which is robust to outliers such as salt-and-pepper noise.

Figure 58.3 shows the outputs of various filtering methods for the “Boat” image with $p = 30\%$. It is clearly observed that the proposed method not only removes the noise but also successfully preserves the detail features in the image.

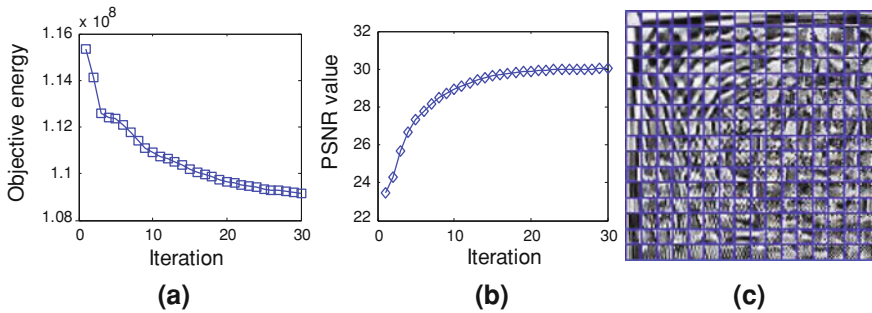


Fig. 58.4 **a** The objective function value and PSNR versus iteration k respectively. **c** The finally learned dictionary with “Boat”. $p = 30\%$

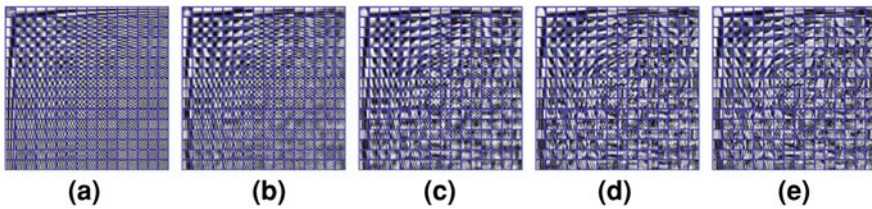


Fig. 58.5 The sequence of dictionaries generated by AL-ADM-DL after 1, 3, 7, 16, 28 iterations, with “Boat” (from left to right). $p = 30\%$

Figure 58.4 displays the objective function 58.2 and the PSNR value versus iterations k respectively. Seen from Fig. 58.4a, the AL based dictionary learning approach demonstrates a notable ability to decrease the fidelity error quickly. Accordingly PSNR value presented in Fig. 58.4b increases with the iterative procedure. We also illustrate the sequence of dictionaries generated by our AL-ADM-DL. As shown in Fig. 58.5, the dictionary sequence generated by AL-ADM-DL is more and more “regular complex”. AL-ADM-DL changes the initial dictionary drastically in the first few stages (e.g. $k = 1, 2, 3, \dots, 13$). Specially, the learned dictionaries generated at 1,3,7,16,28 iterations are depicted. We can observe that from one to seven iterations most of the edge objective atoms are constructed and furthermore from 16 to 28 iterations more and more details are added to the atoms.

58.4 Conclusion

This work has presented a novel AL method for impulse noise removal via learned dictionary. Image denoising experiments demonstrate that the proposed method provides excellent results both visually and quantitatively. Like in Ref. [12], one

can further improve our results by combining some “decision-based” median filters [3, 4] with our method. These extensions will be studied in our forthcoming papers.

Acknowledgments This work was partly supported by the joint project of Chinese NSFC (under 30911130364) and French ANR 2009 (under ANR-09-BLAN-037201).

References

1. Arce GR, Paredes JL, Mullan J (2000) Nonlinear filtering for image analysis and enhancement. In: Bovik AL (ed) Handbook of image video process. Academic Press, London
2. Pratt WK (1975) Median filtering. Image Proc. Institute, University of Southern California, Los Angeles, Technical report
3. Dong Y, Xu S (2007) A new directional weighted median filter for removal of random-valued impulse noise. IEEE Signal Process Lett 14(3):193–196
4. Ghanekar U, Singh A, Pandey R (2010) A contrast enhancement based filter for removal of random valued impulse noise. IEEE Signal Process Lett 17(1):47–50
5. Nikolova M (2004) A variational approach to remove outliers and impulse noise. J Math Imag Vis 20:99–120
6. Takeda H, Farsiu S, Milanfar P (2006) Robust kernel regression for restoration and reconstruction of images from sparse noisy data. In: ICIP
7. Elad M, Aharon M (2006) Image denoising via sparse and redundant representations over learned dictionaries. IEEE Trans Image Process 15(12):3736–3745
8. Bertsekas D (1982) Constrained optimization and lagrange multiplier method. Academic Press, London
9. Yang J, Zhang Y (2009) Alternating direction algorithms for l1 problems in compressive sensing. Technical report, Rice University
10. Yin W, Osher S, Goldfarb D, Darbon J (2008) Bregman iterative algorithms for l1-minimization with applications to compressed sensing. SIAM J Imag Sci 1:142–168
11. Rodriguez P, Wohlberg B (2009) Efficient minimization method for a generalized total variation functional. IEEE Trans Image Process 18(2):322–332
12. Chan RH, Ho C-W, Nikolova M (2005) Salt-and-pepper noise removal by median-type noise detectors and detail preserving regularization. IEEE Trans Image Process 14(10):1479–1485

Chapter 59

Motion Object Segmentation Using Regions Classification and Energy Model

Xiaokun Zhang and Xuying Zhao

Abstract An automatic video object segmentation is proposed. The video scene is partitioned into some homogeneous regions by an automatically cluster to form regions method. Then these regions are initially classified into three categories: moving object, candidate and background using the difference information between two successive frames. The spatio-temporal energy model is constructed to determine the candidate regions to moving object or background. Some post-processing methods are utilized to achieve the more accurate segmentation object. Experimental results show that the spatial accuracy of our proposed algorithm improves about 30–50% and temporal coherency improves about 0.05–0.70 than COST211 AM.

Keywords Video processing · Motion segmentation · Regions selection

59.1 Introduction

Video segmentation, which extracts the shape information of moving object from the video sequence, is a key operation for content-based video coding(MPEG-4), object-based video database querying (MPEG-7), and intelligent signal processing (computer vision). The segmentation of a video sequence into moving regions is

X. Zhang (✉) · X. Zhao (✉)
Beijing Electronic Science and Technology Institute, No.7 Fufeng Road, Fengtai
District, Beijing 100070, People's Republic of China
e-mail: sam@besti.edu.cn

X. Zhao
e-mail: xyzhao@besti.edu.cn

difficult as it is an inverse problem and, furthermore, ill-posed. To date, a variety of approaches have been addressed to solve the problem [1]. These techniques include: (1) intra-frame segmentation with inter-frame tracking, (2) motion field clustering, and (3) frame differencing. These methods often fail to produce semantically meaningful objects since noise and not enough texture. To solve the problem, several research focus on interactive object extraction (semi-automatic video object segmentation) [2]. Semi-automatic methods are flexible and relatively accurate. But human interactivity adds burden to users. And it is not suitable for real time processing system. Developing an accurate and automatic segmentation is necessary to some practical application such as conferencing systems. Region-based segmentation is more robust than pixel-based methods, but some algorithms, which were proposed in previous works [3–5], are relatively complicated. In this paper we focus attention on region-based automatic segmentation with low complexity. The basic idea of our algorithm is to classify and select regions to moving object with the spatio-temporal information.

The paper is organized as follows. In Sect. 2 some regions are achieved by a spatial segmentation. The regions are initially classified with temporal information in Sect. 3. In Sect. 4 a spatio-temporal energy model is utilized to determine which classes the candidate regions should belong to. In Sect. 5, the noise regions and ragged boundaries were eliminated by post-processing. The paper ends with experimental results in Sect. 6 followed by conclusions.

59.2 Spatial Regions Partition

There are many approaches of the spatial regions partition, such as watershed transforms [6, 7], clustering, and Bayesian methods. The watershed algorithm produces very accurate results, but it has two inherent drawbacks: it is extremely sensitive to gradient noise and usually results in over segmentation. Clustering refers to classification of pixels into regions according to certain similar properties. The biggest obstacle to overcome in clustering procedures is the determination of the correct number of classes (regions), which is assumed to be known. In practice, the value of classes is determined by trial and error. The Bayesian methods need many times of iterations to achieve a relatively stable segmentation. We proposed an approach to overcome these drawbacks of the spatial segmentation above-cited. The basic framework of the algorithm is that all pixels is used to automatically cluster to form regions(ACFR), which is not need the iteration and the numbers of regions known in advance. The ACFR algorithm utilizes the differences between the current pixel and its neighborhood to form regions. If the minimum difference in all neighborhoods is less than a predefined threshold T , the current pixel and its neighborhood pixel with the minimum difference are partitioned into the same region, otherwise redefined a new region to the current pixel. Since the order of the process of partition is scanning pixels in an image from up to down, and from left to right, In respect to the current pixel, its left, left-up, up and right-up pixels in the

8-pixel neighborhood have been defined and labeled to a certain region. While calculating the neighborhood differences, only these pixels are included. To get a better result, the values of these pixels in neighborhood are respectively substituted by the means of its own region. The details can be described as follows.

- (1) Set a fixed threshold $T_{diff} = 20$.
- (2) Calculate the minimum absolute difference between the current pixel value and the mean of four neighborhood regions: $d = \min(|f(x, y) - m(i, j)|)$ where $m(i, j)$ is the mean of a neighborhood region.
- (3) If the difference $d < T_{diff}$, merge the current pixel to the region in which there is a minimum difference, assign the same region label, and recalculate region mean.
- (4) If all neighborhood regions don't satisfy $d < T_{diff}$, a new region has appeared, and then a new region label is assigned. Since only a pixel in the region, the current pixel value is considered as the region mean.
- (5) Repeat for all pixel in a given frame image.

59.3 Initial Regions Classification

Motion information is the main clues that are used for segmenting video sequences. However, extracting and coupling motion information are inaccurate to areas with low texture. Therefore, unlike in many other motion segmentation algorithms, we do not directly determine whether each region in the initial partition is part of a moving object or part of the background, but mark it as one of three classes: moving object, candidate and background. The motion of each region is estimated by differences between two consecutive frames [8]. In order to avoid losing of slowly moving regions, the skipping frames difference may be calculated in a local sum of a window of observation centered at the current pixel. A region is classified as a moving object if more than 70% of its pixels are marked as changed, and background if less than 10%, otherwise it is marked as candidates. This initial classification serves two main purposes. First, the define of the candidate regions avoid incorrect segmentation. Second, it reduces the computational load of the following region energy estimation phase by eliminating most of the background regions.

59.4 Spatio-Temporal Energy Model for Candidate Regions Selection

The candidate regions will sequentially be selected to be moving object or background. The motion information of each candidate regions is used to define an energy model along with spatial information. The selection problem to the

candidate regions is formulated as the spatio-temporal energy and a solution is easily obtained. The energy can be expressed as $E_{temporal}$ and $E_{spatial}$, which stands for the temporal and spatial energy respectively. The temporal energy, which is caused by the change between successive frames in a sequence, is defined as the sum of pixels changed in a region.

$$E_{temporal}^i = \sum_{(x,y) \in M_i} V_1^i(x,y) \quad (59.1)$$

where i denote the label of the candidate $V_1(x,y)$ is the motion changed pixel, M is a set of pixels in the candidate region.

The spatial energy term, which expresses the spatial constraints from the neighborhood regions, is defined as follows:

$$E_{spatial}^i = \sum_{(x,y) \in M_i} V_2(x,y) \quad (59.2)$$

where $V_2(x,y)$ is the local energy term, which is defined as follows:

$$V_2(x,y) = \begin{cases} -1 & n(x,y) \in B \\ 1 & n(x,y) \in M \\ 0 & n(x,y) \in C \end{cases} \quad (59.3)$$

where $n(x,y)$ is 4-pixel neighborhood, while B , M , and C are denoted as the region of background, moving object, and candidate, respectively. If its neighborhood has not the only one of B , M , and C , the prior order should be from B to M , and to C . The total energy is defined as follows:

$$E_{total}^i = \frac{1}{2N} (E_{temporal}^i + E_{spatial}^i) \quad (59.4)$$

where N is the number of pixels in the candidate region. The total energy term of the candidate region represents the near degree to moving object or background. So it is regarded as a fundamental of the candidate regions selecting. If E_{total}^i is lower than a predefined threshold T , the region is selected to be background, otherwise it is marked as moving object. Experimentally, we have found that using a parameter T from 0.5 to 0.7, leads to good results.

59.5 Post-Processing

After the former steps, an initial binary mask, which describes the moving object with "1" and background with "0" respectively, is generated. Due to the camera noise and irregular object motion, there exist some noise regions and not smooth

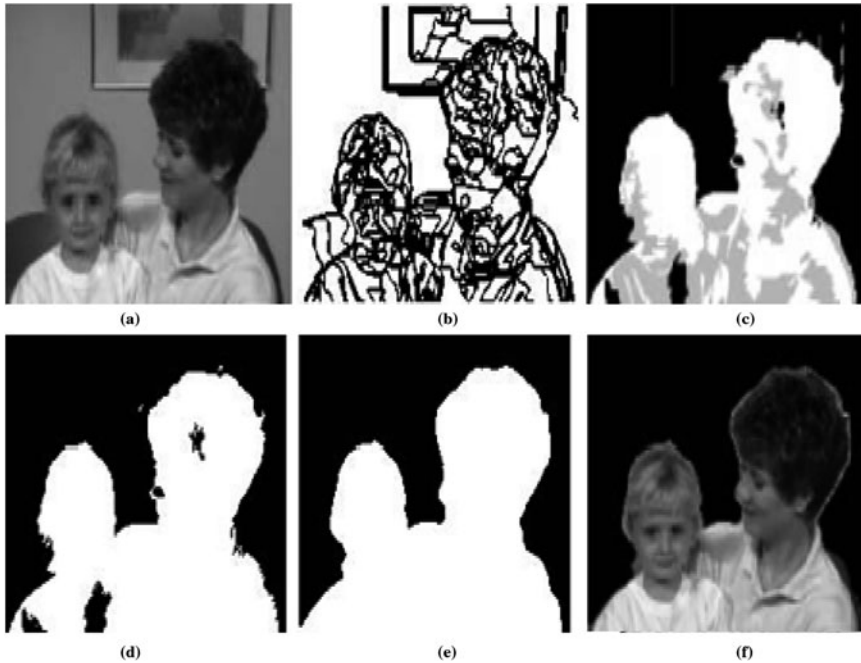


Fig. 59.1 Segmentation process on the first frame of test sequence. Mother & daughter. **a** original frame **b** regions partition **c** initial classification **d** energy selection **e** post processing **f** segmentation object

boundary in the initial object mask. Therefore, a post processing is necessary. A traditional way to remove the noise regions is using the morphological operations to filter out smaller regions. However, it may degrade the precision of the object boundary. We utilize three new techniques to solve the problem: (1) spatial grad compensating, (2) Markov processing, and (3) bigger region labeling and filling. In spatial partition the means of the neighborhood regions is applied to blur some faint edge, and then ragged boundary is formed in mask. To reduce errors of the mask, the boundary pixels are filled with “1” where there is a higher gradient value than around others pixels. Markov and Gibbs random field (MRF/GRF) are used to eliminate the noise region. The binary mask is regarded as the initial label field and the observer fields. The final label field is obtained by maximization a posterior probability (MAP). The optimization is carried out by the iterative conditional modes (ICM). By the former process, some separate regions with the bigger size are still in the mask. We label these “holes” and fill them with “1” if it is encircled by “1”, otherwise with “0”. Finally, we get a clean object mask, which is used to extract moving objects in a video sequence.

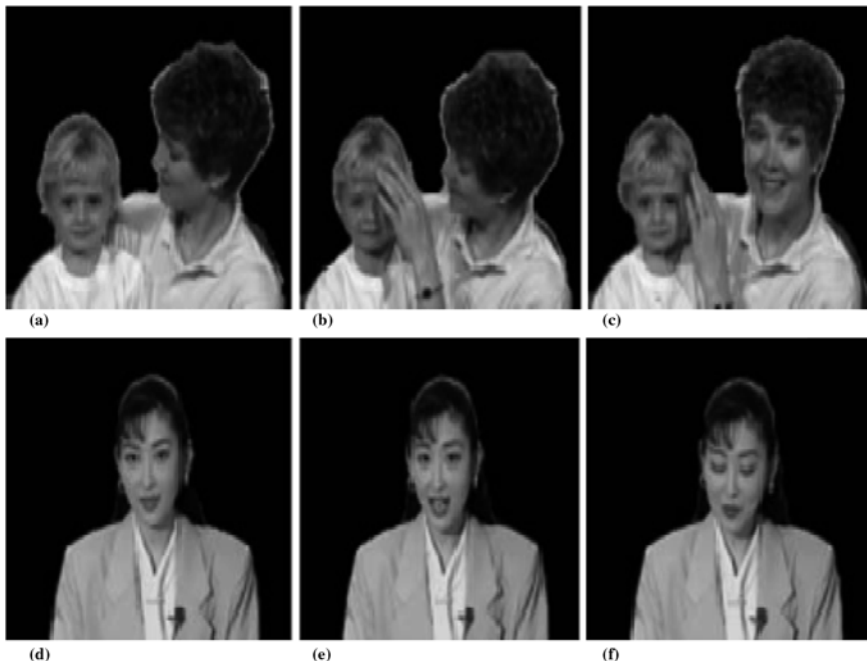


Fig. 59.2 Experimental results on Mother & daughter and Akiyo sequences. **a** 10# frame **b** 20# frame **c** 50# frame **d** 10# frame **e** 30# frame **f** 80# frame

59.6 Experimental Result

To assess the validity of the proposed method, it was tested on several MPEG-4 test sequences such as Akiyo and Mother & daughter. Mother & daughter is a typical test sequence, which is used to show the process to video segmentation. Figure 59.1 shows the segmentation of the first frame in Mother & daughter sequence. After four steps: regions partition, initial classification, spatio-temporal energy selection, and post processing, the motion object (Mother & daughter) is accurately extracted. The others experimental results are illustrated in Fig. 59.2. We randomly extract some segmentation frames from Mother & daughter and Akiyo sequences so that satisfactory results are obtained.

In order to objectively evaluate the performance of the proposed algorithm, the segmentation accuracy (SA) is defined as follows [9]:

$$SA(t) = 1 - \frac{\sum_{(x,y)} A_t^{est}(x,y) \oplus A_t^{ref}(x,y)}{\sum_{(x,y)} A_t^{ref}(x,y)} \quad (59.5)$$

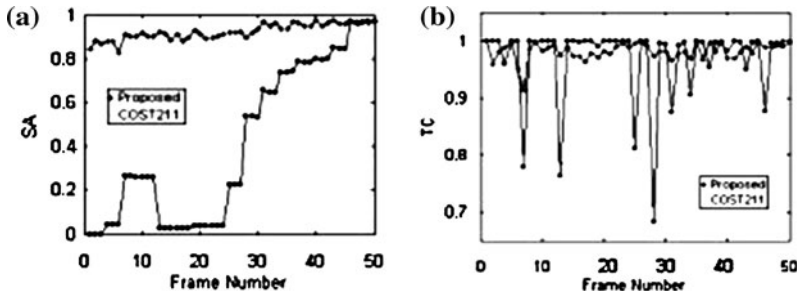


Fig. 59.3 Performance evaluation of our algorithm with COST211 AM on Mother & daughter sequence a SA b TC

Where \oplus is the logical “XOR” operation, and A_t^{est}, A_t^{ref} is the segmentation mask and reference mask, respectively. The MPEG-4 not provides the reference mask of any sequences, so we manually segment some references mask of Mother & daughter sequences using PHTOSHOP 7.0(image processing software).

Temporal coherency (TC) reflects the ability when segmenting long video sequence. Since TC is the extent, to which every segmented mask resembles the reference mask, i.e., the spatial accuracy of every frame. The difference value of SA between successive frames t and $t-1$ is defined:

$$TC(t) = 1 - |SA(t) - SA(t - 1)| \tag{59.6}$$

$$t = 2, 3, \dots, n$$

The segmentation results by COST211 AM are used as ground truth [10]. Performance evaluation on Mother & daughter sequence is illustrated in Fig. 59.3. From it, we can see that SA of the proposed approach improves about 0.05 ~ 0.80 and TC improves about 0.05 ~ 0.70 than COST211 AM. So better segmentation results are obtained by the proposed approach than COST211 AM.

59.7 Conclusions

In this paper, an automatic video segmentation is presented. The algorithm is designed based on regions classification. Regions partition is simple and fast in respect to others methods, which just meets the requirements for automatic video object segmentation. To overcome the errors of regions classification between moving object and background, the candidate regions is introduced as one of three classes. Energy model based on the motion and the spatial constraints is built to perform the candidate regions selecting. A post processing is utilized to get the accurate object. Experimental results show the effectiveness of our approach.

References

1. Zhang DS, Lu GJ (2001) Segmentation of moving object in image sequence: a review. *Circuits Syst Signal Process* 20(2):143–183
2. Gu C, Lee M-C (1998) Semiautomatic segmentation and tracking of semantic video objects. *IEEE Trans Circuits Syst Video Technol* 8(5):572–584
3. Tsaig Y, Averbuch A (2002) Automatic segmentation of moving objects in video sequences: a region labeling approach. *IEEE Trans. Circuits Syst Video Technol* 7(12):597–612
4. Gelon M, Bouthemy P (2000) A region-level motion-based graph representation and labeling for tracking a spatial image partition. *Pattern Recognit* 33:725–740
5. Patras I, Hendricks EA, Lagendijk RL (2001) Video segmentation by MAP Labeling of watershed segments. *IEEE Trans Pattern Anal Mach Intell* 23(3):326–331
6. Vincent L, Serra J (1991) Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Trans Pattern Anal Mach Intell* 13(6):583–589
7. Wang D (1998) Unsupervised video segmentation based on watersheds and temporal tracking. *IEEE Trans Circuits Syst Video Technol* 8(5):539–546
8. Mech R, Wollborn M (1998) A noise robust method for 2D shape estimation of moving objects in video sequences considering a moving camera. *Signal Process* 66:203–217
9. Wollborn M, Mech R (1998) Refined procedure for objective evaluation of video object segmentation algorithms [R]. Doc. ISO/IEC JTC1/SC29/WG11 M3448, March 1998
10. COST211 AM. Working site for sequences and algorithms exchange. <http://www.tele.ucl.ac.be/exchange>

Chapter 60

A Digital Image Scrambling Method Based on Hopfield Neural Network

Li Tu and Xuehua Huang

Abstract In order to improve the deficiencies of existing image restoration algorithms and accelerate its speed, through the study of Hopfield neural network, we proposed a digital image scrambling method of nonlinear mapping characteristics, and by using overlapped block technique, the time and space complexity reduced. Compared with the Arnold transformation, Fibonacci transformation, affine transformation and other scrambling methods, this method has better effective on noise resistance, JPEG compression, shear resistance attacking than the above three methods.

Keywords Image scrambling · Hopfield neural network · Overlapped block · Image restoration

60.1 Introduction

The degradation of the image is that a variety of reasons make a clear image blur in its acquisition and transmission, or the inappropriate condition causes the image obtained cannot achieve the proper quality specification [1]. The reduction in image quality will affect its following processing. The image restoration is to restore the degraded image as much as possible to the original image. The

L. Tu (✉) · X. Huang
Department of Computer Science, Hunan City University, Yiyang,
413000 Hunan, China
e-mail: tulip1907@163.com

X. Huang
e-mail: 40545092@QQ.com

traditional image restoration methods are Kalman filtering, Wiener filtering, inverse filtering, maximum entropy restoration method, etc., these methods have certain shortcomings [2].

The image restoration can be converted as an optimization problem, so the genetic algorithm is widely used in processing image restoration [3]. In some literature, the genetic algorithm was used to restore degraded images, and achieved good results, but the number of iterations were larger than 1000, the computational was large.

60.1.1 Hopfield Neural Network

The method of image restoration based on Hopfield neural network converted an image restoration problem as a optimization problem, the image restoration of neural network is based on the contraction of the network energy function, This completely avoids a matrix inversion problem of image restoration, and it has general applicability, but the establishment of the neural network increased the space complexity and time complexity greatly.

In this study, under the condition of without affecting the effect of image restoration, the traditional Hopfield network is improved [4], that make the space complexity and time complexity of algorithm reduced greatly.

60.2 Method of Image Processing

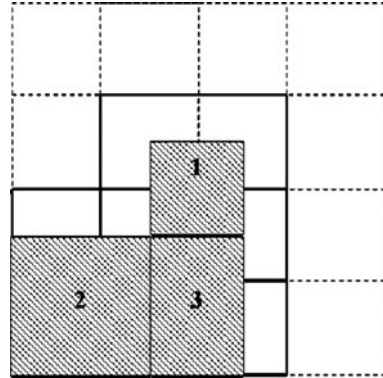
60.2.1 Image Block

Grayscale image is a limited set of function $\{x(i,j)|1 \leq i \leq M, 1 \leq j \leq N, x(i,j) = 1, 2, \dots, G\}$, $x(i,j)$ is the grayscale of the i row j -column pixel image in the image, the image height is M , the width is N , and the gray level is G [5]. The commonly method of image scrambling is a geometric transformation (including the Arnold transformation, Fibonacci transformation, etc.), affine transformation, the baker transformation, magic square transformation. The traditional methods of image restoration will inevitably dimensional calculation of equations, or to restore the process to meet the assumption of generalized stationary process, which makes the fuzzy image restoration problem more difficult.

In order to reduce the time complexity and the space complexity, the image block processing was used in this study. First, blocked the whole image, and then processed each block with neural network recovery algorithm, combined the results, and we can get a complete image.

Each pixel only is affected by the surrounding pixels within a certain range, so we can use technique of overlapping blocks in image restoration, this reduce the space complexity and time complexity of the algorithm.

Fig. 60.1 Image block sample



Blocked the image from the lower left corner (0, 0), set block size as $K \times K$, then the lower left corner of each block was $(i \times K/2, j \times K/2)$, and $\{i, j = 0, 1 \dots [2(L1/K)]\}$. If L cannot be divisible by K , then add the last make a pixel gray value of each row or column in blocking area without pixel.

After recovered each block of the image separately, to the blocks of the region without containing the image edge, saved the size of $(K/2) \times (K/2)$ area start from $(K/4, K/4)$ point in each sub-block; And to the blocks of the region containing the image edge, Saved them with an additional edge region. It had eight cases, the sub-blocks containing four corners and four sides of the image. For example, to the block containing the lower left corner of the image, saved the size of $(3 K/4) \times (3 K/4)$ region start from $(0, 0)$; And we can get the result of other blocks in the same way.

Figure 60.1 shows the examples of the image block. Assuming the image size is 4×4 , the block size is 2×2 , and the image is divided into nine. Such as the block does not contain the edges of the image, the block in the middle of the image, We saved the shadow areas in the middle of the image, this is the part 1; For the lower left corner of the block, we save the shaded part in the left bottom of the image, this was the part 2; And for the edge of the block, we save the shaded part at the below, it is the part 3.

60.2.2 Image Scrambling Method Based on Hopfield Neural Network

Digital image scrambling is essentially a class of image coding and decoding [6, 7], it can be seen as the mapping of the original image to the scrambling image. Hopfield neural network can realize the nonlinear mapping from input signal to output signal. In order to restore the image, it needs to create a neural network. It uses scrambling images as input signal, the original image as output signal.

Set the original image as I , the algorithm of blocking image restoration based on Hopfield neural network is described as below:

- (1) Set the memory model, divided the original image into non-overlapping blocks of 32×32 pixels, coded them, got the memory mode of value 1 and -1 : $U_k = [U_1^k, U_2^k, \dots, U_n^k]^T; k = 1, 2, \dots, m$; Then transformed the two-dimensional sub-block into p by a line stacking way, p is a one-dimensional vector.
- (2) Construct a Hopfield neural network and designed its weights, each node of the neural networks in a possible state (1 or -1), when stimulation of the neuron suffered more than its threshold, the neurons are in a state (such as 1), or the neurons will always be in another state (such as -1):

$$W_{ij} = \begin{cases} \frac{i}{N} \sum_{u=1}^M U_u^k, & U_j^k, j \neq i \\ 0, & j = i \end{cases} \quad (60.1)$$

W_{ij} is the connection weight between neuron i and neuron j , once calculated, the weight will remain the same;

- (3) Initialized the status of the network. Set the image mode $U' = [u'_1, u'_2, \dots, u'_n]^T$ as the initial state of the network, namely: $v_i(0) = u'_i$, $v_i(0)$ is the state of any neurons i at time of $t = 0$; Input the one-dimensional matrix P to the network, After processing the scrambling image obtained is the I' .
- (4) After treated the scrambling image of I' with method (1), got sample set P' ;
- (5) Iterated, according to the formula:

$$v_i(t + 1) = Sgn \left[\sum_{j=1}^N w_{ij} x_{j(n)} \right], \quad t = t + 1 \quad (60.2)$$

60.3 Simulation Results

The following are gray scrambled image results. The original image was a cameraman gray image of 64×64 pixels., it was shown in Fig. 60.2a1, the scrambled image was shown in Fig. 60.2b1. Through training and learning, established the relationship between scrambled image and restored image. The training objectives: the largest training times was 3000, the learning rate was 0.01, the target error is 0.0001. Trained on the cameraman image, saved the weights and the threshold. The restored image was shown in Fig. 60.2c1.

The experimental results showed:

The algorithm can scrambling the original image well, the illegal users are difficult to identify the original image from the image scrambled;



Fig. 60.2 Scrambling on the gray image and recovery them **a1** original cameraman image **b1** scrambled cameraman image **c1** restored cameraman image (PSNR = 50.9343 dB)

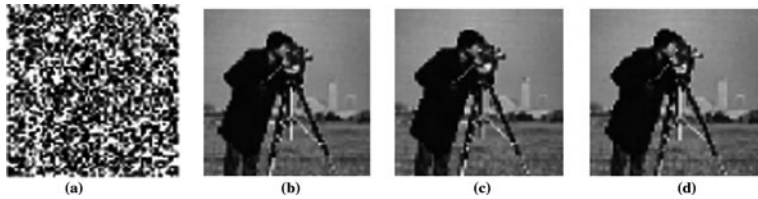


Fig. 60.3 Noise attacking experimental. **a** scrambled cameraman image **b** recovered image after salt and pepper noise attack, PSNR = 50.9300 dB **c** recovered image after Gaussian noise attack, PSNR = 33.8386 dB **d** recovered image after multiplicative speckle noise attack, PSNR = 50.9131 dB

Although the training time is slightly longer once the network was trained. The speed of image restoration was fast.

The effect of image restoration was good (PSNR values are more than 50 dB).

60.3.1 Anti-Attacking Test

The following images were scrambling to recover after attacked of different images, and compared it with Arnold transformation, Fibonacci transformation, affine transformation.

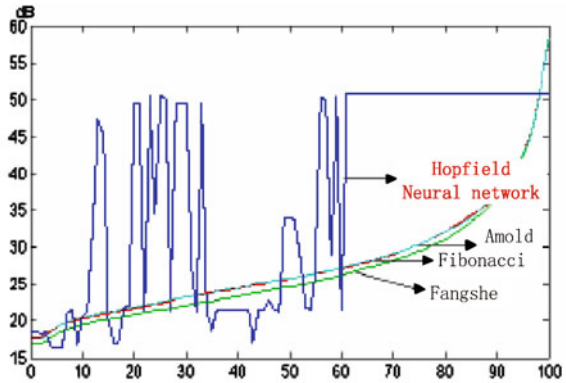
60.3.2 Noise Attack

Figure 60.3a is a cameraman scrambling image, Fig. 60.3b–d were the recovered image of scrambled image (a) by adding intensity is 0.001 salt and pepper noise, mean 0 and variance 0.01 Gaussian noise, mean 0 and variance 0.005

Table 60.1 The results of noise attack

	Hopfield (dB)	Amod (dB)	Fibonacci (dB)	Fangshe (dB)
Gaussian noise	33.8386	20.1489	20.0550	20.2936
Salt and pepper noise	50.9300	30.7883	37.7913	33.8787
Multiplicative speckle noise	50.9195	28.3928	28.4045	28.4351

Fig. 60.4 The relationship between PSNR and the JPEG quality retention



multiplicative speckle noise. The simulation results showed, the scrambling image can be recovered well by using the algorithm after polluted by noise.

Table 60.1 showed the recovered images, the original image was grayscale cameraman images of 64×64 pixels.

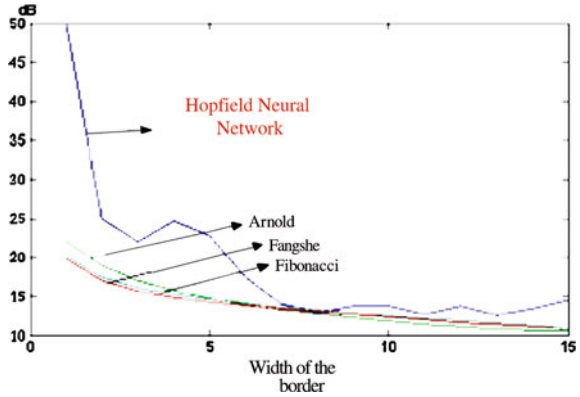
60.3.3 JPEG Compression Attack

The following was the experimental data of the JPEG compression attacking experiments. The original image was the cameraman image. Figure 60.4 shows the experimental results of JPEG compressing attacks. The abscissa was the quality factor of JPEG compression in this picture, values was from 0 to 100, the smaller the quality factor, the worse the image quality after compressing.

60.3.4 Cropping Attack

Figure 60.5 shows the experimental results of cropping (the original image was grayscale images of 64×64 pixels). The method of attacking is cut off the border of the image. In the figure the abscissa was width of the shear, the ordinate was

Fig. 60.5 The relationship between PSNR and the border width to be cut



PSNR values of the image. It can be seen from the figure the algorithm on the cropping attacking was significantly stronger than the other three scrambling methods.

60.4 Conclusion

In the algorithm of image restoration based on Hopfield neural network, the space complexity and the time complexity related to the size of the image image, using block processing can reduce the space complexity and time complexity.

References

1. Ma C, Ma L-Y (2009) A catastrophic genetic algorithm for super-resolution image restoration [J]. *J Image Graph* 14(8):1510–1515 (in Chinese)
2. Zhu J-B, Xu T-Z, Huang C (2006) A new method for image blind restoration based on PDE [J]. *Acta Electron Sin* 34(5):887–891 (in Chinese)
3. Zeng Y-S, Kang L-S, Ding L-X (2003) A close-to-optimal image restoration technique based on regularization method [J]. *Softw* 14(3):689–696 (in Chinese)
4. Yang G, Xu X, Yang G, Zhang J (2010) Semi-supervised classification by local coordination. *Lecture notes in computer science. Neural information processing models and Applications*, vol 6444, pp 517–524
5. Yang G, Xu X, Yang G, Zhang Jianming (2010) Research of local approximation in semi-supervised manifold learning. *J Inform Computat Sci* 7(13):2681–2688
6. Pok G, Liu J-C, Sanju Nair A (2003) Selective removal of impulse noise based on homogeneity level information. *IEEE Trans Image Process* 12(1):86–99
7. Petitcolas F, Anderson R, Kuhn M (1998) Attacks on copyright marking systems. *Lecture notes in computer science. Information hiding*, vol 524, pp 218–238

Chapter 61

Lattice Boltzmann Anisotropic Diffusion Model Based Image Segmentation

Qian Wu, Yu Chen and Gang Teng

Abstract This chapter introduces a new method of image segmentation based on Lattice Boltzmann Anisotropic Diffusion Model (LBADM). Anisotropic diffusion model is build by the membrane media. Active contour model is simulated by thermal diffusion equation, and the heat field is established. Then, by the LBADM it can be obtain the edge of object segmentation. The experiment and discussion is showed that our algorithm can accurately solve the convection–diffusion equation. It can obtain closed curves and deal with topology changes well. Contrasting with the level set method and narrowband level set method for computing speed, the result of the experiment is confirmed that the algorithm greatly reduces the computation of segmentation.

Keywords Image segmentation • Cellular automata • Lattice Boltzmann model • Anisotropic diffusion • Membrane media

Q. Wu (✉)

Department of Mechanic-Electronic Engineering, Suzhou Vocational University,
Suzhou 215014, Jiangsu, China
e-mail: sunniwu@163.com

Y. Chen

School of Electrical and Information Engineering, Jiangsu University,
Zhenjiang 212013, Jiangsu, China
e-mail: tzcy@163.com

G. Teng

Department of Computer Engineering, Suzhou Vocational University,
Suzhou 215014, Jiangsu, China
e-mail: fixxxer@163.com

61.1 Introduction

Image segmentation plays a critical role from image processing to image analysis. Nowadays, there are many methods related to different types of image segmentations, such as morphology, statistical methods and the method based on partial differential equations [1]. However, for these methods, the fundamental idea is that an image is separated into a number of non-overlap areas based on the regional similarities and the differences between regions. Along with the development of image analysis, some new image segmentation methods come after another. For example, the simple threshold setting technique, certain optimum solution according to certain mathematical models, two-dimensional image and three-dimensional image [2, 3]. Although these methods are all widely used in practice, there are still some disadvantages like only available to certain types of image, not widely applicable, sometimes in order to improve the image segmentation performance, several methods have to work together. All these shortcomings restrict the usage of image segmentation technique.

With the development of cellular automata, it provides a method to directly build a link between a known phenomenon and the evolution of cellular automata, and by using spatial structure and image similarity, it is easy to achieve the parallel process in image processing [4, 5]. This chapter introduces the cellular automata into image processing, and poses a new contour segmentation model developed from Lattice Boltzmann Anisotropic Diffusion Model. Due to the high efficiency and parallelism of the Boltzmann models, the amount of calculation will be reduced significantly and also access to a closed partition curve.

61.2 Lattice Boltzmann Model

Cellular Automata (CA) was created by Ulam and Von Neumann in the forties of last century's [6]. It is a dynamical system which is discrete both in time and space, the grid cell to synchronize all the changes according to the rules, every cell has a finite number of states, and the next state are only decided by the current state and the neighboring state of phase [7]. Since the space of cellular automata is very similar to the image structure, and the evolution of image has locality rule, CA is a good tool in image processing which is easy to use for parallel processor.

Lattice Boltzmann Model is a new numerical tool for solving partial differential equations. Lattice Boltzmann simulation of fluid movement originated in the law of the lattice gas cellular automata, and it also did the following improvements: (1) the square cellular mesh grid is developed from the triangle mesh grid to improve and simplify the cellular space, also easier to program; (2) the state space of cellular space is changed from discrete space to continuous space to eliminate the lattice gas cellular automata (LGA) Boolean properties; (3) the collision term is estimated by a single relaxation time Bhatnagar–Gross–Krook (BGK), which significantly reduced the amount of computation.

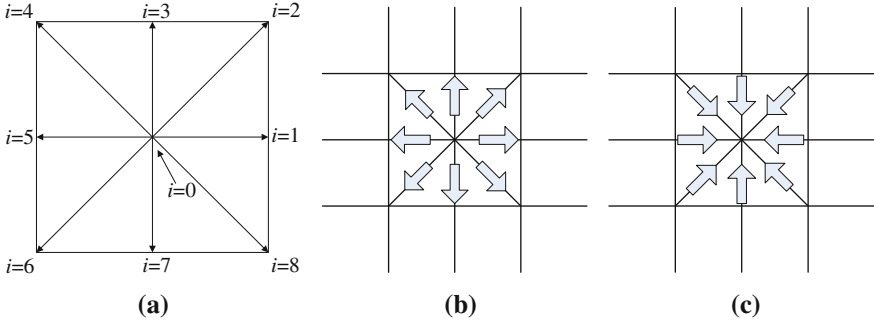


Fig. 61.1 Cellular space of LBM. The particles move and collision. **a** Cellular space. **b** Particles move. **c** Particles collision

Lattice Boltzmann automaton is one of the common cellular space shown in Fig. 61.1. Each cellular has eight adjacent neighbor cellular in east, west, south, north, southeast, southwest, northeast and northwest directions. Each cellular can move to one of the eight directions ($i = 1, 2, \dots, 8$) or stay in the original location ($i = 0$), and has collision in each grid.

Moving phase can be expressed as:

$$f_i^*(\vec{r} + \Delta h \cdot \vec{C}_i, t) = f_i(\vec{r}, t), \quad 1 \leq i \leq 8 \tag{61.1}$$

Collision phase can be expressed as:

$$f_i(\vec{r}, t + \Delta t) = f_i^*(\vec{r}, t) + \frac{1}{\zeta} \left[f_i^{(0)}(\vec{r}, t) - f_i^*(\vec{r}, t) \right], \quad 1 \leq i \leq 8 \tag{61.2}$$

According to formula (61.1) and (61.2), Lattice Boltzmann automata evolution rules can be expressed as:

$$f_i(\vec{r} + \Delta h \cdot \vec{C}_i, t + \Delta t) = f_i(\vec{r}, t) + \frac{1}{\zeta} \left[f_i^{(0)}(\vec{r}, t) - f_i(\vec{r}, t) \right] \tag{61.3}$$

Among them, $f_i(\vec{r}, t)$ is particle density distribution function; $f_i^{(0)}(\vec{r}, t)$ is local equilibrium distribution function which demonstrates the equilibrium state of particles after collision state. ζ is the relaxation time, which demonstrates the time of particle density within the cell that tends to balance the distribution function, and characterize the speed that model tends to equilibrium. Compared with the traditional numerical method, LBM has good stability, easily programmable, and high efficiency. Except that, LBM is a natural discrete system, therefore it is an excellent tool for image processing.

61.3 Lattice Boltzmann Model for Anisotropic diffusion

In order to build the anisotropic diffusion model on diffusion model, first of all, we introduce the semi-permeable membrane selective permeability mechanism. The so called semipermeable membrane is a class of existence that permeating small molecules and large molecules cannot pass through the film in general. Suppose a moment that macromolecules and small molecules moving from a semi-permeable membrane from outside to inside of the semipermeable membrane and the quantities are n_1 and n_2 , then after this specific time the small molecules through the semipermeable membrane, while the macromolecules were blocked outside the semi-permeable membrane. If the probability of particles passing through each cell g is not the same, it will lead to particles in vivo throughout the different diffusion coefficients. Based on this idea, we regard every cellular automata cells as one cell, add a semipermeable membrane between each cell factitiously that may be called “transmembrane media”, as shown in Fig. 61.2. When the particles move to the neighbor cell, with probability g through the membrane media, and with probability $1-g$ turned back. So, the model evolution equation can be written as:

$$\begin{aligned}
 f_i(\bar{r} + \Delta h \bar{C}_i, t + \Delta t) = & g_i(\bar{r}) \left\{ f_i(\bar{r}, t) + \frac{1}{\zeta} [f_i^{(0)}(\bar{r}, t) - f_i(\bar{r}, t)] \right\} \\
 & + [1 - g_i(\bar{r})] \left\{ f_i(\bar{r} + \Delta h \bar{C}_i, t) \right. \\
 & \left. + \frac{1}{\zeta} [f_i^{(0)}(\bar{r} + \Delta h \bar{C}_i, t) - f_i(\bar{r} + \Delta h \bar{C}_i, t)] \right\}
 \end{aligned} \tag{61.4}$$

Here we use another partial equilibrium function:

$$f_i^{(0)}(\bar{r}, t) = \begin{cases} \frac{M \rho(\bar{r}, t)}{9}, & i \neq 0 \\ 1 - \frac{8M \rho(\bar{r}, t)}{9}, & i = 0 \end{cases}$$

Expanse left part of Eq. 61.4 toward time and space to first-order Taylor and second order infinitesimal, simultaneous equations, we can obtain the macroscopic equations of the model:

$$\frac{\partial}{\partial t} \rho = \frac{M \Delta h^2}{18 \Delta t} \sum_i g_i(\bar{r}) (\bar{C}_i \cdot \nabla)^2 \rho - \frac{M \Delta h}{9 \Delta t} \sum_i g_i(\bar{r}) (\bar{C}_i \cdot \nabla) \rho - \frac{\varepsilon^2}{\Delta t \zeta} \sum_i g_i(\bar{r}) f_i^{(2)} \tag{61.5}$$

If we ignore the Second-order infinite small items $-\frac{\varepsilon^2}{\Delta t \zeta} \sum_i g_i(\bar{r}) f_i^{(2)}$ of Eq. 61.5, we can obtain

$$\frac{\partial}{\partial t} \rho = \frac{M \Delta h^2}{18 \Delta t} \sum_i g_i(\bar{r}) (\bar{C}_i \cdot \nabla)^2 \rho - \frac{M \Delta h}{9 \Delta t} \sum_i g_i(\bar{r}) (\bar{C}_i \cdot \nabla) \rho \tag{61.6}$$

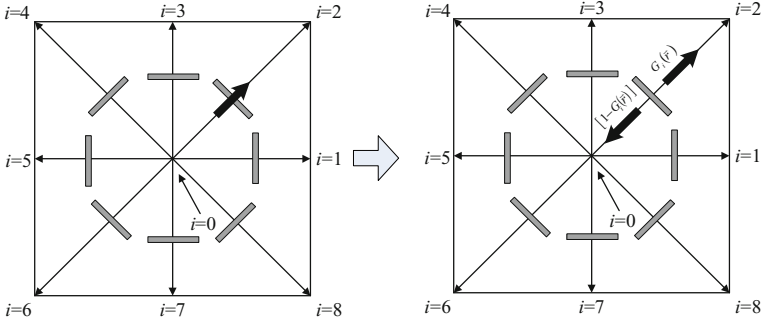


Fig. 61.2 Each particle through the membrane medium is denoted by the probability $g_i(\vec{r})$ and of return to original location by the probability $[1 - g_i(\vec{r})]$

If we expand anisotropic diffusion equation $\partial\rho/\partial t = \text{div}[c(|\nabla\rho|)\nabla\rho]$ proposed by Perona and Malik, we can see it is consistent with Eq. 61.6 in formally.

61.4 Experimental Analysis

61.4.1 Image Segmentation Algorithm Based on LBADM

1. Initialize the curve Γ .
2. Initialize the value of thermal field $\rho(\vec{r}, 0)$ with the following condition:

$$\rho(\vec{r}, 0) = \begin{cases} d(\rho(\vec{r}, 0), \Gamma) & \vec{r} \in S_1 \\ -d(\rho(\vec{r}, 0), \Gamma) & \vec{r} \in S_2 \\ 0 & \vec{r} \in \Gamma \end{cases}$$

where $d(\rho(\vec{r}, 0), \Gamma)$ is the distance between the node \vec{r} and the curve Γ . S_1 and S_2 denote the sets of the nodes inside and outside curve Γ , respectively.

3. Initialize the $f_i^{(0)}(\vec{r}, 0)$ according to function:

$$f_i^{(0)}(\vec{r}, t) = \begin{cases} M \frac{\rho(\vec{r}, t)}{9}, & i \neq 0 \\ 1 - \frac{8M\rho(\vec{r}, t)}{9}, & i = 0 \end{cases}$$

4. Set the initial value of $f_i(\bar{r}, t)$ with function: $f_i(\bar{r}, 0) = f_i^{(0)}(\bar{r}, 0)$.
5. Set the maximum step of the iteration N .
6. While $n < N$ do
7. Streaming with function (2-3): $f_i^*(\bar{r} + \Delta h \bar{C}_i, n\Delta t) = f_i(\bar{r}, n\Delta t)$, Where $\bar{r} \in S_1 \cup S_2, 1 \leq i \leq 8$
8. Collection with function (4):

$$f_i(\bar{r} + \Delta h \bar{C}_i, (n+1)\Delta t) = g_i(\bar{r}) \left\{ f_i^*(\bar{r}, n\Delta t) + \frac{1}{\zeta} \left[f_i^{(0)}(\bar{r}, t) - f_i^*(\bar{r}, n\Delta t) \right] \right\} + \left[1 - g_i(\bar{r}) \right] \left\{ f_i^*(\bar{r} + \Delta h \bar{C}_i, n\Delta t) + \frac{1}{\zeta} \left[f_i^{(0)}(\bar{r} + \Delta h \bar{C}_i, n\Delta t) - f_i^*(\bar{r} + \Delta h \bar{C}_i, n\Delta t) \right] \right\}$$

9. Update the state of the cells with:

$$\rho(\bar{r}, (n+1)\Delta t) = \sum_{i=0}^8 f_i(\bar{r}, (n+1)\Delta t)$$

10. Update curve Γ by finding the nodes satisfying $\rho(\bar{r}, t) = 0$.
11. If the node of the curve Γ reaches the boundary of thermal field, reinitialize the thermal field.
12. $n = n + 1$.
13. Compute the equilibrium function with:

$$f_i^{(0)}(\bar{r}, n\Delta t) = \begin{cases} M \frac{\rho(\bar{r}, n\Delta t)}{9}, & i \neq 0 \\ 1 - \frac{8M\rho(\bar{r}, n\Delta t)}{9}, & i = 0 \end{cases}$$

14. End While
15. Output curve Γ .

61.4.2 Preferences

Among them, $d(\rho(\bar{r}, 0), \Gamma)$ represent the shortest distance (Fig. 61.3) of the points \bar{r} at heat field to the zero temperature Γ , S_1 and S_2 represents the area inside and

Fig. 61.3 The *thick real line* represents the edge of the thermal field. The *dashed line* represents the curve Γ

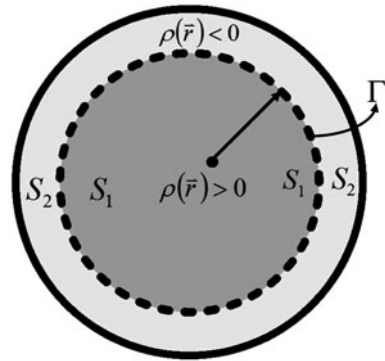
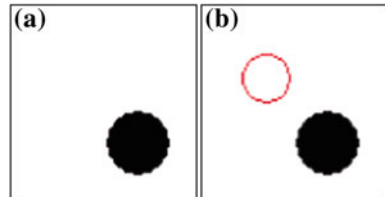


Fig. 61.4 The figure of comparison test



outside zero temperature Γ . In Fig. 61.3, dotted line represents zero temperature Γ and solid line represents the edge of the heat field.

The conformation of the heat field decides the direction of the zero temperature’s expanding. To ensure the zero temperature line has evolved in one direction, we can modify the area S_2 to achieve this goal. We always maintain the distance between the zero temperature to heat field as the width of this two pixels.

As the iteration, line of zero temperature keeps on expanding. To achieve the image segmentation, we want the zero temperature line stop expand at the area where the gradient is larger and continuing evolvement where the gradient is smaller. The bigger the n , the more sensitive it will be with the gradient, and also, it sensitive to noise interference. The smaller n is not very sensitive to the gradient but prone to over-segmentation. In the test, we assume $n = 2$. The other parameters of the model are selected as $\Delta h=1, \Delta t = 0.0833, M = 1$.

61.4.3 Experimental Results and Analysis

In order to evaluate the computational speed, the following algorithm is given only to the level set method and the narrow band level set method for computing time. We used three methods to split the black circle area in Fig. 61.4a. Table 61.1 shows the calculation time for each method. The experiment was based on

Table 61.1 The calculation time of LBADM, Narrow band level set method and level set method

	LBADM (s)	Narrow band level set method (s)	Level set method (s)
Time	561	2,122	3,746

Windows XP platform, used Matlab software. Figure (a) is a 82×82 binary image, (b) is the initial segmentation contour. It is obvious in Table 61.1, by using the algorithm proposed in this chapter, the total calculation time is 561 s, which is much less than that of level set method and the narrow band level set method. This shows that using our algorithm, image segmentation speed can be greatly improved.

61.5 Conclusion

The proposed algorithm, which is based on anisotropic diffusion model of active contour, overcomes the problem which segmentation contour is not continuous, and easy to mark partition defects. Image segmentation results show that the algorithm can implement great image segmentation on different images including CT images, magnetic resonance images, and digital subtraction angiography images and so on. Model can achieve comparable good segmentation results, and similar with the level set, it can well handle the image the same topology change.

Meanwhile, comparing the calculation time with level set method and the narrow band level set method, the proposed algorithm has high processing speed, which reflects the simplicity and high efficiency of the algorithm based on LBADM.

Acknowledgments The authors appreciate the support provided by Scientific Research Foundation for Young Teacher Project of Suzhou Vocational University.

References

1. Gastaud M, Barlaud M, Aubert G (2004) Combining shape prior and statistical features for active contour segmentation[C]. IEEE Trans Circuits Sys Video Technol 14(5):726–734 Special session on audio and video analysis for interactive multimedia services
2. Parker BJ, Feng DG (2005) Graph-based Mumford-Shah segmentation of dynamic PET with application to input function estimation [C]. IEEE Trans Nucl Sci 52(1):79–89
3. Mitiche A, Sekkati H (2006) Optical flow 3D segmentation and interpretation: a variational method with active curve evolution and level sets [C]. IEEE Trans Pattern Anal Mach Intell 28(11):1818–1829
4. Yuan WF, Tan KH (2007) An evacuation model using cellular automata[J]. Phys a Stat Mech Appl 384(2):549–566

5. Yue H, Hao H, Chen X, Shao C (2007) Simulation of pedestrian flow on square lattice based on cellular automata model [J]. *Phys A* 384(2):567–588
6. Walus K, Budiman RA, Jullien GA (2004) Split current quantum-dot cellular automata-modeling and simulation [C]. *IEEE Trans Nanotechnol* 3(2):249–255
7. Xiao X, Shao SH, Ding YS, Huang HD, Chen XJ, Chou KC (2005) An application of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation [J]. *J Theor Biol* 235(4):555–565

Part VI
Optimization and Scheduling

Chapter 62

Research on New Distributed Solution Method of Complex System Based on MAS

Wei jin Jiang, F. Guo Fan and Luo Zhong

Abstract Generalized decision function-based decomposition criterion in multi-agent systems and multi-agent task specification decomposition features are put forward in this paper. It is pointed out that problem of tasks decomposition in multi-agent systems is equivalent to that forming a multiply sectioned Bayesian (MSBN) related to information agent. The relation between d-cutset of MSBN and decomposition criterion is proved, and decomposition method of MSBN related to multi-agent task specification is put forward. Feasibility on the method put forward is verified with an example in the end.

Keywords Distribution computation · Information agent · Generalized decision function · Task specification decomposition · Bayesian network

62.1 Introduction

The multi-agent system was considered is dendritic structure which is composed by the agent marking node and the agent task specification, root pitch point marking is root-agent, leaf marking for leaf-agent, the non-root, non-leaf marking for guest-agent, information system [1] is composed by the universe of discourse

W. Jiang (✉) · F. Guo Fan
School of Computer and Electronic Engineering, Hunan University of Commerce,
Changsha, 410205, China
e-mail: ljxwhj@163.com

W. Jiang · L. Zhong
School of Computer Science and Technology, Wuhan University of Technology,
Wuhan, China

and the attribute collection, is called description information system agent information agent, takes the complex information system root-agent the task specification.

Distributional intelligence decision support system (DIDSS) [2, 3] may be composed by different node agent, is more than a agent system, the DIDSS decision-making question is a multi attribute system, is concealing the different type attribute collection and the between relations, the goal which the task specification decomposition is root-agent which corresponds to the multi attribute collection institute carries on the task specification decomposition, found corresponding leaf-agent and its the task specification, and takes it as the knowledge unit to save uses in the partial inference, thus decomposes its task specification based on agent DIDSS to the task specification policy-making performance into root-agent to become a series of sub-task specification, the sub-task specification which obtains assigns for corresponding leaf-agent, the leaf-agent antithetical couplet duty carries on the decision-making, and submits the policy-making result for root-agent, carries on by the root-agent antithetical couplet duty policy-making result builds up. Therefore, the key question is the task specification decomposition.

Generalized decision function (GDF) [3] can to the multi attribute, the multi-universe of discourse information system carry on the decomposition, produces the different subsystem, and does not send in the decomposition process has the overall situation information to lose, obtains the decomposition model using GDF equally in the theory of probability the condition independent rule, thus the multi attribute information system decomposes the subsystem the process equally into seek Bayesian Network (BN) the [4] decomposition process.

Regarding one the information system which assembles with multi-agent, the available big variable collection description, each agent has a variable subset correspondence knowledge, regarding in single agent situation, available BN expression domain knowledge, and in the observation situation which assigns carries on the inference, when the information system is very big (variable many) must carry on the decomposition to the system, uses the multi-agent coordination to solve the problem, how is the key question carries on a big information system between multi-agent the decomposition as well as each agent how expresses own sub-question.

Xiang [5] proposed multiply sectioned Bayesian (MSBN), MSBN was the big information system correspondence information agent task specification decomposition has provided an effective method, MSBN is (Bayesian subnet) is composed by group of related SBN, SBN expressed a agent subsystem corresponded the knowledge, between multi-agent composed an ultra tree structure, between such agent inference could by distributional carry on, thus the information agent duty decomposition question might sum up for established one MSBN and the solution correspondence SBN question which corresponded with the information agent task specification.

62.2 Information Agent Formal Description

Defines 1 *Ins* (information system) [1]: is a triple, namely $Ins = (U, A, d)$, U is the individual closed region, a is the condition attribute collection, d is the policy-making attribute. Regarding $a \in A$, has $a: U \rightarrow V_a, V_a$ for the condition attribute value collection, gathers $V_A = \cup_{a \in A} V_a$ is the condition attribute collection a value region; $d: U \rightarrow V_d, V_d$ for policy-making attribute d value collection, mark \rightarrow expresses between parameter mapping relations [6].

Regarding information system $Ins = (U, A, d)$, $B \subseteq A$ determines information function $Inf_B: U \rightarrow \times_{a_{ij} \in B} V_{a_{ij}}$, may determine by the under type $Inf_B(u) = \{(< a_{i1}, a_{i2}, \dots, a_{i|B|} >, < a_{i1}(u), a_{i2}(u), \dots, a_{i|B|}(u) >)|a_{ij} \in B\}$, which symbols \times expressed card accumulates, $u \in U|B|$ is the gathers B cardinal number, $B = (a_{i1}, a_{i2}, \dots, a_{i|B|})$ is in gathers on $A = \{a_1, a_2, \dots, a_{|A|}\}$ mapping.

Information agent has the general agent [6] characteristic, its particular performance for carries on the plan and the decision-making to the goal for the information system task specification, below structurally carries on the definition to information agent.

Defines 2 Information agent may define is six triple, namely($Ag, ag_A, In, CM, IP, Out$), among Ag is agent name: ag_A the expression for by the attribute is the task specification which A information system Ins is composed; in is the inductor, reflects information agent to the environment or other agent information sensation behavior; CM is the correspondence mechanism, uses for to realize between the multi-agent news transmission; IP is the information processor, in the goal table task specification carries on the plan, the decision-making through IP , information agent to the task specification plan, the policy-making performance for to the information system plan, the decision-making; out is the effector, IP after the task specification plan, the decision-making, forms a series of movements, through effector to environment or other agent react[7–9].

Defines 3 Regarding information agent, Ag, ag_A of mandate standardized corresponding to the $Ins_A = (U, A, d)$ of information system [10–12], if has task specification ag_{A1}, ag_{A2} to correspond separately to information system $Ins_{A1} = (U_1, A_1, d), Ins_{A2} = (U_2, A_2, d)$, satisfies condition

$$\left\{ \begin{array}{l} A = A_1 \cup A_2, a(u) = \\ \left\{ \begin{array}{l} a(u_1), \text{ if } a \in A_1 \\ a(u_2), \text{ if } a \in A_2 \end{array} \right. \\ u = \{u = (u_1, u_2) \in u_1 \times u_2| \\ Inf_{A_1 \cap A_2}^{Ins_A}(u_1) = \\ Inf_{A_1 \cap A_2}^{Ins_A}(u_2)\} \end{array} \right. \quad (62.1)$$

Called ag_{A1}, ag_{A2} is the sub-task specification which ag_A decomposes, among $u_1 \in U_1, u_2 \in U_2, u = (u_1, u_2) \in U, a \in A, U_1 \times U_2$ expresses U_1, U_2 card accumulates.

62.3 Information Agent Task Specification Decomposition Method

62.3.1 Based on GDF Information Agent Task Specification Decomposition Criterion

Defines 4 GDF [2]: Regarding information system which assigns $Ins = (U, A, d)$, $B \subseteq A$, the generalized decision function GDF definition is $\hat{\partial}_B^d: V_B \rightarrow 2^{V_d}$, namely $\hat{\partial}_B^d(\omega_B) = \{d(u): Inf_B(u) = \omega_B\}$

V_B is gathers the B attribute value region, ω_B is gathers the B attribute value into gather in the A attribute value mapping, 2^{V_d} for V_d Power collection $\hat{\partial}_B^d(\omega_B)$ the simple form is $\hat{\partial}_B$

If $A_1, A_2, X, Y, Z \subseteq A$, GDF has following nature:

- (1) $\hat{\partial}_A \supseteq \hat{\partial}_{A_1} \cap \hat{\partial}_{A_2}$
- (2) $Inf_A^{-1} = \varphi, \hat{\partial}_{A_1} \cup \hat{\partial}_{A_2} = \varphi$
- (3) $\hat{\partial}_{XUY} = \bigcup_{\omega_z \in V_z} \hat{\partial}_{XUYUZ}$
- (4) $\hat{\partial}_{XUYUZ} \subseteq \hat{\partial}_{XUY}$
- (5) $\hat{\partial}_{XUY} \cap \hat{\partial}_{XUYUZ} = \hat{\partial}_{XUYUZ}$

Decomposition criterion: supposes ag_A, ag_{A_1}, ag_{A_2} expression by information system $Ins_A = (U, A, d)$, $Ins_{A_1} = (U_1, A_1, d)$, $Ins_{A_2} = (U_2, A_2, d)$ corresponds task specification, A, A_1, A_2 are the corresponding attribute collection, $A_1, A_2 \subset A, A_1 \cup A_2 = A$, if satisfies $\hat{\partial}_{A_1} \cap \hat{\partial}_{A_2}$, then ag_{A_1}, ag_{A_2} in order to satisfy defines 3 information agent the task specification ag_A decomposition condition the sub-task specification, called ag_1 decomposable is ag_{A_1} and ag_{A_2} , the mark is $ag_A = ag_{A_1} \oplus$ and ag_{A_2} .

Supposes ag_A, ag_{A_1} are information system $Ins_A = (U, A, d)$, $Ins_{A_1} = (U, A, d)$ task specification, $A_1 \subseteq A$, and n attribute subset sum aggregate satisfies $\bigcup_{i=1}^n A_i = A$, if n subset GDF occurring together satisfies $\bigcap_{i=1}^n \hat{\partial}_{A_i} = \hat{\partial}_A$, then ag_{A_i} is n sub-task specification which ag_A decomposes, the mark is $ag_A = ag_{A_1} \oplus ag_{A_2} \oplus \dots \oplus ag_{A_n}$.

Defines 5 Assigns agent, the task specification is ag_A , corresponds system $Ins_A = (U, A, d)$, attribute subset $P, Q, R \subseteq A$, if $\hat{\partial}_{P \cup Q \cup R} = \hat{\partial}_P \cup \hat{\partial}_Q \cup \hat{\partial}_R$, $P \cup Q \cup R$ is composition task specification, $ag_{Q \cup R}$, records is $ag_{P \cup Q \cup R} = ag_{P \cup R} \oplus ag_{Q \cup R}$.

Obtains the following duty decomposition nature by the GDF nature:

Nature 1 if $ag_{P \cup Q \cup R} = ag_{P \cup R} \oplus ag_{Q \cup R}$, $ag_{P \cup Q \cup R} = ag_{Q \cup R} \oplus ag_{P \cup R}$.

Nature 2 if $ag_{P \cup Q \cup R \cup S} = ag_{P \cup S} \oplus ag_{Q \cup R \cup S}$, $ag_{P \cup Q \cup S} \oplus ag_{Q \cup S}$, and $ag_{P \cup R \cup S} = ag_{P \cup S} \oplus ag_{R \cup S}$.

Nature 3 if $ag_{p \cup Q \cup R \cup S} = ag_{p \cup R \cup S} \oplus ag_{Q \cup R \cup S}$, and $ag_{p \cup R \cup S} = ag_{p \cup S} \oplus ag_{R \cup S}$, $ag_{p \cup Q \cup R \cup S} = ag_{p \cup S} \oplus ag_{p \cup R \cup S}$.

The decomposition nature had reflected the information agent task specification and between the decomposition task specification transformation relations, are restricted in the article the length, nature proof omitting [6-9].

The above indicates information agent the task specification to be possible through GDF decomposition criterion $\hat{\partial}_A = \hat{\partial}_{A1} \cap \hat{\partial}_{A2}$ decomposes, how is the key question regarding task specification ag_A which assigns solves satisfies the decomposition criterion sub-task specification $\hat{\partial}_{A1}$, $\hat{\partial}_{A2}$, the this article following research indicated may use BN description information agent the task specification, but BN d-cutset is deciding decomposition criterion $\hat{\partial}_A = \hat{\partial}_{A1} \cap \hat{\partial}_{A2}$ sub-task specification ag_{A1} , ag_{A2} .

62.3.2 Based on the BN Task Specification Expressed and Decomposes

Defines 6 Bayesian Network (BN) is a triple, namely $BN = (N, G, P)$ [13], N for variable collection; G has to the non-link chart, $G = (V, E)$, V is the node collection, E has to the side collection; in the E element has to the side e reflection node variable between cause and effect dependence relations, is x pair of y has the direct cause and effect influence from the node x to node y having to the side direct-viewing meaning, x is the y father node, y is x child node; P is the node probability distribution, indicated between the node the cause and effect influence intensity, each node all has a condition probability table, the quota describes his/her all father node to this node function effect, regarding each node x, $Pa(x)$ is x father node, then assigns in the $Pa(x)$ situation, the node x conditional probability is $P(x|Pa(x))$ [14, 15].

Supposes $G = (V, E)$ is BN having to the non-link chart, $X, Y, Z \subseteq V$ is the node collection, if Z divides any pair of $x \in X \setminus Z$ and $y \in Y \setminus Z$, Z d-divides X and Y, records is $(X, Y/Z)$.

Supposes $G_i = (V_i, E_i)$ (i=1, 2) is two charts, chart $G = (V_1 \cup V_2, E_1 \cup E_2)$ is called is the G_1, G_2 merge, records is $G = G_1 \cup G_2$.

Defines 7 Multi-module Bayesian Network (MSBN) is a triple, namely $MSBN = (N, G, P)$. In the formula $N = \cup N_i$ is composed by certain variable subset N_i sum aggregate; $G = (V, E) = \cup G_i$, is certain has to the non-link chart G_i merge becomes, each $G_i = (V_i, E_i)$ node collection V_i element r_i uses variable subset N_i the element marking; P is the joint probability, supposes x is a variable, its occurring together is $Pa(x)$, x conditional probability is $P(x|Pa(x))$, joint probability $P = \prod_{i=1}^n p_{G_i}$, P has to non-link sub graph G_i in the node probability distribution [16, 17].

If $SBN = (N_i, G_i, P_{G_i})$ the expression with the Bayesian subnet which i marks, then MSBN is merges by n Bayesian subnet [10, 11], namely $MSBN = \bigcup_{i=1}^n SBN_i = \left(\bigcup_{i=1}^n N_i, \bigcup_{i=1}^n G_i, \bigcup_{i=1}^n P_{G_i} \right)$, called MSBN by the decomposition is n SBN_i .

The MSBN decomposition key is in MSBN having to the non-link chart G decomposition, may decompose MSBN through d -cutset having to the non-link chart.

Supposes $G = G_1 \cup G_2 \dots \cup G_n = (V, E)$ is has by n to the non-link sub graph merge, $G_i = (V_i, E_i)$, V, V_i in order to correspond the chart the node collection, E, E_i have to the side collection, regarding random sub graph G_i, G_j , the node collection is V_i, V_j occurs together $I = V_i \cap V_j$, if regarding each $x \in I$ in G father collection $Pa(x)$, either $Pa(x) \subseteq V_i$, either $Pa(x) \subseteq V_j$, either $Pa(x) \subseteq V_k, k \neq i, k \neq j$, then the name occurs together I_{ij} is each element $x \in I$ is called the cut point in G_i, G_j d -cutset, d -cutset, called G is divided by I_{ij} certain sub graph $\{G_1, G_2, \dots, G_n\}$.

Describes the information agent task specification performance with BN for to use BN the node description information agent task specification attribute, has to describes between the attribute probability dependence relations, MSBN is composed by group of related SBN, decomposes the information root-agent task specification much the information leaf-agent task specification, each information leaf -agent task specification corresponds SBN, thus regarding more than attribute information agent, its task specification available MSBN indicated; decomposes certain SBN MSBN, each SBN expression information leaf -agent task specification [18, 19].

Directs the principle assigns the information system $Ins = (U, A, d)$, attribute collection $A = \{a_1, a_2, \dots, a_{|A|}\}$, $G = (V, E)$ in order to have to the non-link chart, in V element correspondence information system Ins attribute collection A element, has to the side $E = \bigcup_{i=1}^n \{(a_i, a_i) : a_i \in M_i\}, M_i$ is in A_i a a_i decomposition boundary, $X, Y, Z \subseteq A$, if exists $(X, Y/Z), \partial_X \cup_Y \cup_Z = \partial_X \cup_Y \cap \partial_Y \cup_Z$

By directs the principle, may obtain the following theorem and the deduction [17]:

Theorem *Supposes G_1, G_2, G is has to the non-link chart, and $G_1 = (V_1, E_1), G_2 = (V_2, E_2), G = G_1 \cup G_2 = (V, E), V = V_1 \cup V_2, E = E_1 \cup E_2$, if $I = V_1 \cap V_2$ is G_1, G_2 a d -cutset, $(V_1, V_2/I)$.*

Proof Shear the collection definition by the d - division and d - to obtain I division $x \in V_1 \setminus I$ and $y \in V_2 \setminus I$, namely $\langle V_1, V_2/I \rangle$.

Under by directs the principle and the theorem, this article obtains deduces:

Deduces 1 *supposes G_1, G_2, G is has to the non-link chart, $G_1 = (V_1, E_1), G_2 = (V_2, E_2), G = G_1 \cup G_2 = (V, E), V = V_1 \cup V_2, E = E_1 \cup E_2$, if $I = V_1 \cap V_2$ is G_1, G_2 a d -cutset, then $\partial_V = \partial_{V_1} \cap \partial_{V_2}$.*

Proof Because $I = V_1 \cap V_2$, $V_1 = V_1 \cup I$, $V_2 = V_2 \cup I = V_1 \cup V_2 \cup I$. By theorem knowledge $\langle V_1, V_2/I \rangle$, again by directs the principle to obtain $\hat{\partial}_V = \hat{\partial}_{V_1} \cap \hat{\partial}_{V_2}$.

Deduces 2 supposes G_1, G_2, G is has to the non-link chart, $G_1 = (V_1, E_1)$, $G_2 = (V_2, E_2)$, $G = G_1 \cup G_2 = (V, E)$, $V = V_1 \cup V_2$, $E = E_1 \cup E_2$, if $I = V_1 \cap V_2$ is G_1, G_2 a d-cutset, then $ag_V = ag_{V_1} \oplus ag_{V_2}$.

Proof by deduces 1 to obtain with the decomposition criterion definition.

Deduces 3 supposes G_1, G is has to the non-link chart, $G_i = (V_i, E_i)$, $G = \bigcup_{i=1}^P G_i = (V, E)$, $V = \bigcup V_i$, $E = \bigcup E_i$, In G regarding random G_m, G_n , if $I_{ij} = V_m \cap V_n$ is G_m, G_n d-cutset, then $ag_V = ag_{V_1} \oplus ag_{V_2} \dots \oplus ag_{V_P}$.

The lemma has established between the decomposition criterion and the BN relations, and pushes by the theorem 1, 2, 3. Obtains the conclusion is MSBN d-cutset was deciding satisfies the decomposition criterion $\hat{\partial}_A = \hat{\partial}_{A_1} \cap \hat{\partial}_{A_2}$ sub-task specification ag_{A_1}, ag_{A_2} , thus the information agent task specification decomposition question may sum up for establishes one MSBN which corresponds with the information agent task specification to solve its d-cutset and the corresponding SBN question.

62.4 Conclusions

Under the distributed environment may use the agent technology establishment based on agent DIDSS, its key question lies in the task specification which assigns decomposes into a series of sub-task specification, and dispenses it for different node agent, forms multi-agent the task specification, thus the DIDSS performance for to the decomposition sub-duty decision-making and the synthesis, the article decomposes agent using the generalized decision function as the criterion the task specification, proposed used MSBN establishment information agent the task specification decomposition network architecture, the method which this article proposed is DIDSS under the environment information agent task specification decomposition has provided one novel method.

Acknowledgments This work is supported by the Natural Science Foundation of Hunan Province of China No. 10JJ5064, the Society Science Foundation of Hunan Province of China No. 07YBB239, the Sci. and Technology Plan Project of China under of Hunan Grant No.2009GK2002, and the Technology Innovation Foundation under Science and Technology Ministry Grant No. 09C26214301947.

References

1. Polkowski L, Skwron A (1996) Rough metrology: a new paradigm for approximate reasoning, *J. Int J Approx Reason* 15(4):333–365
2. Jiang WJ, Wang P (2006) Research on distributed solution and correspond consequence of complex system based on MAS [J]. *J Comput Res Dev* 43(9):1615–1623

3. Sleazk D (2006) Approximate reducts in decision tables [A]. IMPU-96 Information processing and management of uncertainty on knowledge-based systems [C], Granada, Spain, 1–5(3):1159–1164
4. Pearl J (1988) A probabilistic reasoning in intelligence system: networks of plausible inference [M]. Morgan Kaufmann, San Moteo
5. Xiang Y (1996) A probabilistic frame work for cooperative multi-agent distributed interpretation and optimization of communication [J]. *Artif Intell* 87(1–2):295–342
6. Wooldridge M, Jennings NR (1995) Intelligent agent: theory and practice [J]. *Knowl Eng Rev* 10(2):115–152
7. Xiang Y (2008) Bayesian network repository [EB/OD]. <http://snowwhite.cis.uoguelph.ca/faculty/info/yxiang/index.html>
8. Xu K, Wang Y, Wu C (2007) Test technology and application of model of chain of services of Grid [J]. *Sci China*, Vol E, 37(4):467–485
9. Weicai Z, Jin L, Mingzhi X et al (2004) A multi agent genetic algorithm for global numerical optimization [J]. *IEEE Trans Syst Man Cybern* 34(2):1128–1141
10. Busoniu L, Babuska R, Schutter BD (2008) A comprehensive survey of multiagent reinforcement learning. *IEEE Trans Syst Man Cybern-Part C Appl Rev* 38(2):156–172
11. Gou Y, Huan JZ, Rong H (2005) Adaptive grid job allocation with genetic algorithm. In: *Future Genre Comp Sys*. Elsevier Press, London 21:151–161
12. Matsui M (1993) A generalized model of convey-serviced production station (CSPS). *J Jpn Ind Manag Assoc* 44(1):25–32
13. Bredin J, Kotz D, Rus D, Maheswaran RT, Imer C, Basar T (2003) Computational markets to regulate mobile-agent systems [J]. *Auton Agents Multi-Agent Syst* 6(3):235–263
14. Nakase N, Yamada T, Matsui M (2002) A management design approach to a simple flexible assembly system. *Int J Prod Econ* 76:281–292
15. Jiang WJ, Wang P, Lianmei Z (2009) Research on grid resource scheduling algorithm based on MAS cooperative bidding game. *Chin Sci F* 52(8):1302–1320
16. Jiang WJ, Wang P (2006) Research on distributed solution and correspond consequence of complex system based on MAS [J]. *J Comput Res Dev* 43(9):1615–1623
17. Jie C, Greiner R, Kelly J et al (2002) Learning Bayesian network from data: an information-theory based approach [J]. *Artif Intell* 137:43–90
18. Choi SPM, Liu J (2007) A genetic agent-based negotiation system [J]. *Comput Netw* (37):195–204
19. Guiquan L, Xiaoping C, Yan F et al (2007) A formal model of multi-agent cooperative systems. *J Comput* 24(5):529–535

Chapter 63

An Order-Searching Genetic Algorithm for Multi-Dimensional Assignment Problem

Yazhen Jiang, Xinming Zhang and Li Zhou

Abstract This paper proposes an order-searching genetic algorithm to solve multi-dimensional assignment problem by fusing the order-searching algorithm and the genetic algorithm. The new algorithm searches the optimal solution in the whole solution space firstly, then memorizes and improves the better solution by using cross and mutation operations, so that the convergence speed is accelerated and the accuracy of the data association is improved. Simulation results show that the new algorithm is feasible and effective.

Keywords Multi-dimensional assignment problem · Multi-layer order-searching algorithm · Genetic algorithm · Solution space

63.1 Introduction

Data association problem of multi-sensor multi-target location system can usually be transformed into a multi-dimensional (S - D) assignment problem [1, 2]. Order-searching algorithm is a random searching algorithm to solve the multidimensional assignment problem, and the multi-layer order-searching algorithm is an heuristic searching algorithm which is based on the order-searching algorithm [3]. The multi-layer order-searching algorithm is more suitable for the better detection

Y. Jiang (✉) · L. Zhou
School of Information Science and Engineering, Ludong University,
Yantai 264025 China
e-mail: zxyzlw_99@sina.com

X. Zhang
Foreign Languages Institute, Ludong University, Yantai 264025, China

environment. In the medium density target scenario, the convergence speed becomes slow. While in dense target and clutter scenario, it can get a near optimal solution in a short time, but it is usually not the best one.

Using genetic algorithm to solve multi-dimensional assignment problem, since of the inadequate information deriving from suboptimal solutions at the beginning of the algorithm, the convergence speed is very slow [4]. This paper proposes an order-searching genetic algorithm by fusing the order-searching algorithm and the genetic algorithm. This algorithm can not only guarantee to search the feasible solutions of better quality by using order-searching algorithm and multi-layer order-searching algorithm in the global solution space early in the algorithm, but also use cross operation in the genetic algorithm to integrate the useful information contained in the optimal solutions in the middle and late stages of the algorithm. And use mutation operator to further optimize part of scattered components of the solution.

63.2 Multi-Dimensional Assignment Model

There are many ways to construct a multi-dimensional assignment model of data association for multi-sensor multi-target location system. A relatively effective method is to construct the multi-dimensional assignment model by using the negative log likelyhood ratio based on measurement division. And the brief introduction of the model construction of multi-dimensional assignment problem based on measurements of bearing-only passive sensors is described as follows [1, 2, 5, 6].

The likelihood function of bearing-only measurements $\mathbf{Z}_{i_1 i_2 \dots i_S} = \{\mathbf{Z}_{1i_1}, \mathbf{Z}_{2i_2}, \dots, \mathbf{Z}_{Si_S}\}$ originates from a same target is given as

$$A(\mathbf{Z}_{i_1 i_2 \dots i_S} | t) = \prod_{s=1}^S [P_{ds} \cdot p(\mathbf{Z}_{si_s} | \omega_t)]^{u(i_s)} \cdot [1 - P_{ds}]^{[1-u(i_s)]} \quad (63.1)$$

Where, P_{ds} is the detect probability of sensor s , $u(i_s)$ is a binary indicator function. If $i_s = 1$, i.e., sensor s detect target t , it is 1. Otherwise, it is 0. $p(\mathbf{Z}_{si_s} | \omega_t)$ is the probability density function of \mathbf{Z}_{si_s} being from target t .

The likelihood function that the measurements are all spurious or unrelated to target t is

$$A(\mathbf{Z}_{i_1 i_2 \dots i_S} | t = \phi) = \prod_{s=1}^S \left[\frac{1}{\Psi_s} \right]^{u(i_s)} \quad (63.2)$$

where, Ψ_s is the field of the view of sensor s . The association cost of associating the S -tuple of measurement $\mathbf{Z}_{i_1 i_2 \dots i_S}$ with target t is given by

$$c_{i_1 i_2 \dots i_S} = -\ln \frac{\Lambda(\mathbf{Z}_{i_1 i_2 \dots i_S} | t)}{\Lambda(\mathbf{Z}_{i_1 i_2 \dots i_S} | t = \varphi)} \tag{63.3}$$

where, ω_t in (63.1) is usually replaced by its least-square estimate $\hat{\omega}_t = \arg \max_{\omega_t} \Lambda(\mathbf{Z}_{i_1 i_2 \dots i_S} | t)$. Then the cost of associating $\mathbf{Z}_{i_1 i_2 \dots i_S}$ with target t is

$$c_{i_1 i_2 \dots i_S} = \sum_{s=1}^S [u(i_s) \left(\ln \left(\frac{\sqrt{2\pi} \cdot \sigma_s}{P_{ds} \cdot \Psi_s} \right) + \frac{1}{2} \left(\frac{Z_{s i_s} - \hat{\theta}_{st}}{\sigma_s} \right)^2 \right) - (1 - u(i_s)) \cdot \ln(1 - P_{ds})] \tag{63.4}$$

Then the ociation problem can be transformed into the following generalized S-D assignment problem:

$$\min_{\rho_{i_1 i_2 \dots i_S}} \sum_{i_1=0}^{n_1} \sum_{i_2=0}^{n_2} \dots \sum_{i_S=0}^{n_S} c_{i_1 i_2 \dots i_S} \rho_{i_1 i_2 \dots i_S} \tag{63.5a}$$

subject to

$$\left\{ \begin{array}{l} \sum_{i_2=0}^{n_2} \sum_{i_3=0}^{n_3} \dots \sum_{i_S=0}^{n_S} \rho_{i_1 i_2 \dots i_S} = 1; \quad \forall i_1 = 1, 2 \dots n_1 \\ \sum_{i_1=0}^{n_1} \sum_{i_3=0}^{n_3} \dots \sum_{i_S=0}^{n_S} \rho_{i_1 i_2 \dots i_S} = 1; \quad \forall i_2 = 1, 2 \dots n_2 \\ \vdots \\ \sum_{i_1=0}^{n_1} \sum_{i_2=0}^{n_2} \dots \sum_{i_{S-1}=0}^{n_{S-1}} \rho_{i_1 i_2 \dots i_S} = 1; \quad \forall i_S = 1, 2 \dots n_S \end{array} \right. \tag{63.5b}$$

where, $\rho_{i_1 i_2 \dots i_S}$ are binary correlation variables. If the S -tuple of measurement $\mathbf{Z}_{i_1 i_2 \dots i_S}$ originates from a real target, it is 1; otherwise, it is 0.

In an assignment problem whose all elements of the cost matrix are involved in the optimal assignment process, the number of feasible solutions grows exponentially with the increase of the dimension of the assignment problem. It is obviously not very satisfactory to search for the optimal solution in the huge solution space composed of all the feasible solutions. It is no doubt that the efficiency of the searching algorithm will be improved to some extent if we search the optimal solution according to some heuristic information. To solve this problem, we propose the following order-searching genetic algorithm.

63.3 Order-Searching Genetic Algorithm

63.3.1 Shortcomings of Order-Searching Algorithm

Considering that most of the assignment models are transformed from practical problems, and according to the concrete meaning of practical application, usually we only permit part elements of the cost matrix in the optimal assignment process. In certain specific circumstances, elements permitted involving in the optimal assignment process are few and not very regular, it is difficult to find a feasible solution to meet actual requirements in a short period of time using basic order-searching algorithm. Even we adopt multi-layer order-searching algorithm, there is not an obvious improvement. To solve this problem, we can separate the feasible solutions out with the obtained infeasible results. But it takes a large amount of time. If the valid information in feasible solutions of better quality can be adopted and the inferior information in feasible solutions of less quality can be varied in time, it can undoubtedly accelerate the convergence rate and improve the convergence of order-searching algorithm.

63.3.2 Genetic Algorithm

In order to solve assignment problems using genetic algorithm, first of all is to encode the solutions of multi-dimensional assignment problem into the encoding string suitable for genetic algorithm. In general, genetic algorithms used to solve combinatorial optimization problems are binary encoding form. Taking consideration of a large number of decoding operations in the process of using binary code form to solve multi-dimensional assignment problem [4], this paper employs decimal encoding. For example, for a $n \times n \times n$ cost matrix of a 3-D assignment problem, one of its feasible solutions can be composed of three lines of coding sequences. To facilitate that, we let the three lines of coding sequences side by side, after that, each column corresponds to one gene of the genetic algorithm, and all the genes make up of a chromosome. Take $n = 8$ for example, codes of two chromosomes A_1, A_2 can be denoted as:

$$A_1 = \begin{pmatrix} 4 & 2 & 3 & 6 & 8 & 5 & 7 & 1 \\ 4 & 2 & 3 & 5 & 6 & 8 & 7 & 1 \\ 2 & 7 & 3 & 4 & 6 & 5 & 8 & 1 \end{pmatrix}$$

$$A_2 = \begin{pmatrix} 7 & 1 & 5 & 2 & 3 & 4 & 8 & 6 \\ 7 & 2 & 1 & 4 & 3 & 6 & 8 & 5 \\ 7 & 4 & 1 & 5 & 2 & 6 & 8 & 3 \end{pmatrix}$$

Each chromosome corresponds to a feasible solution of the assignment problem. The ordered 3-tuple of different rows but the same column of A_1 or A_2 was a gene of the chromosome, which just corresponds to a component of a feasible solution of the 3-D assignment problem. Noted that each line of a chromosome is a full array of 1 to n (dimension of the cost matrix), so it usually needs a random search to find a set of feasible solutions of the assignment problem at the beginning of the traditional genetic algorithm. This can be realized by using the *randpermz* instruction in Matlab three times, in other words, it can construct three different arrays of 1 to n natural numbers, and these just constitute a set of feasible solution meeting the constraint conditions of 3-D assignment problem, which is as well as a chromosome of the genetic algorithm.

63.3.3 Order-Searching Genetic Algorithm

Using random searching to generate a chromosome at the beginning of the genetic algorithm, the fitness function (the reciprocals of the value of objective function) is usually not be very good. Even crossover and mutation operations are used to improve the efficiency of the algorithm, the convergence speed is still slow. Therefore, just using genetic algorithm to solve assignment problem with larger scale, the time the algorithm spent is far from our satisfaction.

In order to search feasible solution of good quality containing local information at the beginning of the genetic algorithm as soon as possible, we can use order-searching algorithm to search in the global solution space in advance. And take good quality feasible solutions gotten from random order-searching algorithm through several times as individuals in the initial group of the genetic algorithm. By the performance of order-searching algorithm, we know that these individuals not only contains better gene fragments, but there is no exchange of information between each other, which has a global searching effect.

Using genetic algorithm to solve assignment problem, we must ensure that individuals through no matter crossover or variation still meet the feasible constraints. As A_1 and A_2 show, arbitrary exchange of gene segments of the two chromosomes would break the feasible constraints of the assignment problem. Therefore, we take a multi-operators restructuring strategy. That is, from m feasible solutions of good quality obtained from order-searching algorithm, randomly selected m_1 ($2 \leq m_1 < m$) chromosomes to form a group, totally there are n_1 groups. Take out the same gene segments of several chromosome, i.e., among the m_1 chromosomes, once if there are more than $\lceil (m_1 + 1)/2 \rceil$ chromosomes contain the same gene segment, write down the location of this gene segment. After that, may be there are no 1 appears on some cost plane of the solution matrix, so we cross out all the cost planes which contain 1, then get a dimension-decreased

matrix. Repeat the multi-operators crossover operations above until the dimension of the matrix is less than 10. Then we can directly implement the multi-layer order-searching algorithm on the dimension-decreased cost matrix to get some variate optimization genes segments.

During the next iteration, in order to maintain the scale of groups of size m , in addition to the n_1 chromosomes above, we also use order-searching algorithm again to generate n_2 chromosomes. And choose m chromosomes of maximum value of fitness function among the original m chromosomes, n_1 chromosomes generated from evolution, the new generated n_2 chromosomes. And for these m chromosomes, repeat the multi-operator crossover and variation operations on the cost matrix. Cycle and so it goes on until it reaches the maximum iterations or the value of the fitness function does not change any more after five iterations.

Detailed description of the algorithm is given in the following.

$$\text{Step 1. } f = -\infty, \quad \max \text{ iter} = T, \quad k = 0, \quad l = 0$$

- Step 2. Generate M chromosomes by order-searching algorithm, and form the initial group of m ($M > m$) chromosomes which have bigger value of the fitness function. Choose m_1 ($2 < m_1 < m$) chromosomes from m chromosomes.
- Step 3. Take crossover and variation operations on each chromosome. Through retention, we record the good gene segments, cross out the cost plane contains one element in a certain segment.
- Step 4. If the dimension of the dimension-decreased cost matrix is still higher than a certain number given in advance, repeat the multi-operator crossover operation on the dimension-decreased cost matrix. Otherwise, use multi-layer order-searching algorithm on the dimension-decreased cost matrix. Compare the value of the fitness function of the corresponding feasible solution among n_1 chromosomes and n_2 chromosomes generated from order-searching algorithm, if $\max f_i > f, i = 1, 2, \dots, n_1 + n_2$, then $f := \max f_i$, and the optimal solution is denoted as $X = X_{\max f_i}$, let $l = 0$, turn to Step 5. Otherwise, let $l := l + 1$, when $l = 4$, stop, the optimal solution is $X^* = X$, and the optimal value $z^* = -f$. Otherwise, turn to Step 5.
- Step 5. If $k \neq T$, put n_1 solutions formed through evolution, n_2 feasible solutions generated from order-searching algorithm and the original m feasible solutions together, choose m feasible solution corresponding bigger fitness function value to construct the group of the next iteration. Let $k := k + 1$, repeat Step 2–Step 4. Otherwise, the optimal solution is X , the optimal value is $-f$.

Fig. 63 1 Position of four sensors

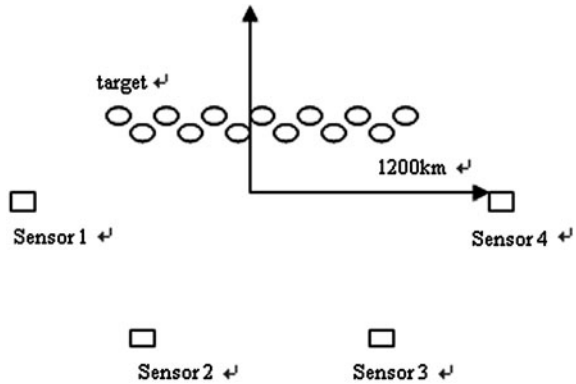


Table 63.1 Simulation results of 4-D assignment algorithms (average of 50 times)

		Accuracy (%)	Run time (s)	RMS (km)
$\pi/180$	General genetic algorithm	49.8	92.7	20.75
	Order-searching algorithm	72.1	197.6	7.36
	Order-searching genetic algorithm	78.4	112.8	5.67
$\pi/90$	General genetic algorithm	44.7	90.4	24.03
	Order-searching algorithm	62.5	171.3	10.96
	Order-searching genetic algorithm	68.2	108.6	9.30

63.4 Simulations

63.4.1 Simulation Model

Suppose four bearing-only passive sensors are used to locate targets, the cross location is shown in Fig. 63.1. For convenience, let the false alarm probability is 0 and the detection probability is 1. The number of target is 20, the interval of each target is 150 km. The measurement error of different sensors is the same, it is taken as $\pi/180$ and $\pi/90$. A multi-group iteration strategy is used in order-searching algorithm, and the number of groups is taken as 5. In 4-D assignment algorithm, the maximum iteration T is taken as 400, 320 to the corresponding two measurement errors.

63.4.2 Simulation Result Analyses

As is shown in Table 63.1, compared with order-searching genetic algorithm, the accuracy of data association of general genetic algorithm is much lower. The main reason is that the absence of heuristic information early in the algorithm process,

which leads to concentrations of individuals of low fitness, and the convergence rate is slow. In order to improve the accuracy of data association of the general genetic algorithm, we must increase the number of iterations.

From the results of Table 63.1, we can also see that order-searching genetic algorithm is better than order-searching algorithm in data association accuracy, and the procedure time decreases relatively. This result was mainly due to the recombination and variation of the multi-operator in genetic algorithm. It leads the searching towards the optimum solution, and accelerates the convergence speed.

63.5 Conclusions

In this paper, we take further research on two algorithms solving multi-dimensional assignment problem: order-searching algorithm and genetic algorithm. By fusing the ideas of both order-searching algorithm and the genetic algorithm, we propose an order-searching genetic algorithm to solve multi-dimensional assignment problem. The simulation results show that compared with simulation results using only one algorithm, the fused algorithm can effectively improve the data association accuracy of multi-target, and further improve the location accuracy.

Acknowledgments Supported by the National Natural Science Foundation of China (Grant No.60802088 and 60672140), the Natural Science Foundation of Shandong Province (Grant No.ZR2009GM001), the Technology Projects of Shandong University (Grant No.J09LG01), and the discipline construction project funding of Ludong University.

References

1. He Y, Wang G, Lu D et al (2010) Multisensory information fusion with applications. Electronic Industry Press, Beijing
2. Deb S, Yeddanapudi M, Pattipati K (1997) S-D assignment algorithm for multisensor-multitarget state estimation. *IEEE Trans Aerosp Electron Syst* (33):523–537
3. Zhou L, He Y, Weihua Z (2007) Order searching algorithm to solve multidimensional assignment problem with applications. *J Optoelectron Laser* (18):364–368
4. Chen Y, Yushu L, Jie F (2005) A niching genetic and ant colony optimization algorithm for generalized assignment problem. *J BeiJing Inst Technol* 25(6):490–494
5. Pattipati S, Deb KR, Bar-shalom Y (1993) A multisensor-multitarget data association algorithm for heterogeneous sensors. *IEEE Trans Aerosp Electron Syst* (29):560–568
6. Chummun MR, Kirubarajar T, Pattipati KR et al (2001) Fast data association using multidimensional assignment with clustering [J]. *IEEE Trans Aerosp Electron Syst* (37):898–913

Chapter 64

The Heuristic Algorithm of Stacking Layer for the Three-Dimensional Packing of Fixed-Size Cargoes

Wang-sheng Liu, Hua-yi Yin and Mao-qing Li

Abstract According to the actual operation of working people, a heuristic algorithm of stacking layer that meets the requirements of stability and convenience for load-and-unload was proposed. Firstly, choose the stacking direction according to the position of the compartment door. Secondly, optimize the combination of length, width and height along the stacking direction to minimize the remaining space. Finally, optimize each layer's layout. In each layer's layout, adopt long-short edge combination mode for each edge. Considering the flatness and stability of loading and unloading, the number of long-short edge is related. Experiment results show that the algorithm can maintain the requirements of stability and convenience of loading and unloading, and also has nice space utilization.

Keywords Fixed-size cargoes · Three-dimensional packing · Stacking layer method · Heuristic algorithm

64.1 Introduction

Currently, the logistics and distribution services have become important components of Enterprise competitive power, as the quality of freight loading has a direct influence towards to the efficiency of distribution, and further impact the working

W. Liu (✉) · H. Yin · M. Li
School of Computer and Information Engineering, Xiamen University,
Xiamen 361005, China
e-mail: kollzok@yahoo.com.cn

W. Liu
Research Centre of Modern Logistics, Jimei University,
Xiamen 361021, China

efficiency of the distribution center, increasingly importance has been attached to the study of freight loading. With the Standardization and normalization of the logistics and distribution services, especially the widely use of great tonnage container transportation, in order to improve the loading efficiency, more and more standardized packaging has been used. Therefore, the research of the fix-size loading is an urgent issue. However, in previous researches, loading problem with different size is more explored, such as [1–10], the algorithms proposed are for different-size 3-dimension loading problem. The optimization algorithms are very complicated, and are not suitable for the fix-size loading problem. Loading problem is a combination optimization problem with complex constraints. It is a NP hard problem which difficult to solve. However, it's much easier for fix-size loading problem, which object is to load as much as possible. Currently there is little research focus on fix-size loading problem. George [11] proposed a Heuristic algorithm framework and give a upper bound for optimal, Sui et al. [12] describe a Heuristic approach by using nest loop in 2D matrix arrangement. For the cargo that side tumbling is allowed, Yang [13] introduced a mirror copy method for 2D arrangement, optimizing each possible layout, then optimize of trends in the three coordinate directions, choose the best coordinate direction to maximize the space utilization. Xu et al. [14] improved Yang's method, propose a simpler method, reduce 3-direction optimization to height direction optimization. In 2D arrangement optimization, they take the model proposed in Niu et al. [15]. However, practical loading needs consideration of the stability and convenience for loading and unloading. The algorithms above only consider maximize the volume of cargo accommodated, which is equivalent to the high space usage, the arrangement is usually irregular or instable, and is not good for loading and unloading. From the intuition of that objects should be kept even and stable, this chapter propose an Heuristic Algorithm of Stacking Layer for the Three-Dimensional Packing of Fixed-Size Cargoes, which is convenient and stable for loading and unloading.

64.2 Problem Definition

There are two research directions for loading problem, one is focus on minimizing the number of containers, the other one is focus on the arrangement of objects in a container, try to maximal the total volume of cargo for each container. For the objects with fix-size, the main consideration is how to fill the car to maximize the objects in the car, such that the number of cars is minimal. So for fix-size loading problem, the two directions have the same goal, we only need to find a best arrangement of objects.

The fix-size of object is a box, so the problem can be described as: given a infinite set of boxes with fix-size, try to find a best loading method to a container with known size, to maximize the efficiency of the loading space utilization (or the total volume of cargo), with the consideration of the convenience and stability of loading and unloading. There are several assumptions in this following discussion:

- (1) the box size is less than the container size;
- (2) boxes can be arbitrary rotated and arranged in the container;
- (3) the boxes can be stacked.

64.3 Stacking Layer Method Design

64.3.1 The Concept of Stacking Layer Method

Stacking layer method simulates the idea that each layer is ensured to be more “straight” during the real packing process. Each pile is packed along the same direction, and each layer cannot be filled until the previous layer is fully filled. Each layer must be smooth and tidy. There are two categories according to the direction when packing: the parallel layer stacking and vertical layer stacking. Parallel layer stacking keeps each layer of the cargo paralleled with the compartment door. And vertical layer stacking keeps each layer vertical to the compartment door. The vertical layer stacking method in the chapter specifically means that the stacking is along the direction of the compartment height.

The compartment door of the truck is usually set in three ways:

- (1) in the rear compartment;
- (2) in the side;
- (3) two doors both in the rear and the side.

Vertical layer stacking is preferred for the stability of the load because the load-bearing surface holds the whole bottom of the compartment. And for the parallel layer stacking, each layer bears a much smaller area that only a small strip of the bottom of the compartment is taken. But for (1) and (2), workers need to load and unload the cargos layer by layer since there is only one door and step on the cargos when working on the layers far away from the door. Loading and unloading cannot be done if there is not enough space in the compartment. Therefore, in order to facilitate loading and unloading easily, parallel layer stacking method is more appropriate. The layer farthest from the door can be filled first, which means the filling is from inside out. And the cargos can be unloaded from outside layer by layer during the distribution. Because there are two doors in (3), it is much easier when loading and unloading. The layers far away from the back door can be load and unloaded through the side door, so vertical layer stacking method is preferred.

64.3.2 The Two-Step Solving Algorithm for Stacking Layer Method

Based on the earlier packing algorithms, a two-step solving algorithm for stacking layer method is introduced in this chapter. First, a triangular combinatorial

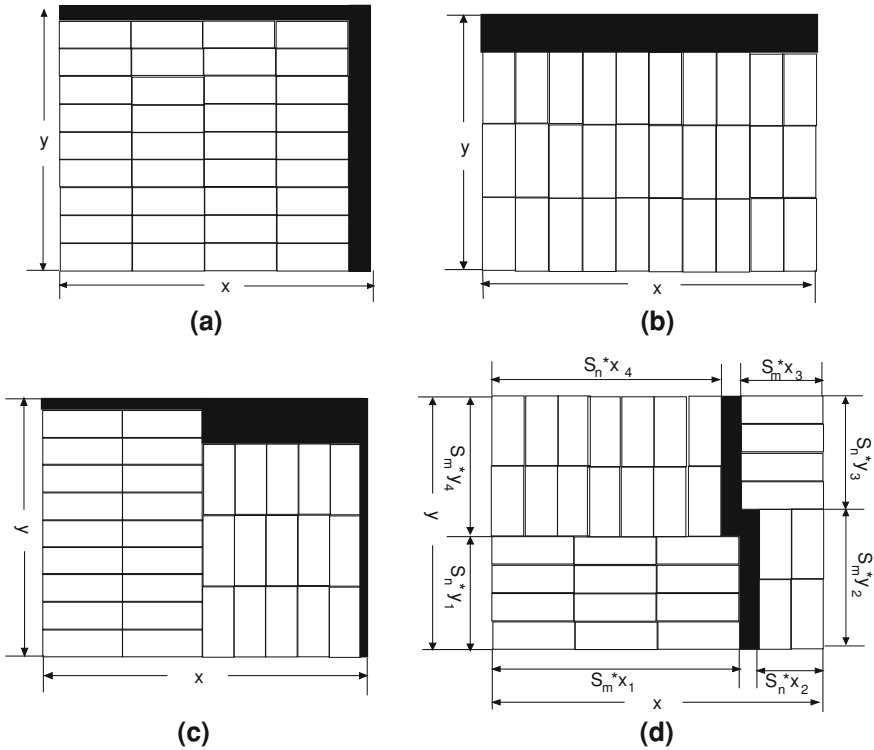


Fig. 64.1 General packing model. **a** Stacking Model 1. **b** Stacking Model 2. **c** Stacking Model 3. **d** Stacking Model 4

optimization along a certain direction of the compartment is operated. Then the layout is further optimized. Solving the first step is similar to the one-dimensional cutting stock problem. The length, width and height of the compartment is ordered from largest to smallest, denoted as s_1, s_2 and s_3 ($s_1 \geq s_2 \geq s_3$). The loading direction of each layer is denoted as z (stands for the length or width or height of the compartment). We seek for the linear combination of s_1, s_2 and s_3 that minimizes the remaining space along the loading direction. The objective function is

$$\text{Min } FZ = Z - (a \times s_1 + b \times s_2 + c \times s_3) \tag{64.1}$$

which subjects to the constraints condition of $0 \leq a \leq \lfloor z/s_1 \rfloor, 0 \leq b \leq \lfloor z/s_2 \rfloor, 0 \leq c \leq \lfloor z/s_3 \rfloor$ (a, b, c are integers), where FZ is the remaining space and “ $\lfloor \rfloor$ ” means to round down.

Step two is actually a layout problem that the layout space is located in (x, y) and the objective rectangular to be filled is (s_m, s_n) . Suppose $s_m \geq s_n$, there are four types of stacking models illustrated in Fig. 64.1. The black filler represents the remained gap.

In fact, the former three stacking models are the special case of the fourth packing model. When $x_2 \sim x_4, y_2 \sim y_4$ or $x_1, x_2, x_4, y_1, y_2, y_4$ or x_2, x_4, y_2, y_4 are all

zeros, it becomes to stacking model 1; when $x_1, x_3, x_4, y_1, y_3, y_4$ or $x_1, x_2, x_3, y_1, y_2, y_3$ or x_1, x_3, y_1, y_3 are all zeros, it becomes to stacking model 2; when x_3, x_4, y_3, y_4 or x_1, x_2, y_1, y_2 are all zeros, it becomes stacking model 3. So the stacking method can be summarized as: find the parameters in packing model 4 $x_1 \sim x_4, y_1 \sim y_4$ to minimize the remaining space and maximum the number of rectangles embedded whose edges are donated as s_m and s_n .

Stacking method 4 combines the length of side edge to fill the space, which is equivalent to variety stacking models (see Ref. [14]) and mutation models. The stacking model in [12] is a mutation model of stacking method 4, which is similar to this chapter. But in [12], y_2 is not bind to $s_n \times y_1$ and x_3 is not bind to $s_n \times x_2$. But they are bind to each other in this chapter, which is more consistent to the idea that people maintain the stability, tidily and smooth in the actual loading. In the stacking model of this chapter, if the parameters x_1, y_1 are know, then

$$x_2 = \lfloor (x - s_m x_1) / s_n \rfloor \quad (64.2)$$

$$y_4 = \lfloor (y - s_n y_1) / s_m \rfloor \quad (64.3)$$

When using parallel stacking method, to maintain the stability of the loading, there is

$$y_2 = \min(\lceil s_n y_1 / s_m \rceil, \lfloor y / s_m \rfloor) \quad (64.4)$$

“ $\lceil \rceil$ ” means to round up. To use the smaller value is to ensure that y_2 is not greater than $\lfloor y / s_m \rfloor$.

When $y_2 = \lfloor y / s_m \rfloor$, there is $x_3 = y_3 = 0$ and

$$x_4 = \lfloor (x - s_n x_2) / s_n \rfloor \quad (64.5)$$

Otherwise

$$y_3 = \lfloor (y - s_m y_2) / s_n \rfloor \quad (64.6)$$

and for the pile of (x_3, y_3) , when $s_n x_2 - s_m \lfloor s_n x_2 / s_m \rfloor \leq s_m / 2$, if we load cargos horizontally, the center of gravity will fall outside of the cargo supporting surface (x_2, y_2) , so we let

$$x_3 = \lfloor s_n x_2 / s_m \rfloor \quad (64.7)$$

where x_4 is equal to (64.5). When the gap between pile (x_3, y_3) and pile (x_4, y_4) is satisfied with $x - s_m x_3 - s_n x_4 > s_n$, the cargos can be placed vertically (as illustrated in Fig. 64.2), we name this cargo heap as pile (x_5, y_5) .

64.4 Space Utilization Test

In this section, the space utilization of packing with the layer stack heuristic algorithm is mainly tested. We realize the algorithm combining C# with Matlab2009a and run the program on a PC with the Core Duo processor 2.27 GHz,

Fig. 64.2 Example of appearing (x_5, y_5) cargo heap

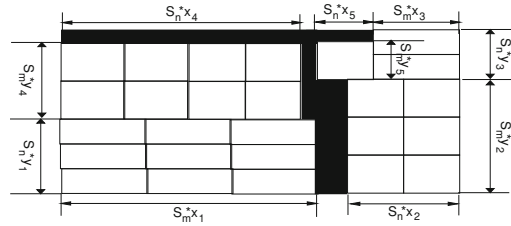


Table 64.1 Results comparison of some algorithms

Size of container	Size of box	Trivial algorithm		Great fox software		[11]’s heuristic	
		Numbers	Spc util (%)	Numbers	Spc util (%)	Numbers	Spc util (%)
5800,	390,320,310	756	89.49	804	95.17	806	95.41
2300,	510,330,290	560	83.63	640	95.57	645	96.32
2450	530,310,470	390	92.15	401	94.75	402	94.98
	560,370,310	432	84.9	490	96.30	492	96.69
	600,530,310	288	86.87	312	94.1	316	95.31
1201,	60,50,30	644	86.66	692	93.12	700	93.92
233,	60,50,40	460	82.50	524	94.02	524	94.02
239	61,56,39	476	94.82	480	95.62	487	97.01
	63,87,32	316	82.87	361	94.67	361	94.67
	87,68,52	172	78.81	196	90.16	212	93.1

4G of memory. Table 64.1 compares the space utilization using the algorithm described in Literature [12], general batch method and major domestic packing software. Use our layer stack method to solve the above packing problem, and the result is shown in Table 64.2.

From Tables 64.1 and 64.2, we can find that the heuristic algorithm of literature [12] has the best solution with highest utilize rate and loading number. The utilize rate showed in Table 64.2 is generally higher than that of general batch method, and maximum utilization of each row is about the same or even higher than that of [12]. For example, when the size of container is (1201, 233, 239) and the size of the small box is (60, 50, 30), the maximum number obtained by our algorithm is 720, which is more than the result of [12] by 20, and the utilization rate is 96.89%, which is 20 more than the result of [12] in number and higher by 2.97%. Especially when the size of container is (1201, 233, 239) and the small box is (61, 56, 39), our algorithm achieves the highest utilization rate 99.20%.

Also, from Table 64.2, we can conclude that the utilization rate of our algorithm relates with the position of the compartment door. Different position of the door determines different direction of the layer, and further produce different utilization rate. In the view of the experiment, when packing the layer in the length direction of the compartment, the utilization rate will generally lower than the case along with the width and height direction. However, sometimes the layer packing

Table 64.2 Computing results of stacking layer method

Container	Size of Box	Back Door		Side Door		Back and Side Door	
		Numbers	Spc util (%)	Numbers	Spc util (%)	Numbers	Spc util (%)
5800,	390,320,310	782	92.57	801	94.82	806	95.41
2300,	510,330,290	580	86.61	636	94.98	632	94.38
2450	530,310,470	380	89.78	408	96.40	400	94.51
	560,370,310	486	95.51	492	96.69	483	94.92
	600,530,310	301	90.79	310	93.50	316	95.31
1201,	60,50,30	720	96.89	648	87.20	653	88.55
233,	60,50,40	500	89.71	524	94.02	525	94.20
239	61,56,39	498	99.20	488	97.21	487	97.01
	63,87,32	329	86.28	350	91.79	358	93.88
	87,68,52	174	80.04	196	90.16	191	87.86

along with the length direction may result in high utilization rate, for example in the experiment with the highest utilization rate 99.20% and the experiment in which the size of container is (1201, 233, 239), the small box is (60, 50, 30) and its highest utilization rate is 96.89%.

64.5 Conclusion

On packing the specifications of the goods, the previous algorithms only consider the maximization of space utilization. Although high utilization rate can be obtained, yet it is hard to achieve actually due to irregular arrangement. Our algorithm takes the convenience of loading before distribution, and the convenience of unloading in the delivery into account, and put forward the layer along with the direction of easy loading and unloading, according to the location of the compartment doors. If the car has only one door, make the layer parallel to the door, and if the car is equipped with two doors, make the layer perpendicular to the height direction of compartment door. The algorithm considers the actual packing factors better and is easier for loading and unloading. Compared with other packing algorithm, it is a more practical algorithm with high space utilization rate.

References

1. Bortfeldt A, Gehring H (2001) A hybrid genetic algorithm for container loading problem. *Eur J Oper Res* 131:143–161
2. Pisinger D (2002) Heuristic for the container loading problem. *Eur J Oper Res* 141:143–153
3. Bortfeldt A, Gehring H, Mack D (2003) A parallel tabu search algorithm for solving the container loading problem. *Parallel Comput* 29(5):641–662
4. Mack D, Bortfeldt A, Gehring H (2004) A parallel hybrid local search algorithm for the container loading problem. *Int Trans Oper Res* 11(5):511–533

5. Gendreau M (2006) A tabu search algorithm for a routing and container loading problem. *Transp Sci* 40(3):342–350
6. Crainic T, Perboli G, Tadei R (2009) TS2PACK: a two-level tabu search for the three-dimensional bin packing problem. *Eur J Oper Res* 195(3):744–760
7. Egeblad J, Pisinger D (2009) Heuristic approaches for the two- and three-dimensional knapsack packing problem. *Comput Oper Res* 36(4):1026–1049
8. Yang H, Shi J (2010) A Hybrid CD/VND algorithm for three-dimensional bin packing. In: *The 2nd international conference on computer modeling and simulation*. IEEE Press, Sanya
9. Almeida AD, Figueiredo MB (2010) A particular approach for the three-dimensional packing problem with additional constraints. *Comput Oper Res* 37(11):1968–1976
10. Bu L, Yin C, Pu Y (2002) A genetic and simulated annealing algorithm for optimal sequential casing of less-than-carload freights. *J Southwest Jiaotong Univ* 37(5):531–535
11. George JA (1992) A method for solving container packing for a single size of box. *J Oper Res Soc* 43(4):307–312
12. Sui S, Shao W, Gao Z (2005) A Heuristic algorithm for dimensional container packing problem of fixed-size cargoes. *Inf Control* 34(4):490–494
13. Yang D (2007) Optimum algorithm for container loading one type of objects. *J Trans Eng Inf* 5(2):17–23
14. Xu L, Ji Z, Xia J (2008) The optimum algorithm for the container loading problem with homogeneous cargoes. *J Shangdong Univ (Eng Sci)* 38(3):1–4
15. Niu Y, Fan Y, Xu E (2004) Method of container loading in order model. *Logist Tech* 5:47–49

Chapter 65

Research on a Novel Hybrid Optimization Algorithm Based on Agent and Particle Swarm

Wei jin Jiang, Xingjun Jiang and Lina Yao

Abstract A novel multi-agent particle swarm optimization algorithm (MA-PSO) is proposed for optimal reactive power dispatch and voltage control of power system. The method integrates multi-agent system (MAS) and particle swarm optimization algorithm (PSO). An agent in MA-PSO represents a particle to PSO and a candidate solution to the optimization problem. All agents live in a lattice-like environment, with each agent fixed on a lattice-point. In order to decrease fitness value, quickly, agents compete and cooperate with their neighbors, and they can also use knowledge. Making use of these agent interactions and evolution mechanism of PSO. MA-PSO realizes the purpose of ' minimizing the value of' objective function. MA-PSO applied for optimal reactive power is evaluated on power system. It is shown that the proposed approach converges to better solutions much faster than the earlier reported approaches.

Keywords Multi-agent system · Particle swarm optimization · Reactive power optimization · Hybrid optimization method

65.1 Introduction

The Multi-agent system (MAS) is a network architecture composed by those Agent which has many loose couplings, has the sensation ability, the question solution ability can correspond and interactive with other Agent in the system [1].

W. Jiang (✉)

School of Computer and Electronic Engineering, Hunan University of Commerce,
Changsha, 410205, China
e-mail: ljxwhj@163.com

X. Jiang · L. Yao

Department of Computer, Hunan Radio and TV University, Changsha, China

These Agent in physics or the logic is dispersible, its behavior is autonomous, they completed the complex control task or the solution complex question through the consultation, the coordination and the cooperation. Particle swarm optimizes algorithm (PSO) is one kind of the evolve technical, the source is from the research on the behaviors of preys of the bird group, it essentially belongs to the iteration the stochastic search algorithm, has the parallel processing characteristic, robustness good, is easy to realize, in the principle may find the optimized question by a bigger probability. The overall situation can be optimal solution, the counting yield is higher too, it had been successfully applied to solves each kind of complex optimized question [2]. Like, the PSO algorithm is introduced the electrical power system, the literature [2–4] separately uses into solve the unit to combine, the plan of electrical network expands, as well as compensates the capacitor to optimize electrical power system and so on as this kind of electrical power system optimization question, all has obtained a ideal effect. This article unify PSO and the MAS technology structure one brand-new algorithm: multi-agent and particle swarm optimizes the algorithm (MA-PSO) to solve the electrical power system idle work optimization question. First structure a lattice-like environment, all Agent are survives in this environment, each Agent is equal to a particle in the particle swarm optimized algorithm, they are fixed in a lattice, through competition and the cooperative operation with neighbor and from the study operation, unifies the PSO algorithm the evolution mechanism, unceasingly through Agent alternately mutual influence between Agent and the environment, to renews each Agent in the solution space position, enables its quickly, precisely to restrain to the overall situation optimal solution [5]. Finally has carried on the simulation computation take the IEEE 30 pitch points systems as the experimental system, and carries on the comparison with other some methods optimized result, finally indicated this algorithm has the convergence rate to be quick, the computation precision high prominent merit.

65.2 Hybrid Optimization Algorithm

The PSO algorithm is the people receives one superior algorithm enlighten and proposed by the behavior which in the real world is the bird group search food, to revises the individual motion strategy by summarizes through community's between information sharing and the individual own experience, finally strives for to take the solution of optimized question [6]. PSO initial changes into crowd of stochastic granules, then found the optimal solution through the iteration. In each iteration, the granule through tracks two “the extreme value” to renew oneself. The first extreme value is called individual extreme value P_{Best} . Another extreme value is an optimal solution which entire population at present found, this extreme value is overall situation extreme value g_{Best} . Each granule basis following formula renews own speed and in the solution space position

$$v_{d+1} = w \cdot v_d + \varphi_1 \cdot rand() \cdot (p_{Best} - x_d) + \varphi_2 \cdot rand() \cdot (p_{Best} - x_d) \quad (65.1)$$

$$x_{d+1} = x_d + v_{d+1} \quad (65.2)$$

Among them, the subscript d expression iteration number of times, x_d expressed time the d th iteration granule space position, v_d expressed time the d th iteration granule speed, w is the inertia constant, φ_1, φ_2 is the study factor, $rand()$ is situated between (0, 1) between random number.

Agent is one kind of system which has the sensation ability, the question solution ability, can correspond and interactive with other Agent in the system, thus completes software entity with one or many functions goals. Agent usually specifically serves following several typical characteristics [7]:

- (1) Agent is usual stay in a specific environment, and Agent only can work in this environment.
- (2) Agent can sense partial environment of itself.
- (3) Agent should have the good ability of autonomy, has the domination to its own behavior or the movement, independently completes its specific task without exterior intervention.
- (4) Agent should have the reacting capacity of sense environment and make the corresponding movement.

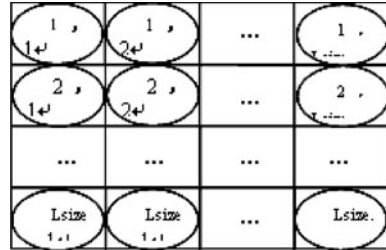
MAS is the network architecture which is composed by many loose couplings Agent. These Agent in physics or the logic is dispersible, its behavior is autonomous, they complete the complex control task or the solution complex question through the consultation, the coordination and the cooperation. In usual situation, MAS system needs to define following four elements when solution question: (1) Each Agent's intention and goal; (2) Agent's "housing" environment; (3) Because each Agent merely can sense its own partial environment, therefore should define each Agent's partial environment; (4) In order to realize the Agent own intention and the goal, each Agent can adopt motion strategy.

Under will act according to these four essential factors to come the MA-PSO algorithm which specify this article proposed.

65.2.1 MA-PSO Algorithm

The MAPSO algorithm is unified the PSO algorithm and the MAS main characteristic to structure one brand-new algorithm, first is to structure Agent survival environment, each Agent not only may competition and the cooperation operation with the neighbor, from the study operation, moreover also has absorbed the PSO algorithm evolution mechanism, carries on information sharing with overall situation with most superior Agent, and summarizes according to own experience revises Agent the motion strategy, enables its quickly, precisely to restrain to the overall situation optimal solution.

Fig. 65.1 Structure of the agent environment



(1) Agent intention’s definition

In the MAPSO algorithm, supposition its anyone Agent is α it is equal in the PSO algorithm a granule, it has an optimized adaption value which the question decided. When solution the idle work’s optimization question, the Agent α adaption value decided by the under type, namely

$$f(\alpha) = F_Q \tag{65.3}$$

The Agent α ’s goal is in satisfies the movement condition under the limit to reduce its adaption value as far as possible. In order to realize its goal, Agent will act according to the environment which oneself will locate to make the corresponding movement the response, the maximization reduces its adaption value.

(2) Environment definition

Regarding the environment definition, this article has designed one kind of center extremely simple lattice structure environment, like Fig. 65.1 shows. Each Agent all “housing” in this environment, and is fixed in its small check. Figure 65.1 center each circle represents Agent, in the circle data represents Agent in the environment position, each Agent own should contain two data, include algorithm’s each granule speed and position in PSO. L_{size} is a positive integer, may know from the chart, the always standard fraction is $L_{size} \times L_{size}$, it is equal in the PSO algorithm the population law.

(3) Agent partial environment’s definition

In the MAS system, Agent usually can sensate independently to adopt the motion strategy in the partial environment information to complete its intention and the goal, therefore the Agent partial environment definition is quite important [8]. In this article structure MAS, supposition L_{ij} is the check coordinates for (i,j) Agent, $i, j = 1, 2, \dots, L_{size}$, L_{ij} neighbor N_{ij} is defined as

$$N_{ij} = \{L_{i1j}, L_{ij1}, L_{i2j}, L_{ij2}\} \tag{65.4}$$

In the formula $i_1 = \begin{cases} i - 1i \neq 1 \\ L_{size}i = 1 \end{cases} \quad j_1 = \begin{cases} j - 1j \neq 1 \\ L_{size}j = 1 \end{cases}$

$$i_2 = \begin{cases} i + 1i \neq 1 \\ 1i = L_{size} \end{cases} \quad j_2 = \begin{cases} j + 1j \neq 1 \\ 1j = L_{size} \end{cases}$$

In the MAPSO algorithm, the type (65.4) has defined each Agent four neighbors, they have composed this Agent partial environment.

(4) Agent motion strategy

In the MAPSO algorithm, each Agent first acts according to its partial environment if the neighbor to carry on the competition and the cooperation operation. Supposition Agent L_{ij} , also $L_{ij} = (l_1, l_2, \dots, l_n)$ is it in the optimization asked in the exercise key space the position, M_{ij} is in the L_{ij} four neighbors has the smallest adaption valueAgent, also $M_{ij} = (m_1, m_2, \dots, m_n)$. If L_{ij} satisfying -like (65.5), then it is a winner, otherwise it is Loses.

$$f(L_{i,j}) \leq f(M_{i,j}) \quad (65.5)$$

If L_{ij} is a winner, it is invariable in the solution space maintain position. If it will be Loses, L_{ij} (65.6) is changed in the solution space position according to the type.

$$l_k = m_k + rand(-1, 1)(m_k - l_k) \quad k = 1, 2, \dots, n \quad (65.6)$$

The in the formula $rand(-1, 1)$ is situated between $(-1, 1)$ between the random number. If, then $l_k = x_{k \min}$; If $l_k < x_{k \max}$, then $l_k = x_{k \max}$. $l_k < x_{k \min}$. $x_{\min} = (x_{1 \min}, x_{2 \min}, \dots, x_{n \min})$ is optimizes the question feasible solution space the under limiting value, $x_{\max} = (x_{1 \max}, x_{2 \max}, \dots, x_{n \max})$ is above the limiting value. Equation (65.6) may know from the type, if Agent is Loses, it still retained own original useful information, moreover also fully has absorbed the neighbor M_{ij} beneficial information, causes it further to reduce its adaption value.

Agent with the neighbor's competition and the cooperation operate is equal in the Agent environment gradually transmission, but this kind of information transmission merely limits to its partial environment, the useful information transmission speed relative is slower. The PSO algorithm inspiration, Agent not only if the neighbor is carried on the competition and the cooperation operation, moreover also carries on the exchange of information with overall situation most superior Agent, and summarizes according to own experience revises Agent the motion strategy, like this will speed up Agent in the entire environment, information transmission speed, will enhance its algorithm the astringency.

The method proposed in this article, straightly takes over the use of the PSO algorithm formula (65.1), (65.2) carries on the renewal to each Agent in the solution space position, in the formula P_{Best} is the optimal solution individual extreme value which Agent at present found, g_{Best} is in the entire environment at present found optimal solution Agent the overall situation extreme value.

(5) Agent from study operation

Agent may not only carry on the competition and operation in its partial environment with the neighbor, moreover also may study through the knowledge which own has, further enhances the ability for solute question. The literature [9] center used the small scope heredity algorithm to obtain a better effect for its partial search. Receives this method of the inspiration, this article applies the small scope of the search technology to realize Agent from the study function.

Supposition Agent $L_{i,j}$, in the solution space position is $L_{i,j} = (l_1, l_2, \dots, l_n)$, structure is similar the size which shows to chart 1 for $sL_{size} \times sL_{size}$ Agent environment. In chart each $sL_{i,j}$, $i, j = 1, 2, \dots, sL_{size}$ (65.7) obtains in the solution space position according to the type.

$$sL_{i,j} = \begin{cases} L_{i,j} & i = 1, j = 1 \\ LL_{i,j} & other \end{cases} \quad (65.7)$$

The in the formula $LL_{i,j} = (ll_{i,j,1}, ll_{i,j,2}, \dots, ll_{i,j,n})$, also $ll_{i,j,k}$ (65.8) calculates based on the type obtains

$$ll_{i,j,k} = \begin{cases} x_{k \min} l_k \cdot rand(1 - sR, 1 + sR) < x_{k \min} \\ x_{k \max} l_k \cdot rand(1 - sR, 1 + sR) < x_{k \max} \\ l_k \cdot rand(1 - sR, 1 + sR) & other \end{cases} \quad (65.8)$$

In the formula sR is the small scope partial search radius, also $sR \in [0, 1]rand(1 - sR, 1 + sR)$ is situated between $(1 - sR, 1 + sR)$ between the random number.

In the study operation, Agent $L_{i,j}$ the solution space position current is its knowledge, according to the type (65.7), (65.8) structure other MAS, its essence was take sR as the partial search radius, small scope expansion this search space. Because in this MAS, the Agent number relative are less, it only needs to carry on the competition and the cooperation operation with the neighbor, Agent can fast the information transmission in the entire environment, no longer introduces the PSO evolution mechanism, effective saves the computation cost. After a s_{Gen} iteration, seeks has its smallest adaption value Agent $L_{i,j}$ substitutes $L_{i,j}$, has completed $L_{i,j}$ from the study operation.

65.3 Calculates Example

Use The MA-PSO algorithm which proposed this article applies in a electrical power system idle work optimization, carries on the idle work optimization computation to the IEEE30 pitch point demonstration system, confirms the MA-PSO algorithm the optimized effect. In this system has 41 legs, six generators pitch points and 22 loads pitch points. These six generators pitch point is 1, 2, 5, 8, 11, 13, the pitch point 1 took the balance pitch point and other are the P-V pitch point, in the system other pitch points is the P-Q pitch points; May adjust the transformer leg is 6-9, 6-10, 4-12, 27-28; Shunted capacitor pitch point 3, 10, 24. System always load $P_{load} = 2.834$, $Q_{load} = 1.262$. The system parameter Ref. [10], various variables about limiting value like Tables 65.1, 65.2 and 65.3 shows. In MA-PSO algorithm L_{size} usually may the value be 3-10, this article $L_{size} = 6$, biggest iteration number of times $T_{max} = 300$, inertia constant $w = 0.7298$, study factor $\varphi_1 = \varphi_2 = 1.49618$, from study operation in $sL_{size} = 3$, partial search radius $s_R = 0.2$ biggest iteration number of times $s_{Gen} = 10$.

Table 65.1 Parameters and limits of generations

Number	Q_{Gmax}/pu	Q_{Gmin}/pu	V_{Gmax}/pu	V_{Gmin}/pu
1	0.596	-0.298	1.1	0.9
2	0.480	-0.24	1.1	0.9
5	0.6	-0.3	1.1	0.9
8	0.53	-0.265	1.1	0.9
11	0.15	-0.075	1.1	0.9
13	0.155	-0.078	1.1	0.9

Table 65.2 Limits of PQ nodes and transformer taps

$V_{load.max}/pu$	$V_{load.min}/pu$	$Tk.max/pu$	$Tk.min/pu$
1.05	0.95	1.05	0.95

Table 65.3 Limits of reactive power sources and voltage

$Qc.max/pu$	$Qc.min/pu$	$Vc.max/pu$	$Vc.min/pu$
0.36	-0.12	1.05	0.95

Table 65.4 Comparison of optimal results for different methods

	$\sum P_G/pu$	$\sum Q_G/pu$	Ploss/pu	Qloss/pu	Psave/pu	$\eta_{SAVE}/\%$
Broyden	2.88986	0.93896	0.055860	-0.32304	0.00402	6.71000
SGA	2.88380	1.02774	0.049800	-0.23426	0.01008	16.8400
AGA	2.88326	0.66049	0.049260	-0.60151	0.01062	17.7400
EP	2.88362	0.87346	0.046930	-0.38527	0.01025	17.1200
PSO	2.88330	0.82500	0.049262	-0.22920	0.01062	17.6200
MA-PSO	2.88270	0.81950	0.048747	-0.22836	0.01113	18.5900

In the initial condition, establishes the generator machine the terminal voltage and the transformer changes compares is 1.0, through the tidal current computation, obtains $\sum P_G = 2.893857$, $\sum Q_G = 0.980199$, $P_{loss} = 0.059879$. Three P-Q node voltage Highest limit respectively is: $V_{26} = 0.932$, $V_{29} = 0.940$, $V_{30} = 0.928$; An idle work electricity generation power Highest limit is: $Q_G = 0.569$.

In order to confirm the validity which the algorithm proposed in this article, separately with the auto-adapted heredity algorithm (AGA) [10], the standard heredity algorithm (SGA) [10], the evolution algorithm (EP) [11] and the Broyden nonlinear programming method [12] carried on the comparison, Table 65.4 has produced the most superior optimized result which each algorithm separately moved 30 time obtained arrives. May see from Table 65.4, after this article proposed the MA-PSO algorithm idle work optimizes the system net which obtains to

damage P_{loss} is 0.048747 and the net damages the descending rate η_{SAVE} is 18.5900%, this result all must be better than other several optimized method, fully had demonstrated this article proposed the method validity, can obtain a better optimal solution.

65.4 Conclusions

This article proposed one brand-new algorithm-MAPSO algorithm. A this method structure MAS environment, each Agent was in a PSO algorithm granule, they carried on the competition and the cooperation operation and from the study operation in this environment with the neighbor, and unified the PSO algorithm evolution mechanism, enabled its to be fast, accurately to restrain to the overall situation optimal solution. Through specifically calculated the example indicated this method has the very good counting yield and restrains the stableility. Therefore took one kind of brand-new attempt, this method to solves the electrical power system to have the highly complex restraint condition the combination optimization question to have the count for much inspiration significance.

Acknowledgments This work is supported by the Natural Science Foundation of Hunan Province of China No. 10JJ5064, the Sci. and Technology Plan Project of China under of Hunan Grant No.2009GK2002, and the Technology Innovation Foundation under Science and Technology Ministry Grant No. 09C26214301947.

References

1. Bredin J, Kotz D, Rus D, Maheswaran RT, Imer C, Basar T (2003) Computational markets to regulate mobile-agent systems. *Auton Agents Multi-Agent Syst* 6(3):235–263
2. Archer A, Tardos E (2001) Truthful mechanisms for one-parameter agents [A]. In: *Proceedings of the 42nd IEEE symposium on foundations of computer science [C]*, IEEE Computer Society Press, Lasvegas, USA, pp 482–491
3. Kennedy J, Eberhart R (1995) Particle swarm optimixation [C]. In: *Proceedings of IEEE international conference on neural networks, Perth, Australia, vol 4, pp 1942–1948*
4. Hu J, Guo C, Cao Y (2004) A hybrid particle swarm optimization method for unit commitment problem [J]. *Proc CSEE* 24(4):24–28
5. Yu X, Li Y, Xiong X et al (2003) Optimal shunt capacitor placement using particle swarm optimization algorithm with harmonic distortion consideration [J]. *Proc CSEE* 23(2):26–31
6. Eberhart RC, Shi Y (2008) Comparing inertia weights and constriction factors in particle swarm optimization [C]. In: *Proceedings of the congress on evolutionary computing*, IEEE Service Center, California, USA, pp 84–88
7. Liu J, Jing H, Tang YY (2002) Multi-agent oriented constraint satisfaction [J]. *Artif Intell* 136(1):101–144
8. Zhong W, Liu J (2004) A multiagent genetic algorithm for global numerical optimization [J]. *IEEE Trans Syst Man Cybern* 34(2):1128–1141
9. Jiang WJ, Wang P, Zhang L (2009) Research on grid resource scheduling algorithm based on MAS cooperative bidding game. *Chin Sci F* 52(8):1302–1320

10. Wu QH, Cao YJ, Wen JY (1998) Optimal reactive power dispatch using an adaptive genetic algorithm [J]. *Int. J. Electr Power Energy Syst* 20(8):563–569
11. Lai LL, Ma J (1997) Application of evolutionary programming to reactive power planning-comparison with nonlinear programming approach [J]. *IEEE Trans Power Syst* 12(1):198–204
12. Das DB, Patvardhan C (2006) Reactive power dispatch with a hybrid stochastic search technique [J]. *Int J Electr Power Energy Syst* 24(9):731–736

Chapter 66

A Novel BCC Algorithm for Function Optimization

Jia-Ze Sun, Guo-Hua Geng, Ming-Quan Zhou and Wang Shu-Yan

Abstract Aiming at improving the performance of bacterial colony chemotaxis (BCC) optimization algorithm, a novel bacterial colony chemotaxis algorithm is introduced through integrating chaotic optimization into bacterial colony chemotaxis (NBCC) optimization algorithm, it greatly enhances the local searching efficiency and global searching performance. Simulation results on standard test functions show that NBCC is pretty efficient to solve complex problems.

Keywords Bacterial colony chemotaxis · Function optimization · Kent map

66.1 Introduction

In the field of optimization algorithm, many researchers have been inspired by biological processes to develop new optimization algorithm methods such as particle swarm optimization algorithms (PSO). The swarm intelligence

J.-Z. Sun (✉) · W. Shu-Yan
School of Computer Science and Technology, Xi'an University of Post
and Telecommunications, Xi'an 710061, China
e-mail: sunjiaze@126.com

W. Shu-Yan
e-mail: wsylxj@126.com

J.-Z. Sun · G.-H. Geng
Institute of Visualization Technology, Northwest University, Xi'an 710127, China
e-mail: ghgeng@nwu.edu.cn

M.-Q. Zhou
School of Information Science and Technology, Beijing Normal University,
Beijing 100875, China
e-mail: mqzhou@nwu.edu.cn

optimization algorithms have successfully been used in many fields such as function optimization, training of neural networks, fuzzy control system and so on.

Li et al. [1] present bacterial colony chemotaxis (BCC) algorithm, based on bacterial chemotaxis (BC) algorithm [2]. BCC algorithm greatly improves the convergence speed and accuracy of BC algorithm and makes it comparable to many other well-used intelligent optimization methods.

In this paper, a novel bacterial colony chemotaxis (NBCC) algorithm based on Kent map is introduced through the technique of hybrid algorithm. By integrating chaotic optimization into bacterial colony chemotaxis optimization algorithm, it greatly enhances global searching performance. Simulation results on standard test functions show that NBCC is pretty efficient to solve complex problems.

66.2 Related Work

66.2.1 BC Algorithm Principle

Bacterial chemotaxis (BC) [2] algorithm is built up by simulating the movement of a single bacterium. Bacteria are single-cell organisms, but they acquire information about their environment, and use this information to survive.

It is proved that bacteria do share information among each other, but individuals and social interaction among bacteria differ from the interaction models for the behavior of social insects. The scientists construct an optimization algorithm based on the simplicity and robustness of the process of bacterial chemotaxis. For optimization purposes, the scientists study microscopic models that consider the chemotaxis of a single bacterium. Several features are added to the basic algorithm using evolutionary concepts in order to obtain an improved optimization strategy, called the bacteria chemotaxis (BC) algorithm [2]. Take 2D system for a minimum point as an example, the BC algorithm basic steps is following:

- (1) Compute the velocity v . In the model, the velocity is assumed to be a scalar constant value.

$$v = \text{const} \quad (66.1)$$

- (2) Compute the duration of the new direction τ from the distribution of a random variable with an exponential probability density function

$$P(X = \tau) = \frac{1}{T} e^{-\tau/T} \quad (66.2)$$

Where the expectation value $\mu = E(X) = T$ and the variance $\sigma^2 = Var(X) = T^2$ The time T is given by

$$T = \begin{cases} T_0, & \text{for } \frac{f_{pr}}{l_{pr}} > 0 \\ T_0 \left(1 + b \left| \frac{f_{pr}}{l_{pr}} \right| \right), & \text{for } \frac{f_{pr}}{l_{pr}} < 0 \end{cases} \tag{66.3}$$

Where T_0 minimal mean time;

$f(x_1, x_2)$ function to minimize (2-D);

f_{pr} difference between the actual and the previous function value;

$l_{pr} = |\vec{x}_{pr}|$, \vec{x}_{pr} vector connecting the previous and the actual position in the parameter space;

b dimensionless parameter;

- (3) Compute the new direction. The probability density distribution of the angle α between the previous and the new direction is Gaussian and reads, for turning right or left respectively

$$P(X = \alpha, v = \mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[-\frac{(\alpha - v)^2}{2\sigma^2} \right] \tag{66.4a}$$

$$P(X = \alpha, v = -\mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[-\frac{(\alpha - v)^2}{2\sigma^2} \right] \tag{66.4b}$$

Where the expectation value $\mu = E(X) = 62^\circ$ and the standard deviation $\sigma = \sqrt{Var(X)} = 26^\circ$ referring to the chemotaxis measurements of the bacterium *Escherichia coli* by Berg and Brown $\alpha \in [0^\circ, 180^\circ]$. The choice of a right or left direction as referring to the previous trajectory is determined using a uniform probability density distribution, thereby yielding a probability density distribution for the angle of α

$$P(X = \alpha) = \frac{1}{2} \cdot [P(X = \alpha, v = \mu) + P(X = \alpha, v = -\mu)]. \tag{66.5}$$

We formulate the following variation of the expectation value and the variance, which is applied only if $\frac{f_{pr}}{l_{pr}} < 0$

$$\mu = 62^\circ (1 - \cos(\theta)) \tag{66.6}$$

$$\sigma = 62^\circ (1 - \cos(\theta)) \tag{66.7}$$

With

$$\cos(\theta) = e^{-\tau_c \tau_{pr}} \tag{66.8}$$

where τ_c is the correlation time, and τ_{pr} the duration of the previous step. Equation (66.8) is based on the first-order approximation of the average value of the cosine of the angle between successive trajectories at times t and $t + \tau_c$ as a function of t . It provides us with a statistical measure of direction-duration dependencies in experimental observations of BC. Note, however the difference between the formulation in (66.8) which the duration of the previous trajectory is used and formulation

$$\cos(\theta) = e^{-\tau_c \tau} \quad (66.9)$$

in which the actual trajectory duration is used, as found in [5, 6]. With (66.8), which is used in the following, the probability distribution of the angle becomes dependent on the duration of the previous step, which is contradictory to the assumption made by Dahlquist et al. [3]. In Sect. 66.4, we discuss results of the application of both algorithms on a number of test problems.

As soon as α is computed, it is possible to obtain the normalized new displacement vector \vec{n}_μ with unit length.

(4) Compute the new position. The length of the path l is given by

$$l = v\tau. \quad (66.10)$$

The normalized new direction vector \vec{n}_u with $|\vec{n}_u| = 1$ is multiplied by l to obtain the displacement vector \vec{x}

$$\vec{x} = \vec{n}_u l \quad (66.11)$$

such that the new location of the bacterium is

$$\vec{x}_{new} = \vec{x}_{old} + \vec{n}_u l \quad (66.12)$$

In summary, the algorithm contains the following parameters: the minimal mean time \mathbf{T}_0 , the dimensionless gradient parameter \mathbf{b} , the correlation time τ_c , and the velocity v . They constitute the optimization strategy parameters that are adjusted as described in the following sections.

66.2.2 BCC Algorithm Principle

BCC algorithm is one of novel heuristic colony intelligent optimization algorithm, and it is gained by establishing information interaction between individual bacterial. It is supposed that the bacterium has a sense limit in its environment. BCC algorithm basic steps can see [1].

66.2.3 Chaos Search Strategy

Chaos is a kind of universal nonlinear phenomena in many systems [4]. Chaotic movement is characterized by ergodicity, randomness and regularity. So chaotic movement could go through every state in certain scale according to its own regularity and ergodicity. It can be introduced into the optimization strategy to accelerate the optimum seeking operation and find the global optimal solution.

One-dimensional Kent mapping function [4]:

$$y_{n+1} = \begin{cases} y_n/a; & 0 \leq y_n < a \\ (1 - y_n)/(1 - a); & a \leq y_n \leq 1 \end{cases} \quad (66.13)$$

is a typical Chaos system, where $a \in (0, 1)$ is bifurcation parameter, $y_n \in [0, 1]$ is the sequence value. Starting the iteration from the random initial value, with the process of taking a series value of from small to large, the results of the initial iteration period doubling bifurcation will occur. When equals 0.4997, the target of Kent sequence achieve to the most excellent, and generate chaos sequence [4].

66.3 Algorithm

66.3.1 Overview

Directly using traditional BCC algorithm to solve the function optimization problem, BCC algorithm has good optimization capabilities, convergence speed, high precision optimization and its performance is better than BC algorithm. But for some multimodal function which has not obvious gradient change, the bacterium will get into the local optimum easily.

In this chapter, a novel bacterial colony chemotaxis (NBCC) algorithm is introduced through the technique of hybrid algorithm. By integrating chaotic optimization into bacterial colony chemotaxis optimization algorithm, it greatly enhances the local searching efficiency and global searching performance.

66.3.2 Integrating Chaotic Optimization into BCC Algorithm

In order to maintain the diversity of bacteria, we add chaos search in original BCC algorithm, which is better at local searching. It would be effective and rational to combine chaos and original BCC algorithm to balance the local and global search. On one hand it can enhance the global search capabilities and get out of the local optimum easily. While on the other hand, it will not reduce the convergence speed and search accuracy at the same time.

When the center position of other bacteria which have better objective function value in the sense limit, the center position is expressed as follow:

$$center_position(i) = Kent()dis(xi, k, center(xi, k)) \quad (66.14)$$

(66.14) where $dis(xi, k, center(xi, k))$ is the distance between the bacteria i and the $center_position(i)$, and $Kent()$ is a chaotic sequence number meeting the Kent map equation in interval $(0, 1)$.

66.3.3 NBCC Algorithm for Function Optimization Problem

In summary, the novel NBCC algorithm for function optimization problem is as following:

Step 1: Initialize the number of bacteria colony, the position of individual bacterium and the sense limit, In particular, the initial position of individual bacterium is generated by Kent map based on chaotic sequence as mentioned above.

Step 2: In the initial conditions, the objective function fitness of individual bacterium i in the bacteria colony is calculated.

Step 3: The bacterium i which moves at the k th step, apperceives the information around it, identifies the center position of other bacteria which have better objective function fitness value in the sense limit. The center position is expressed as follow:

$$center_position(i) = Kent().dis(xi, k, center(xi, k)) \quad (66.14)$$

where $dis(xi, k, center(xi, k))$ is the distance between the bacteria i and the $center_position(i)$, and $Kent()$ is a chaotic sequence number is generated by Kent map based on chaotic sequence as mentioned above.

Step 4: The bacterium i which moves at the k th step, gains another position $bc_position(i)$ according to single bacterium BC algorithm.

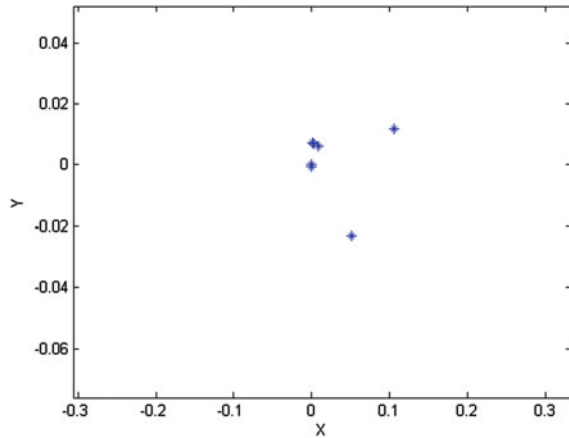
$$bc_position(i) = current_position(i) + next(i) \quad (66.15)$$

where $current_position(i)$ denotes the position of bacteria i at this time; $next(i)$ denotes the expected changing value of the next location: $next(i) = vt \times e^{j\alpha}$, v , t and α are the velocity, the duration and the rotation angle of movement respectively.

Step 5: Compare the objective function fitness values of the two positions $center_position(i)$ and $bc_position(i)$, then bacteria i moves to the position which value is better at number $k+1$ step.

Step 6: Update the optimal position and the related parameters. Repeat step 3–5, until the termination conditions satisfied, jump out of the cycle.

Fig. 66.1 Bacterium location at the 50th iteration



66.4 Numerical Experiments

To test the performance of the NBCC algorithm for solving the function optimization problem, the author developed program for it. In this chapter, four typical example in literature [2] is adopted and the experiment results are compared with result of this chapter.

Concerning the NBCC algorithm, simulation experiment parameters are as follows: the scale of bacterial colony is 20, and the maximum iterative times are 250, the precision $\epsilon = 0.000001$ and the initial position of bacterium is generated by Kent map based on chaotic sequence.

Tests on following function for search minimization:

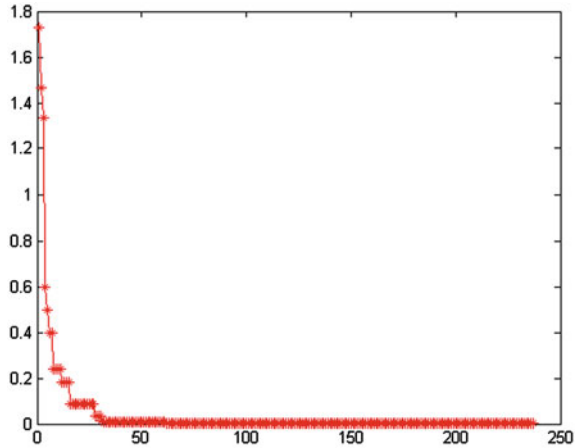
$$F_1(x, y) = (x^2 + y^2)^{0.25} \cdot (\sin^2(50(x^2 + y^2)^{0.1}) + 1.0) \quad (x, y) \in [-20, 20]$$

The function $F_1(x, y)$ reaches the global minimum value 0 at point (0, 0). There are innumerable local minimum points in function interval, the general optimization algorithm can easily fall into those local minimum. Figure 66.1 is each bacterium current location after the 50th iteration. As can be seen, most bacterium current locations have almost reached the global minimum point.

Figure 66.2 gives the global optimal fitness of function $F_1(x, y)$ changing with steps during an iterative process. Generally speaking, in contrast with BCC algorithm the number of iterations is less and the convergence speed is faster. So the convergence of this algorithm is effective.

The function is solved by BCC algorithm and this novel NBCC algorithm. As can be seen, the number of iterations for solving the function optimal problem is less and the convergence speed is faster than BCC. So the convergence speed of the novel NBCC algorithm is faster than BCC.

Fig. 66.2 Global optimal fitness of function



66.5 Conclusions

This paper presents a novel chaotic hybrid bacterial colony chemotaxis (NBCC) algorithm for the function optimization problem, which combines BCC algorithm with a chaos searching strategy and elitist strategy. By integrating elitist strategy and chaotic optimization into bacterial colony chemotaxis optimization algorithm, it greatly enhances the local searching efficiency and global searching performance. Simulation results on standard test functions show that NBCC is pretty efficient to solve complex problems. It has high optimization efficiency, good global performance, and stable optimization outcomes. The performance of NBCC is evidently better than BCC and BC. Therefore, the NBCC algorithm provides the function optimization problem with a novel and efficient solution.

Acknowledgments This work was supported in part by project “Research on Key Problem of Combinatorial Software Testing optimization Based on Swarm Intelligence” (61050003) from National Natural Science Foundation of China, by project “Smart Combinatorial Soft Testing method “ (ZL2009-9) from Natural Science Foundation of XUPT, by project “Smart Combinatorial Embedded Soft Testing Platform” (2009K08-26) from Key Technologies R&D Programmed Foundation of Shan xi Province.

References

1. Li W, Wang H, Zou Z (2005) Function optimization method based on bacterial colony chemotaxis. *J Circuits Syst* 10:58–63
2. Sibylle D, Jarno M, Stefano A et al (2002) Optimization based on bacterial chemotaxis. *IEEE Trans Evol Computation* 6(1):16–29

3. Dahlquist FW, Elwell RA, Lovely PS (1976) Studies of bacterial chemotaxis in defined concentration gradients—a model for chemotaxis toward L-serine. *J Supramol Struct* 4:329(289)–342(302)
4. Li J, Wu W, Xu S-Q (2006) Analysis of three types chaotic biphased codes performance. *J China Acad Electron Inf Technol* 6:527–532
5. Chen WB, Liu YJ, Wang L, Liu XL (2009) A study of the multi-objective evolutionary algorithm based on elitist strategy. In: *Proceedings of 2009 Asia-Pacific conference on information processing*, pp 136–140
6. Sun J, Wang S, (2010) A novel chaos discrete particle swarm optimization algorithm for test suite reduction. In: *Proceedings of 2010 2nd international conference on information science and engineering (ICISE)*, pp 1–4

Chapter 67

Time–Space Hybrid Markov Model

Xin-yao Zou

Abstract Markov model has been used to model user navigational behavior on the World Wide Web. Most existing Markov models based on recommendation algorithms make recommendation just through Web server logs without Web topology. In this chapter, a time–space hybrid Markov model (TSHMM) recommendation algorithm is proposed. First we build both time model from Web log and space model from Web topology respectively. Then we build time–space hybrid model by combining time model with space model. Hybrid model can make recommendation according to user preferences. Some experiments are conducted to validate the hybrid model. The experimental results show that hybrid model can be successfully applied to Web recommendation.

Keywords Markov model · Time–space hybrid model · Recommendation

67.1 Introduction

Web mining has received a great deal of attention because it can extract useful patterns from Web data. Web usage mining [1–3], an important part of Web mining, uses Web logs to find out user navigation patterns. The usage pattern extracted from Web logs can be applied to a wide range of applications, such as Web personalization, site modification, system improvement, business intelligence discovery and usage characterization etc. At an online shop, customers may have

X. Zou (✉)

Department of Electron and Information Engineering, Guangdong AIB
Polytechnic College, Guangzhou 510507, China
e-mail: madelinexy@163.com

difficulty to choose what they need among the great amount of products. On the other hand, it is not easy for retailers to survive due to more and more business competition. If retailers know more about the demand of their customers, online retailers can offer personalized products and services to users according to their preferences.

Recommendation algorithms include collaborative filtering (CF) [4] and Web mining. Most existing Markov-based recommendation algorithms recommend only based on Web logs and neglect that it is Web topology where users navigates on. In this chapter, we build time model from Web logs and space model from Web topology respectively. By combining time model with space model, a time-space hybrid Markov model (TSHMM) recommendation algorithm is proposed.

Section 67.2 explains the basic concepts of Web recommendation and Markov model. Section 67.3 details the TSHMM algorithm. Experimental results are given in Sect. 67.4. Conclusion remarks are drawn in Sect. 67.5.

67.2 Problem Description

Web site is a resource set R , including Web pages, documents, images and sounds. Web server logs store data about user behavior. The first step is preprocessing for Web log based recommendation algorithms, which includes data cleaning, user identification and session identification. After preprocessing, Web log is transformed to a transactional table. Each row of the table is a resource access sequence. A Web log is as follows.

$$\text{Log} = \{L_i\} = \{(L_{i1}, L_{i2}, \dots, L_{iL}) | L_{ij} \in R, j \in [1, L]\} \quad (67.1)$$

Markov model based recommendation algorithms build Markov model from user historical behaviors. Each node in the link graph can be regarded as a state in a finite discrete Markov model. The model can be defined by a tuple $\langle R, A, B \rangle$, where R is the state space containing all the nodes in the link graph, A is the probability transition matrix containing one-hop transition probabilities between the nodes, and B is the initial probability distribution on the states in R . The first order Markov model only considers page-to-page transition probabilities:

$$p(x_2|x_1) = \Pr(X_2 = x_2 | X_1 = x_1) \quad (67.2)$$

The K -order Markov model considers conditional probability that a user transitions to the n th page given his or her previous k page visits:

$$p(x_n|x_{n-1}, \dots, x_{n-k}) = \Pr(X_n = x_n | X_{n-1} = x_{n-1}, \dots, X_{n-k} = x_{n-k}) \quad (3)$$

A Markov model depends on the initial distribution vector B and the transition matrix A . $A = [a_{ij}] = P(X[t] = R^j | X[t-1] = R^i)$, $B = [b_i] = P(X[0] = R^i)$. The user's navigation in the Web site can be seen as a stochastic process $\{X_n\}$, which

has R as the state space. The transition matrix is used to predict the probability of pages accessed next by users.

67.3 TSHMM Recommendation Algorithm

Former Web log based recommendation algorithms use Web logs to build the Markov model, which is regarded as our time Markov model. The time model discovers patterns of user behavior. Combining with Web topology, we further build space Markov model to improve recommendation results. The space model exhibits associations in Web topology. Accordingly the TSHMM recommendation algorithm is proposed.

67.3.1 Time model

By viewing a Web user's navigation in a Web site as a Markov chain, we can build a Markov model for link prediction based on past users' visit behavior recorded in the Web log file. Time Markov model can be built from user sessions. Here we use wireless application protocol (WAP) to get the frequent 1-sequence sets S_1 , 2-sequence sets S_2 and their supports. An n -sequence means a sequence at the length of n . Let $S_1 = \{s_i\}$, $S_2 = \{(s_{i1}, s_{i2})\}$, the transition matrix $A = [a_{ij}] = \left[\frac{\sup(a_{i s_1}, a_{i s_2})}{\sup(a_{i s_1})} \right]$, a_i^* , $a_j^* \in S_1$, $(a_i^*, a_j^*) \in S_2$, where a_i^* is the i th row item of A , $\sup(a_i^*, a_j^*)$ is the support of the 2-sequence (a_i^*, a_j^*) , $\sup(a_i^*)$ is the support of a_i^* . The above items can be deduced from S_1 and S_2 .

The printing area is 122 mm \times 193 mm. The text should be justified to occupy the full line width, so that the right margin is not ragged, with words hyphenated as appropriate. Please fill pages so that the length of the text is no less than 180 mm, if possible.

Use 10-point type for the name(s) of the author(s) and 9-point type for the address(es) and the abstract. For the main text, please use 10-point type and single-line spacing. We recommend the use of Computer Modern Roman or Times. Italic type may be used to emphasize words in running text. Bold type and underlining should be avoided.

Papers not complying with the LNCS style will be reformatted. This can lead to an increase in the overall number of pages. We would therefore urge you not to squash your paper.

67.3.2 Space model

One-hop and multi-hop model

The space model includes one-hop model and multi-hop model. One-hop Markov model is a one-step transition model. It can find out direct associations among Web objects in the Web topology. Multi-hop Markov model is a multi-step transition model. It can identify indirect associations among Web objects in the Web topology.

One-hop model

Web topology can be regarded as a space Markov model. Its transition matrix is $M = [m_{ij}]$, where m_{ij} can be calculated with the equal probability model. $m_{ij} = \frac{1}{L_i+1}$, $L_i \geq 0$, where L_i is the count of links in page m_i^* . When a user is at page m_i^* , he (or she) can clicks on any of the L_i links in the page or leave the Web site with equal probability.

Multi-hop model

One-hop model can only obtain direct associations among pages. But it is necessary to calculate the indirect associations, which describe the real associations among pages in Web topology. The transition matrix is $B_1 = \sum_{t=0}^{\infty} B_t$, where $(B_t)_{ij}$ is the association probability from $(B_t)_i^*$ to $(B_t)_j^*$ passing t nodes. B_{ij} is the association probability from B_i^* to B_j^* passing any count of nodes. It has been proved that

$B_1 = \sum_{t=0}^{\infty} B_t = \sum_{t=0}^{\infty} M^{t+1} = E - 1(E - M) - E$. From $(E - M) B_1 = M$, so an iteration method can be used to calculate matrix B_1 from M . The iteration formula is $B_1(k+1) = MB_1(k) + M$, where $\{B_1(k)\}$ is convergent. A detailed explanation has been given in [5]. The converged matrix B_1 and A need make equal dimension. B_1 also needs to be normalized, as the one-hop model. Then Delete the k th row and column from B_1 , where $k \in \{k | (B_1)_{k*} \notin S_1\}$. It can be proved that the +dimension of A and B_1 is equal. Next normalize B_1 to B , where $\forall i, a_i^* = b_i^*$. The normalization ensures that the same row item of A and B is the same. In the end, matrix B can be obtained.

67.3.3 TSHMM recommendation algorithm

Time Markov model and space Markov model can be combined to a hybrid model, which can reveal the user interest more accurately. The factual user behavior depends on two factors: user interest and Web topology. In the former part, we have obtained the transition matrix A of the time model and the transition matrix B of the multi-hop space model. Time model reflects the factual access behavior depending on the Web topology. Hybrid model can reflect the user's real interest, not depending on the Web topology. Define the matrix $C = A - B$, which reflects interest shift of users and can be used for recommendation. Then set two thresholds, including a positive one $t > 0$ and a negative one $t' < 0$. Accordingly we can get several heuristic rules from C .

- a) If $c_{ij} > t > 0$, the probability from c_{i^*} to c_{j^*} of user access is much larger than that of the Web topology. Users have shown strong desire to navigate c_{j^*} after visited c_{i^*} , although it has never been considered by web site designers. From this rule, we can recommend c_{j^*} when a user is at the page c_{i^*} . The recommendation value is $c_{ij} - t$.
- b) If $c_{ij} < t' < 0$, the probability from c_{i^*} to c_{j^*} of user access is much smaller than that of the Web topology. From this rule, Web site designers can remove the link(or the route) from c_{i^*} to c_{j^*} since it is seldom accessed. The removing value is $t' - c_{ij}$.
- c) If $t' < c_{ij} < t$, the probability from c_{i^*} to c_{j^*} of user access and that of the Web topology is approximately the same. No action is needed.

From these heuristic rules, retailer can not only recommend something valuable to users, but also offer revising comments to Web site designers. The recommendation matrix is $D = [d_{ij}]$, where $d_{ij} = \begin{cases} c_{ij} - t, & c_{ij} \geq t \\ 0, & c_{ij} < t \end{cases}$. The recommendation matrix is generated offline. The online module recommends Web pages to users, with the recommendation value, according to the user's current access.

The whole TSHMM recommendation algorithm is as follows:

Input: Threshold t ($t > 0$), Web log Log , Web topology $Topo$

Output: Recommendation matrix D

Begin

Calculate the transition matrix of the time model A . (Sect. 67.3.1)

Calculate the transition matrix of the one-hop space model M . (Sect 67.3.2)

Calculate the transition matrix of the multi-hop space model B . (Sect 67.3.3)

$C = [c_{ij}] = A - B$; //calculate the transition matrix of the hybrid model

$$D = [d_{ij}], \text{ where } d_{ij} = \begin{cases} c_{ij} - t & c_{ij} \geq t \\ 0 & c_{ij} < t \end{cases} \text{ //recommendation matrix}$$

return D ;

End

Table 67.1 Web Log

TID	Navigation commodity ID							
1	A	B	C	D	E	F	G	H
2	B	I	C	A	J	F		
3	I	A	K	L	M			
4	I	C	N	O	H			
5	B	A	C	P	J	H	G	Q

67.3.4 Discussion

TSHMM regards A as the user's factual behavior. We can also recommend to different user groups and extend the model to the clustering model. Users are clustered into groups A1, A2, ..., AG from Web logs, where G is the number of clusters. A lot of clustering methods can be applied here, such as a mixture hidden Markov Model method stated in [6]. Supposing the Web topology is the same to all users, we can obtain interest conversion patterns of users in each cluster. The transition matrix in cluster is i , $C_i = A_i - B$.

In a dynamic Web site, Web topology is probably not available, but some other information is similar to Web topology, such as product catalog, can be used to get user interest shift patterns. The model can be easily applied to other domains, such as E-commerce.

67.4 Experiments

All experiments are conducted on a 1.8-GHz Pentium PC machine with 512 megabytes main memory, running Microsoft Windows/2000 Server. All the programs are written in Microsoft/Visual C#. Experiment 1 illustrates the TSHMM recommendation algorithm and its results. For explanation, we use artificial synthetic data. The Web log is the same as Table 67.1. The string values in the table are the web page IDs. The Web topology is shown as Fig. 67.1.

Suppose the recommendation threshold is 0.2. After calculation, we get the time model and the space model, as shown in Fig. 67.2. The column items of the matrix are C, A, B, I, H from left to right.

From matrix D , we can get the recommendation rules as shown in Table 67.2, with the recommendation values sorted descending. From Table 67.2, it can be seen that the recommendation value of the rule containing H is high. This is because H is not directly connected to any other pages in the matrix so that association values with other pages are all very small. Let us look at the rule $I \Rightarrow A$. The value in the time model is large, but the recommendation value is small. This is because I and A are directly connected and the corresponding value in the space model is also large. Both user factual behavior and Web topology are

Fig. 67.1 Web topology graph

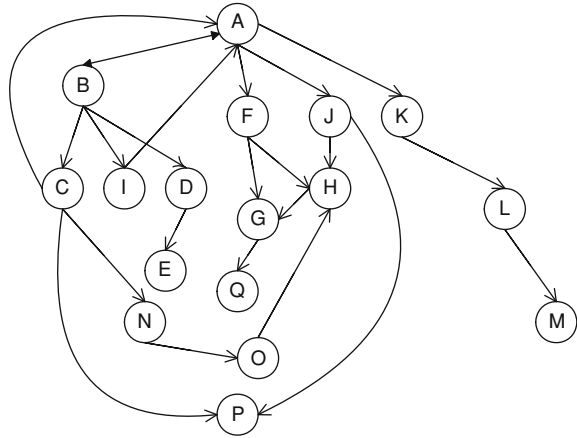


Fig. 67.2 Recommendation matrices

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0.75 \\ 0.5 & 0 & 0.25 & 0 & 0.5 \\ 1 & 0.67 & 0 & 0.33 & 0.67 \\ 0.67 & 0.67 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$B_1 = \begin{bmatrix} 0 & 0.25 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0 & 0 \\ 0.2 & 0.2 & 0 & 0.2 & 0 \\ 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} 0.01 & 0.27 & 0.06 & 0.01 & 0 \\ 0.04 & 0.08 & 0.22 & 0.04 & 0.02 \\ 0.22 & 0.38 & 0.08 & 0.22 & 0.1 \\ 0.02 & 0.54 & 0.11 & 0.02 & 0.01 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$C = \begin{bmatrix} -0.01 & -0.27 & -0.06 & -0.01 & 0.75 \\ 0.46 & -0.08 & 0.03 & -0.04 & 0.48 \\ 0.78 & 0.29 & -0.08 & 0.11 & 0.57 \\ 0.65 & 0.13 & -0.11 & -0.02 & -0.01 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$D = \begin{bmatrix} 0 & 0 & 0 & 0 & 0.75 \\ 0.46 & 0 & 0 & 0 & 0.48 \\ 0.78 & 0.29 & 0 & 0 & 0.57 \\ 0.65 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

important factors when making recommendations. TSHMM synthesize these two factors and propose a new formula for recommendation value. The more different user factual behavior and Web topology is, the higher recommendation value the rule has.

Table 67.2 Recommendation rules

Recommendation rule	Recommendation value
$B \Rightarrow C$	0.78
$C \Rightarrow H$	0.75
$I \Rightarrow C$	0.65
$B \Rightarrow H$	0.57
$A \Rightarrow H$	0.48
$A \Rightarrow C$	0.46
$B \Rightarrow A$	0.29

67.5 Conclusion

Personalization and recommendation can help user to find out interesting contents from vast amounts of data. It has been widely used in many fields, such as E-Commerce Web sites, digital library and so on. Markov model based Web mining methods can be well applied in personalization. In this chapter, we propose the time–space hybrid Markov model and recommendation algorithm. It uses Web log and Web topology to describe interest conversion of users. TSHMM can also extend to clustering model and recommend based on clustering. The experiments illustrate the TSHMM algorithm, which can be well applied in Web mining and Web recommendation algorithms. What’s more, it can be easily applied in other domains, such as E-Commerce.

References

- Heydari M, Helal RA, Ghauth KI (2009) A graph-based web usage mining method considering client side data. In: Proceedings IEEE Symposium Electrical Engineering and Informatics (ICEEI 09), IEEE Press, Aug. 2009 pp 147–153,doi: [10.1109/ICEEI.2009.5254802](https://doi.org/10.1109/ICEEI.2009.5254802)
- Han QT, Gao XY, Wu WG (2008) Study on web mining algorithm based on usage mining. In: Proceedings IEEE symposium the 9th international conference on computer-aided industrial design conceptual design (CAIDCD 08), IEEE press, Nov 2008 pp 1121-1124,doi:[10.1109/CAIDCD.2008.473075-9](https://doi.org/10.1109/CAIDCD.2008.473075-9)
- Salin S, Senkul P (2009) Using semantic information for web usage mining based recommendation. In: Proceedings IEEE symposium the 24th international symposium on computer information sciences (ISCIS 09), IEEE press, Sept 2009 pp 236–241, doi: [10.1109/ISCIS.2009.5291819](https://doi.org/10.1109/ISCIS.2009.5291819)
- Breese J, Heckerman D, Kadie C (1998) Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the 14th conference on uncertainty in artificial intelligence. Morgan Kaufmann Publisher, Madison, WI, May 1998 pp 43–52
- Li YJ, Peng H, Zheng QL, Yang P (2003) A web topology probability matrix approach for interesting association rules discovery. In: Proceedings of the 7th world multi-conference on systemics, cybernetics and Informatics (SCI2003), July 2003 pp 150–155
- Ypma A, Heskes T (2002) Categorization of web pages and user clustering with mixtures of hidden markov models. In: Workshop notes of fourth WEBKDD web mining for usage patterns and user profiles at KDD-2002, pp 31–43

Chapter 68

Research on Turning Parameters Optimization Based on Genetic Algorithm

Youde Wu and Jinchun Feng

Abstract In the process of practical production, especially for machining of large-scale parts, how to choose cutting parameters to improve production efficiency was a difficult problem. On the MATLAB platform, the genetic algorithm was adopted to deal with the optimization of multi-variables objective function. The optimum values were gotten, the difficult problem was solved. The research' findings are of important significance to the confirmation of three parameters of cutting machining.

Keywords Turning parameters · Matlab · Genetic algorithm · Optimization

68.1 Introduction

Metal-cutting includes very complex procedure [1]. The optimization of metal-cutting processing parameters is of important significance to ensure quality, increase productivity and reduce production costs. Nowadays, many factories determine the amount of cutting according to their own experience. However, this option often has randomness, it will vary from person to person [2–4]. Therefore, the most effective way to avoid randomness is to adopt high-efficient and reasonable algorithm to deal with the optimization of turning parameters. The genetic algorithm is one of the intelligent calculation methods, widely used in

Y. Wu (✉)
Southwest Jiaotong University, Cheng Du 610031, China
e-mail: wyd@scetc.net

Y. Wu · J. Feng
Sichuan Engineering Technical College, Deyang 618000, China
e-mail: catsmiling@163.com

recent years. It is a kind of random search algorithm learnt from natural selection and genetic mechanisms. Compared to traditional optimization, it has group search features, parallelism, scalability, and it will not fall into partial optimization. This thesis is based on the MATLAB platform, applying genetic algorithms to solve the problems of turning parameters optimization. Description about the turning parameters optimization.

68.2 The Objective Functions and Design Variables

No matter choosing or optimizing turning parameters, we must take the economy in the cutting machining process into first consideration and determine clear goal for the optimization. Generally speaking, we regard the largest productivity, minimum cost or maximum profit as the goal. In the process of metal cutting, the metal removal capacity in a unit time (Z_w) can be indicated by the following formula:

$$Z_w = 1000.v.f.a_p \quad (68.1)$$

From formula (68.1), we can know that if the cutting parameter or the metal removal capacity in a unit time is increased, we can raise the productivity of cutting machining. It shows that, on the basis of ensuring quality and reducing cost, the reasonable choice and optimization of cutting parameters will keep the product of the three independent design variables, including v (cutting velocity), f (feed) and a_p (the depth of cut) on increasing. But due to the production equipment and production technology constraints, it is not allowed to raise cutting parameters randomly. A real cutting process is limited by many constraints. The central issue of this paper is searching how to seek optimized result under so many constraints [5].

68.3 Constraints

The constraints during cutting process mainly come from two aspects. One is the capability of technics system (such as machine tools, cutting tools, workpiece, etc.) The other is the technical requirement of workpiece (such as machining precision, surface quality, etc.), though very many constraints are emerged technically. According to the property and type of machining (such as roughing or finishing cut, intermittent or continuous processing, etc.), the main constraints are also likely to be different [6]. Their relationship can be described by Fig. 68.1

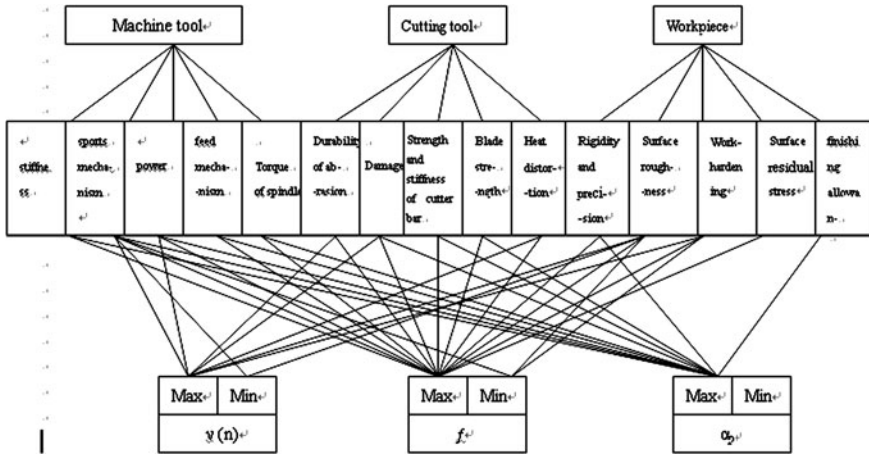


Fig. 68.1 Network chart of constraints

Aspect of cutting tools Constraint inequality of cutting tool durability

$$n \cdot f^{Y_v} \leq \frac{318.31 C_v \cdot K_v}{60^{(1-m)} \cdot d_w \cdot T^m \cdot a_p^{X_v}} \tag{68.2}$$

Constraints inequality of cutter bar strength

$$n^{Z_{Fz}} \cdot f^{Y_{Fz}} \leq \frac{(1000)^{Z_{Fz}} \cdot \sigma_{bb} \cdot B \cdot H^2}{6l \times 9.81 \times 60^{Z_{Fz}} \cdot C_{Fz} \cdot a_p^{X_{Fz}} \cdot (\pi d_w)^{Z_{Fz}} \cdot K_{Fz}} \tag{68.3}$$

Aspect of machine tools Constraints inequality of effective power of the machine tool

$$n^{(Z_{Fz}+1)} \cdot f^{Y_{Fz}} \leq \frac{101.94 \times (1000)^{(Z_{Fz}+1)} \cdot P_E \cdot \eta_m}{60^{Z_{Fz}} \cdot (\pi d_w)^{(Z_{Fz}+1)} \cdot C_{Fz} \cdot a_p^{X_{Fz}} \cdot K_{Fz}} \tag{68.4}$$

Inequality of effective spindle torque constraint

$$n^{Z_{Fz}} \cdot f^{Y_{Fz}} \leq \frac{203.87 \times (1000)^{Z_{Fz}} \cdot M_S}{60^{Z_{Fz}} \cdot C_{Fz} \cdot a_p^{X_{Fz}} \cdot K_{X_{Fz}} \cdot \pi^{Z_{Fz}} \cdot d_w^{(Z_{Fz}+1)}} \tag{68.5}$$

Strength of machine tool's inner body intensity constraint inequality

$$n^{Z_{Fz}} \cdot f^{Y_{Fz}} \leq \frac{(1000)^{Z_{Fz}} \times 2 \times F_j}{9081 \times 60^{Z_{Fz}} \cdot C_{Fz} \cdot a_p^{X_{Fz}} \cdot K_{Fz} \cdot (\pi d_w)^{Z_{Fz}}} \quad (68.6)$$

Machine tool's maximum and minimum axis rotation constraint inequality

$$n_{\min} \leq n \leq n_{\max} \quad (68.7)$$

Machine tool's maximum and minimum feed constrain inequality

$$f_{\min} \leq f \leq f_{\max} \quad (68.8)$$

Aspect of workpiece Constraint inequality of workpiece rigidity.

$$n^{Z_{Fy}} \cdot f^{Y_{Fy}} \leq \frac{(1000)^{Z_{Fy}} \cdot K_O \cdot E \cdot J \cdot f'}{9.81 \times 60^{Z_{Fy}} \cdot C_{Fy} \cdot a_p^{X_{Fy}} \cdot K_{Fy} \cdot (\pi d_w)^{Z_{Fy}} \cdot I^3} \quad (68.9)$$

Constraint inequality of surface roughness

$$f \leq K_R \cdot R_Z^u \quad (68.10)$$

Constraint inequality of maximum and minimum cutting profundity in cutting machining.

$$a_{p \min} \leq a_p \leq a_{p \max} \quad (68.11)$$

68.4 Turning Parameters Optimization

GA is a kind of self-adaptive searching arithmetic to optimize the overall situation. It is formed in the simulation of the natural environment in the biological and genetic evolution. It has the characteristics of simple, parallel, integrated, robustness and it will not fall into partial a optimization. In the mean time, it employs the change rules of probability to guide the search. So the search range is wide and the efficiency is high

Confirmation of GA basic elements

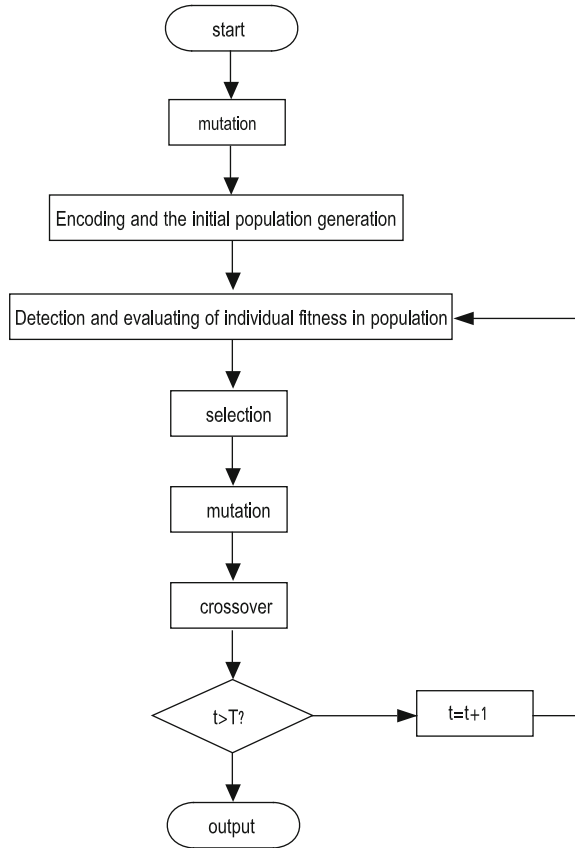
Fitness Functions:

$$Z_w = 1000 \cdot v \cdot f \cdot a_p \quad (68.12)$$

In the formula, v means cutting speed, f means feed rate, a_p means cutting depth.

Other factors. Population size (M):50;cross-probability (P_c):0.8; mutation probability (P_m):0.01; Evolution termination algebra(G):100.

Fig. 68.2 calculation flow chart



Basic operation of GA

There are three basic steps in the process of genetic algorithm, they are selection, crossover and mutation. That is: (1) selection. (2) crossover. (3) mutation.

Main calculation process and flow chart of GA

GA's main calculation process and the flow chart are explained in Fig. 68.2:

68.4.1 Optimization Result of GA Operation

By compiling the * files of fitness functions, penalty functions and constraint conditions, the result calculated by GA can be showed by Figs. 68.3 and 68.4. Figure. 68.3 presents the best fitness and average fitness value of each generation. Figure. 68.4 presents the best individual corresponding to the best fitness of the

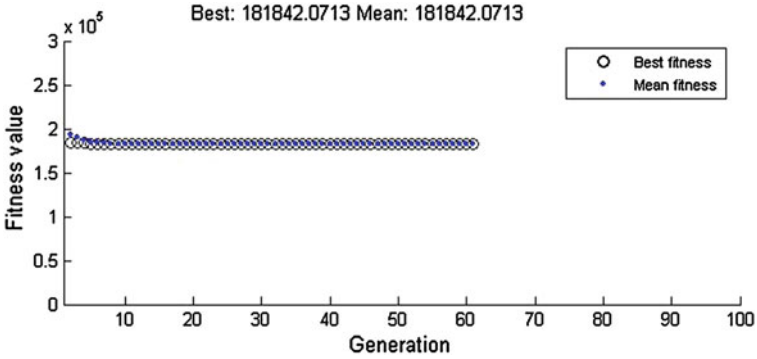


Fig. 68.3 fitness functions

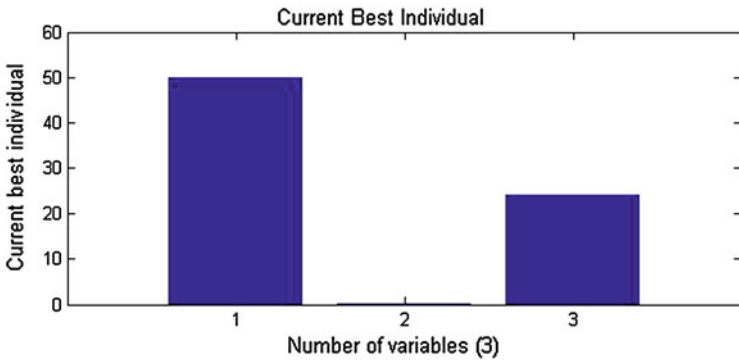


Fig. 68.4 Best individual

current generation. The final optimization result is: the optimal value $z_w^* = 1.8 \times 10^5$; design variables: $v^*, f^*, a_p^* = (50.0276 \ 0.1508 \ 24.0980)$; the calculate result of iteration number is 61.

68.5 Conclusion

The paper aims at large spare component processing and searching the way of choosing the optimal turning parameters of turnings. Many traditional optimization algorithms can achieve the optimization of cutting parameters too. However, they are single-point search and are apt to fall into partial extreme value and low-efficient. On the contrary, GA can deal with several individual during the same time. It evaluates multiple solutions in the searching space, so globally optimal

solution can be obtained in a greater probability. It also possesses higher calculate parallelism [7, 8]. So, making most use of such characteristics as group searching, high speed and high reliability, we can get the function optimum value and solve the problem of multi-objective optimization.

References

1. Pan M-Q, Liu Y-J, Tang Y (2005) The multi-object optimization model of layered turning cut. *Tool Technol* 39(8):29–33
2. Chen Z-A (1995) Cutting parameters optimization of machining. *Mach Manuf Eng* 1:27–28
3. Zhao Y-L, Yang Q, He D-J (2003) Multi object optimization of machining parameters for turning. *Mech Des Manuf* 4:90–91
4. Bin J, Zheng M-L, Li Z-J (2001) Multi-objective optimization of turning parameters. *Harbin Univ Technol*, 6 (2) : 43–46
5. Ermer DS (1997) A century of optimizing machining operations, *ASME. J Manuf Sci Eng* 119:817–822
6. Yang SH (2010) Multi-objective optimization of cutting parameters in turning process using differential evolution and non-dominated sorting genetic algorithm-II approaches. *Int J Adv Manuf Technol* 49(5–8):773–784
7. Solimanpur M, Ranjdoostfard F (2009) Optimisation of cutting parameters using a multi-objective genetic algorithm. *Int J Prod Res* 47(21):6019–6036
8. Savas V, Ozay C (2008) The optimization of the surface roughness in the process of tangential turn-milling using genetic algorithm: *International Journal of Advanced Manufacturing Technology* n 37:335–340

Chapter 69

A Power of a Meromorphic Function Sharing 1 IM with Its Derivative

Wang Gang and Gong Dianxuan

Abstract In this chapter, we deal with the problem of uniqueness of meromorphic functions sharing 1 IM with their derivative. The main result is that give two nonconstant meromorphic functions f and g with an integer $n \geq 5m + 37$, denote $F = f^n(f^m - 1)f'$ and $G = g^n(g^m - 1)g'$, if F and G sharing 1 IM, then $f = g$ or $f = -g$. This improve the result of junfeng Xu and hongxun Yi.

Keywords Meromorphic functions · Entire functions · Derivative · Uniqueness theorem

69.1 Introduction

Let f and g be two non-constant meromorphic functions in the complex plane. It is assumed that the reader is familiar with the standard notations of Nevanlinna's theory such as $T(r, f)$, $m(r, f)$, $N(r, f)$, $\bar{N}(r, f)$, and so on, which can be found in [1]. We use E to denote any set of positive real numbers of finite linear measure, not necessarily the same at each occurrence. The notation $S(r, f)$ denotes any quantity satisfying $S(r, f) = o(T(r, f))(r \rightarrow \infty, r \notin E)$. Let a be a complex number, we say that f and g share the value a CM provided $f - a$ and $g - a$ have

W. Gang (✉)

Shandong Transport Vocational College, Weifang 261206,
Shandong, China
e-mail: jyjcbwg@163.com

G. Dianxuan

College of Sciences, Hebei United University, Tangshan 063009,
Hebei, China
e-mail: gdx216@yahoo.cn

the same zeros counting multiplicities (see [2]). We say that f and g share ∞ CM provided that $1/f$ and $1/g$ share 0 CM. In addition, we also use the following notations. Let k be a positive integer and $a \in C \cup \infty$. We denote by $N_k\left(r, \frac{1}{f-a}\right)$ the counting function of a points of f with multiplicity $\leq k$, by $N_{(k)}\left(r, \frac{1}{f-a}\right)$ the counting function of a -points of f with multiplicity $\geq k$; and denote the reduced counting function by $\bar{N}_k\left(r, \frac{1}{f-a}\right)$, $\bar{N}_{(k)}\left(r, \frac{1}{f-a}\right)$, respectively. Set $N_k\left(r, \frac{1}{f-a}\right) = \bar{N}\left(r, \frac{1}{f-a}\right) + \bar{N}_{(1)}\left(r, \frac{1}{f-a}\right) + \dots + \bar{N}_{(k)}\left(r, \frac{1}{f-a}\right)$.

In 1976, Gross proposed the following question.

Question 1. Whether there exists a finite set S such that $E(S, f) = E(S, g)$ can imply $f \equiv g$ for any pair of nonconstant entire functions f and g ?

Yi [3] gave a positive answer to Question 1. He proved there exist a polynomial set S with seven elements such that $E(S, f) = E(S, g)$ can imply $f \equiv g$ for any pair of nonconstant entire functions f and g .

In 2002, Fang and Fang [4] proved there exists a differential polynomial d such that for any pair of nonconstant entire functions $f \equiv g$ if $d(f)$ and $d(g)$ share one value CM.

Theorem A (See [4]) *Let f and g be two nonconstant entire functions, $n \geq 8$ be an integer. Denote $F = f^n(f - 1)f'$ and $G = g^n(g - 1)g'$. If F and G sharing 1 CM, then $f = g$.*

In 2004, Lin-Yi [5] and Qiu-Fang [6] proved that Theorem A remains valid for $n \geq 7$.

Theorem B *Let f and g be two nonconstant entire functions, $n \geq 7$ be an integer. Denote $F = f^n(f - 1)f'$ and $G = g^n(g - 1)g'$. If F and G sharing 1 CM, then $f = g$.*

Later Fang–Fang obtained the following theorem, they replaced CM with IM.

Theorem C (See [5, Theorem 1]) *Let f and g be two nonconstant entire functions, $n \geq 18$ be an integer. Denote $F = f^n(f - 1)f'$ and $G = g^n(g - 1)g'$. If F and G sharing 1 IM, then $f = g$.*

In 2007, Xu and Yi [7] get following results.

Theorem D (See [7]) *Let f and g be two nonconstant entire functions, $n \geq 4m + 11$ be an integer. Denote $F = f^n(f^m - 1)f'$ and $G = g^n(g^m - 1)g'$. If F and G sharing 1 IM, then $f = g$.*

Theorem E (See [7]) *Let f and g be two nonconstant entire functions, $n \geq 4m + 11$ be an integer. Denote $F = f^n(f^m - 1)f'$ and $G = g^n(g^m - 1)g'$. If F and G sharing z IM, then $f = g$.*

But the results of Theorem D and Theorem E are not true. When m is odd, n is even, let $f = -g$, results are not true. In this chapter, we improve Theorem D and E by obtaining the following results:

Theorem 1.1 *Let f and g be two nonconstant meromorphic functions, $n \geq 5m + 37$ be an integer. Denote $F = f^n(f^m - 1)f'$ and $G = g^n(g^m - 1)g'$. If F and G sharing 1 IM, then $f = g$ or $f = -g$.*

69.2 Some Lemmas

Let f and G be two non-constant meromorphic functions such that F and G share the value 1 IM. Let z_0 be a 1-point of F of order p , a 1-point of G of order q . We denote by $N_L(r, \frac{1}{F-1})$ the counting function of those 1-points of F where $p > q$; by $N_E^1(r, \frac{1}{F-1})$ the counting function of those 1-points of F where $p = q = 1$; by $N_E^2(r, \frac{1}{F-1})$ the counting function of those 1-points of F where $p = q \geq 2$; each point in these counting functions is counted only once. In the same way, we can define $N_L(r, \frac{1}{G-1})$, $N_E^1(r, \frac{1}{G-1})$, and $N_E^2(r, \frac{1}{G-1})$. Particularly, if F and G share 1 CM, then

$$N_L\left(r, \frac{1}{F-1}\right) = N_L\left(r, \frac{1}{G-1}\right) = 0.$$

With these notations, if F and G share 1 IM, it is easy to see that $\bar{N}\left(r, \frac{1}{F-1}\right) = N_E^1\left(r, \frac{1}{F-1}\right) + N_L\left(r, \frac{1}{F-1}\right) + N_L\left(r, \frac{1}{G-1}\right) + N_E^2\left(r, \frac{1}{G-1}\right) = \bar{N}\left(r, \frac{1}{G-1}\right)$

Lemma 2.1 (See [2]) *Let $H = \left(\frac{F''}{F'} - \frac{2F'}{F-1}\right) - \left(\frac{G''}{G'} - \frac{2G'}{G-1}\right)$. where F and G are two nonconstant meromorphic functions. If $H \neq 0$, then*

$$N_E^1\left(r, \frac{1}{G-1}\right) \leq N(r, H) + S(r, F) + S(r, G). \tag{69.1}$$

Let p be a positive integer and $a \in C \cup \{\infty\}$. We denote by $N_p\left(r, \frac{1}{f-a}\right)$ the counting function of the zeros of $f - a$ with the multiplicities less than or equal to p , and by $N_{(p+1)}\left(r, \frac{1}{f-a}\right)$ the counting function of the zeros of $f - a$ with the multiplicities larger than p . And we use $\bar{N}_p\left(r, \frac{1}{f-a}\right)$ and $\bar{N}_{p+1}\left(r, \frac{1}{f-a}\right)$ to denote the corresponding reduced counting functions.

Lemma 2.2 *Suppose that f is a nonconstant meromorphic function and k, p are positive integers. Then*

$$N_p\left(r, \frac{1}{f^{(k)}}\right) \leq T\left(r, f^{(k)}\right) - T(r, f) + N_{p+k}\left(r, \frac{1}{f}\right) + S(r, f), \tag{69.2}$$

$$N_p\left(r, \frac{1}{f^{(k)}}\right) \leq k\bar{N}(r, f) + N_{p+k}\left(r, \frac{1}{f}\right) + S(r, f), \tag{69.3}$$

Lemma 2.3 (See [2]) *Let f be a nonconstant meromorphic function, n be positive integers. $P(f) = a_n f^n + a_{n-1} f^{n-1} + \dots + a_1 f$ where a_i is a meromorphic function satisfying $T(r, a_i = S(r, f))(i = 1, 2, 3, \dots, n)$. then $T(r, P(f)) = nT(r, f) + S(r, f)$.*

Lemma 2.4 *Let f and g be two nonconstant meromorphic functions, $n \geq 5m + 37$ be an integer. Denote $F = f^n(f^m - 1)f'$ and $G = g^n(g^m - 1)g'$. If F and G sharing 1 IM, then $S(r, f) = S(r, g)$.*

This can be easily proved by lemma 2.3.

Lemma 2.5 [9] *If $H \equiv 0$, and*

$$\limsup \frac{\bar{N}\left(r, \frac{1}{f}\right) + \bar{N}\left(r, \frac{1}{g}\right) + \bar{N}(r, f)\bar{N}(r, g)}{T(r)} < 1.$$

then $f \equiv g$ or $fg \equiv 1$, where E is a set of finite linear measure and not necessarily the same at each of its occurrence.

Lemma 2.6 *Let f and g be two non-constant meromorphic functions. Let $n \geq 5m + 37$. Denote $F = f^n(f^m - 1)f'$ and $G = g^n(g^m - 1)g'$. Then $F \neq G$.*

Proof We suppose that

$$f^n(f^m - 1)f'g^n(g^m - 1)g' \equiv 1 \tag{69.4}$$

Let z_0 be a zero of f with multiplicity p . Then z_0 is a pole of g with multiplicity q , say. From (69.1) we get $np + p - 1 = nq + mq + q + 1$ and so

$$kq + 2 = (n + 1)(p - q). \tag{69.5}$$

From (69.5) we get $q \geq \frac{n-1}{m}$ and again from (69.5) we obtain $p \geq \frac{n+k-1}{k}$.

Let z_1 be a zero of $f^m - 1$ with multiplicity p . Then z_1 is a pole of g with multiplicity q , say. So from (69.4) we get $2p - 1 = (m + n + 1)q + 1 \geq n + k + 2$. Hence $p \geq \frac{n+k+3}{2}$. Since a pole of f is either a zero of $g^n(g^m - 1)$ or a zero of g' , we have

$$\begin{aligned} \bar{N} \leq & \bar{N}\left(r, \frac{1}{g}\right) + \bar{N}\left(r, \frac{1}{g^m}\right) + \bar{N}_0\left(r, \frac{1}{g'}\right) + S(r, f) + S(r, g) \leq \frac{k}{n+k-1}N\left(r, \frac{1}{g}\right) \\ & + \frac{2}{n+k+3}N\left(r, \frac{1}{g^m}\right) + \bar{N}_0\left(r, \frac{1}{g'}\right) + S(r, f) \\ & + S(r, g)\left(\frac{k}{n+k-1} + \frac{2k}{n+k+3}\right)T(r, g) + \bar{N}_0\left(r, \frac{1}{g'}\right) + S(r, f) + S(r, g) \end{aligned} \tag{69.6}$$

where $\bar{N}_0\left(r, \frac{1}{g'}\right)$ denotes the reduced counting function of those zeros of g' which are not the zeros of $g(g^m - 1)$. Then by the second fundamental theorem we get

$$mT(r, f) \leq \left(\frac{k}{n+k-1} + \frac{2k}{n+k+3}\right)(T(r, g) + T(r, f)) + \bar{N}_0\left(r, \frac{1}{g'}\right) - \bar{N}_0\left(r, \frac{1}{f'}\right) + S(r, f) + S(r, g) \tag{69.7}$$

$$mT(r, g) \leq \left(\frac{k}{n+k-1} + \frac{2k}{n+k+3}\right)(T(r, g) + T(r, f)) + \bar{N}_0\left(r, \frac{1}{g'}\right) - \bar{N}_0\left(r, \frac{1}{f'}\right) + S(r, f) + S(r, g) \tag{69.8}$$

So $\left(1 - \frac{2}{n+k-1} - \frac{4}{n+k+3}\right)T(r, f) + T(r, g) \leq S(r, f) + S(r, g)$ which is a contradiction.

Using the second fundamental theorem, one can deduce the following two lemmas.

Lemma 2.7 *Let f and g be two non-constant meromorphic functions. Denote $F = f^n(f^m - 1)f'$ and $G = g^n(g^m - 1)g'$. $F_0 = F'$, and $G_0 = G'$. If $F_0 \equiv G_0$. Then $F \neq G$.*

Lemma 2.8 *Let F_0 and G_0 denote as lemma 2.7, if $F_0 \equiv G_0$, and $n \geq 5m + 37$, then $f \equiv g$ or $f \equiv -g$.*

69.3 Proof of Theorem 1

We support $F \neq G$, since F and G share 1 IM. If $H \neq 0$ and by the definition of H and lemma 2.1, we have

$$N_E^{(1)}\left(r, \frac{1}{F-1}\right) \leq N(r, H) + S(r, F) + S(r, G) \leq \bar{N}(r, F) + N_L\left(r, \frac{1}{F-1}\right) + N_L\left(r, \frac{1}{G}\right) + N_0\left(r, \frac{1}{F'}\right) + N_0\left(r, \frac{1}{G'}\right) + \bar{N}_{(2)}\left(r, \frac{1}{F}\right) + \bar{N}_{(2)}\left(r, \frac{1}{G}\right) + S(r, F) \tag{69.9}$$

where $N\left(r, \frac{1}{F'}\right)$ denotes the counting function corresponding to the zeros of F' which are not the zeros of F and $F - 1$, and correspondingly for G . And

$$\begin{aligned} \bar{N}\left(r, \frac{1}{F-1}\right) + \bar{N}\left(r, \frac{1}{G-1}\right) &\leq \bar{N}_{(2)}\left(r, \frac{1}{F}\right) + \bar{N}_{(2)}\left(r, \frac{1}{G}\right) + \bar{N}\left(r, \frac{1}{F}\right) \\ &+ N_E^{(1)}\left(r, \frac{1}{F-1}\right) + 3N_L\left(r, \frac{1}{F-1}\right) + 3N_L\left(r, \frac{1}{G-1}\right) \\ &+ 2N_E^{(2)}\left(r, \frac{1}{G-1}\right) + N_0\left(r, \frac{1}{F'}\right) + N_0\left(r, \frac{1}{G'}\right) + S(r, f) \end{aligned} \tag{69.10}$$

Noting that

$$\begin{aligned} N_E^{(1)}\left(r, \frac{1}{F-1}\right) + 2N_L\left(r, \frac{1}{F-1}\right) + N_L\left(r, \frac{1}{G-1}\right) + 2N_E^{(2)}\left(r, \frac{1}{G-1}\right) \\ \leq T(r, F) + O(1) \end{aligned} \tag{69.11}$$

hence we have

$$\begin{aligned} \bar{N}\left(r, \frac{1}{F-1}\right) + \bar{N}\left(r, \frac{1}{G-1}\right) &\leq \bar{N}_{(2)}\left(r, \frac{1}{F}\right) + \bar{N}_{(2)}\left(r, \frac{1}{G}\right) + \bar{N}(r, F) \\ &+ N_L\left(r, \frac{1}{F-1}\right) + 2N_L\left(r, \frac{1}{G-1}\right) + T(r, F) + N_0\left(r, \frac{1}{F'}\right) \\ &+ N_0\left(r, \frac{1}{G'}\right) + S(r, f) \end{aligned} \tag{69.12}$$

From the second fundamental theorem, we have

$$T(r, F) \leq \bar{N}\left(r, \frac{1}{F}\right) + \bar{N}(r, F) + \bar{N}\left(r, \frac{1}{F-1}\right) - N_0\left(r, \frac{1}{F'}\right) + S(r, F) \tag{69.13}$$

$$T(r, G) \leq \bar{N}\left(r, \frac{1}{G}\right) + \bar{N}(r, G) + \bar{N}\left(r, \frac{1}{G-1}\right) - N_0\left(r, \frac{1}{G'}\right) + S(r, G) \tag{69.14}$$

From the (69.12–69.14) we have

$$\begin{aligned} T(r, F) + T(r, G) &\leq N_2\left(r, \frac{1}{F}\right) + N_2\left(r, \frac{1}{G}\right) + 2\bar{N}(r, F) + \bar{N}(r, G) + N_L\left(r, \frac{1}{F-1}\right) \\ &+ 2N_L\left(r, \frac{1}{G-1}\right) + T(r, F) + S(r, f) \end{aligned} \tag{69.15}$$

$$\begin{aligned} T(r, G) &\leq N_2\left(r, \frac{1}{F}\right) + N_2\left(r, \frac{1}{G}\right) + 2\bar{N}(r, F) + \bar{N}(r, G) \\ &+ N_L\left(r, \frac{1}{F-1}\right) + 2N_L\left(r, \frac{1}{G-1}\right) + S(r, f) \end{aligned} \tag{69.16}$$

$$N_2\left(r, \frac{1}{F}\right) \leq 2\bar{N}\left(r, \frac{1}{F}\right) = 2\bar{N}\left(r, \frac{1}{f}\right) + 2\bar{N}\left(r, \frac{1}{f^m} - 1\right) + 2\bar{N}\left(r, \frac{1}{f'}\right) + S(r, f) \tag{69.17}$$

Let $f^m - 1 = (f - \beta_1) \cdots (f - \beta_m)$ and $\beta_i (i = 1, 2, \dots, m)$ are different. By lemma 2.2, we have

$$\begin{aligned} N_2\left(r, \frac{1}{F}\right) &\leq 2\bar{N}\left(r, \frac{1}{F}\right) = 2\bar{N}\left(r, \frac{1}{f}\right) + 2\sum_{i=1}^m \bar{N}\left(r, \frac{1}{f - \beta_i}\right) + 2\bar{N}\left(r, \frac{1}{f'}\right) + S(r, f) \\ &\leq 6\bar{N}\left(r, \frac{1}{f}\right) + 2\sum_{i=1}^m \bar{N}\left(r, \frac{1}{f - \beta_i}\right) + 2\bar{N}(r, f) + S(r, f) \end{aligned} \tag{69.18}$$

by the same way,

$$N_2\left(r, \frac{1}{G}\right) \leq 6\bar{N}\left(r, \frac{1}{g}\right) + 2\sum_{i=1}^m \bar{N}\left(r, \frac{1}{g - \beta_i}\right) + 2\bar{N}(r, g) + S(r, g) \tag{69.19}$$

$$\begin{aligned} N_L\left(r, \frac{1}{F-1}\right) &\leq N\left(r, \frac{F}{F'}\right) \leq \bar{N}(r, f) + \bar{N}\left(r, \frac{1}{f}\right) + \bar{N}\left(r, \frac{1}{f^m - 1}\right) \\ &\quad + \bar{N}\left(r, \frac{1}{f'}\right) + S(r, f) \end{aligned} \tag{69.20}$$

$$\leq 2\bar{N}(r, f) + 3\bar{N}\left(r, \frac{1}{f}\right) + \sum_{i=1}^m \bar{N}\left(r, \frac{1}{f - \beta_i}\right) + S(r, f) \tag{69.21}$$

$$N_L\left(r, \frac{1}{G-1}\right) \leq 2\bar{N}(r, g) + 3\bar{N}\left(r, \frac{1}{g}\right) + \sum_{i=1}^m \bar{N}\left(r, \frac{1}{g - \beta_i}\right) + S(r, g) \tag{69.22}$$

From (69.16) to (69.21) we have

$$\begin{aligned} T(r, G) &\leq 3\sum_{i=1}^m \bar{N}\left(r, \frac{1}{f - \beta_i}\right) + 3\sum_{i=1}^m \bar{N}\left(r, \frac{1}{g - \beta_i}\right) \\ &\quad + 5\bar{N}(r, g) + 6\bar{N}(r, f) + 9\bar{N}\left(r, \frac{1}{f}\right) + 9\bar{N}\left(r, \frac{1}{g}\right) + S(r, f) + S(r, g) \end{aligned} \tag{69.23}$$

$$\begin{aligned} T(r, F) &\leq 3\sum_{i=1}^m \bar{N}\left(r, \frac{1}{g - \beta_i}\right) + 3\sum_{i=1}^m \bar{N}\left(r, \frac{1}{f - \beta_i}\right) \\ &\quad + 5\bar{N}(r, f) + 6\bar{N}(r, g) + 9\bar{N}\left(r, \frac{1}{g}\right) + 9\bar{N}\left(r, \frac{1}{f}\right) + S(r, f) + S(r, f) \end{aligned} \tag{69.24}$$

From (69.22) and (69.23), we have

$$\begin{aligned}
 T(r, F) + T(r, G) \leq & 6 \left(\sum_{i=1}^m \bar{N} \left(r, \frac{1}{f - \beta_i} \right) + \sum_{i=1}^m \bar{N} \left(r, \frac{1}{g - \beta_i} \right) \right) \\
 & + 11(\bar{N}(r, g) + \bar{N}(r, f)) + 18 \left(\bar{N} \left(r, \frac{1}{f} \right) + \bar{N} \left(r, \frac{1}{g} \right) \right) + S(r, f) + S(r, g)
 \end{aligned}
 \tag{69.25}$$

Since

$$T(r, F) \geq (m + n - 2)T(r, f) + S(r, f) \tag{69.26}$$

and

$$T(r, G) \geq (m + n - 2)T(r, g) + S(r, g) \tag{69.27}$$

From (69.24) and the first fundamental theorem, we have

$$(n - 5m - 37)(T(r, f) + T(r, g)) \leq S(r, f) + S(r, g)$$

which is impossible. Thus, we have $H \equiv 0$. From (69.17) and (69.18), we have

$$\begin{aligned}
 \bar{N}(r, F) + \bar{N}(r, G) + \bar{N} \left(r, \frac{1}{F} \right) + \bar{N} \left(r, \frac{1}{G} \right) \leq & \sum_{i=1}^m \bar{N} \left(r, \frac{1}{f - \beta_i} \right) + \sum_{i=1}^m \bar{N} \left(r, \frac{1}{g - \beta_i} \right) \\
 & + 3(\bar{N}(r, g) + \bar{N}(r, f)) + 2 \left(\bar{N} \left(r, \frac{1}{f} \right) + \bar{N} \left(r, \frac{1}{g} \right) \right) + S(r, f) + S(r, g)
 \end{aligned}$$

From (69.25) and (69.26), we have

$$\bar{N}(r, F) + \bar{N}(r, G) + \bar{N} \left(r, \frac{1}{F} \right) + \bar{N} \left(r, \frac{1}{G} \right) \leq \frac{10 + 2m}{m + -1} T(r) < T(r).$$

where $T(r) = \max(T(r, F), T(r, G))$. From Lemma 2.5, we have $F \equiv G$ or $FG \equiv 1$, and From Lemma 2.6 to Lemma 2.8, we know $f \equiv g$ and $f \equiv -g$.

Acknowledgement The second author is supported by Educational Commission of Hebei Province of China (No.Z2010260).

References

1. Hayman WK (1964) Meromorphic functions. Clarendon Press, Oxford
2. Yi HX, Yang CC (1995) Uniqueness theory of meromorphic functions. Science Press, Beijing
3. Yi Hx (1995) A question of gross and uniqueness of entire functions. Nagoya Math J 138:169–177
4. Fang CY, Fang ML (2002) Uniqueness of meromorphic functions and differential polynomials. Comput Math Appl 44(5–6):607–617
5. Lin WC, Yi HX (2004) Uniqueness theorems for meromorphic function. Indian J Pure Appl Math 35(2):121–132

6. Qiu HL, Fang ML (2004) On the uniqueness of entire functions. Bull Korean Math Soc 41(1):109–116
7. Xu JF, Yi HX (2007) Uniqueness of entire functions and differential polynomials. Bull Korean Math Soc 44(4):623–629

Chapter 70

The Application of Particle Swarm Optimization in Stock Prediction and Analysis

Guorong Xiao

Abstract Data mining is a new emerging research field in recent years which has mixed many subjects, such as database technology, artificial intelligence, and statistics and so on; its ultimate aim is to find hidden information and knowledge from mass data. This article has carried research on association rule based on particle swarm optimization, proposed association rule mining frame based on self-adaptation swarm, applied particle swarm optimization on mass stock data, worked out data algorithm design of this application; through dealing with primitive stocks, it can not only dig hidden rules behind deal data but also verify the efficiency of the algorithm, besides, it can dig association rules totally.

Keywords Particle swarm optimization • Stock prediction • Association rule

70.1 Introduction

Along with the rapid development of national economy, national monetary and securities market has got swift development; financial market has drawn more and more attentions. As a kind of portfolio investment, stock market is a risk investment which is most often seen. To get profit and avoid risk, there are many traditional means to make prediction and analysis on stock. However, monetary and securities market is a huge system which is influenced by many factors; stock market has complex rules; unilateral traditional prediction and analysis cannot

G. Xiao (✉)

Department of Computer Science and Technology, GuangDong University of Finance,
Guangzhou, China

e-mail: newducky@126.com

meet investors' need any more. With the development of advanced technology and intelligence decision support, many financial instruments have begun to apply data mining and artificial intelligence to mine valuable rules from mass deal data so that to provide investors better reference. In transaction of financial market stock dealing, there are mass data remitting into data base; as listed companies exists cooperation and competition, some stock price appear similar or opposite in certain time. To find association rule in stock through data mining, it cannot only make investors know more clear about trends and relation between stocks to make right decision, but also provide support for new application in stock market, also offer new means for new application in financial field.

Swarm intelligence, as an emerging computing technique, has been more and more popular. At present, there are two kinds of main algorithm for swarm intelligence: Ant Colony Algorithm and Particle Swarm Optimization [1, 2]. The former one is the imitation on process of ants' community food gathering, which has been applied on discrete optimization problem; the latter one is the imitation on process of bird flock foraging, which has been developed as a beautiful optimizing tool later. Knowledge acquisition is a key aim of data mining. Nowadays, swarm intelligence has been applied in many fields, such as, multiple target optimization, pattern recognition, decision support and simulation and system identification. Meanwhile, the application of swarm intelligence in knowledge acquisition has drawn more and more attention from people which has become a new research field [3].

Stock trading market data is a kind of time-series data; data mining time-series analysis techniques can find object evolution characteristic and trend of object change in data and apply common qualification analysis and diagram theory analysis. The another character of stock is that it is non-structure or semi-structure text data, so it can apply text data mining technology to find the relations between them and stocks price, and make prediction about future price according to the rule. It can also apply decision-tree to make a decision-tree on every stock and get stock classification. In association rules mining, the application of typical association rules mining can dig relations between stock qualification and stock up and down.

70.2 Stock Prediction and Analysis Means

It is the relation between supply and demand that influence the change of stock market price. Generally speaking, there are two kinds of prediction and analysis: one is basic analysis; and the other one is technology analysis. In stock market, if investors consider whether to invest or which stock to invest, basic analysis can be applied always. It is thought that the relation between supply and demand influence directly stock market; according to economics, finance and banking, financial affairs and investment principles, securities analyst make analysis on basic factors that influence stock market, such as, political economy situation, financial policy,

society economic indicator, political factor, operation characters of distributing and releasing corporation, operation situation, administrative decision, financial position, sales, dividend distribution and competitiveness to evaluate investment. Its theory is based on stability, which thought that every investment has close relation with inner value; for example, stock price is influenced by inner and external factors, but stock market price is changed along with its inner value; the society provide buying or selling opportunity for investors, and this price wave will be finally rectified.

Definition suppose stock data base $D = \{R_1, R_2 \dots R_n\}$; R has recorded stock i 's history quotation, ($1 \leq i \leq n$), $R_i = \{C_i, S_1, S_2, \dots S_m\}$, C_i is R_i 's stock code, m is immediate market quotations quantity, S_j records R_j 's No. j immediate market quotation, $1 \leq j \leq m$. S_j includes time, opening price, closing price, maxi-price, min-price, amount of increase, trading volume and volume of business.

Explanation on stock market predictive variable and relevant variable

Aggregative index number: stand for momentum of whole stock market; it is representative variable of market analysis; divided into Shanghai composite index and Shenzhen composite index and so on.

Opening price: the first deal price on that very day. If no deal price within 30 min, it takes closing price as opening price.

Closing price: it is the last deal price on the very day.

Maximum price: the highest deal price on the very day.

Minimum price: the lowest deal price on the very day.

Trading volume: stock deal quantity. "Hand" is the minimum unit of the deal, one hand stands for 100 shares.

Value of deals: trading volume expressed by currency, equal to transaction price multiply by trading volume.

Daily pricing limits: compare stock price at a moment with that at other moment and get rising range or falling range.

70.3 The Application of Particle Swarm Optimization on Stock Prediction

The premise of application of PSO is: there are many rules in data base [4]. Take every rule as a particle; arrange possible rules in proper order to form searched volume; the whole searched volume is a particle swarm. Searching is to search all rules which are proper for fitness function from data base, and then produce frequent item set from candidate item set; frequent item set which meet the need of min-conf and min-int are strong association rules.

Grain structure concept is the first problem to be solved when applies particle swarm optimization that is also called coding. Code is to transform association rule to particle in particle swarm. Each particle is corresponding to an attribute

association project, that is to say, each particle stands for a rule and particle swarm is rule set. We will discuss code of association rule below; when we find an association rule,

$$X_1 \wedge X_2 \wedge \dots \wedge X_n \Rightarrow Y_1 \wedge Y_2 \wedge \dots \wedge Y_n$$

In fact, there is a rule in transaction: if $X_1 \wedge X_2 \wedge \dots \wedge X_n$ is true, then $Y_1 \wedge Y_2 \wedge \dots \wedge Y_n$ is true; this situation proportion in transaction set is greater than support value; and $Y_1 \wedge Y_2 \wedge \dots \wedge Y_n$ under $X_1 \wedge X_2 \wedge \dots \wedge X_n$ proportion is greater than confidence value. $X_1 \wedge X_2 \wedge \dots \wedge X_n; Y_1 \wedge Y_2 \wedge \dots \wedge Y_n$ is a field in transaction.

So, text applies real number; apply a positive integer to stand for attribute value; suppose that mining rules has something to do with association attribute; after attribute discretization, there are m attribute values ($0 \leq m$, and m is a integer). Take one attribute value from every attribute, and make N attribute together as a real number series; x_i stands for No. i attribute's attribute value (0 states no association). Particle swarm's particle can be expressed by a real number series: $\{x_1, x_2, \dots, x_i, \dots, x_n\}$ ($1 \leq i \leq n, 0 \leq x_i \leq m$).

Apply PSO which is based on dynamic self-adaptation as basic particle swarm algorithmic's improved algorithmic to realize mining of association rule. Below are steps,

Step 1 divide particle into N swarm, initialize every particle;

Step 2 calculate every particle's adaptability in every swarm; save initial position and adaptability;

Step 3 for every particle, compare its adaptability with best position P_i ; take better one as the best one;

Step 4 calculate Δp_i^{1g} and Δp_i^g ; renew particle speed and position;

Step 5 judge whether get greatest iterations; if so, turn to step 6, if not, turn to step 2;

Step 6 end iteration, and get best value.

Apply particle swarm algorithm to mine association rule; below are mining flow chart for every swarm,

According to above algorithm, we make association mining's application. In actual example, to mine some association in extreme situation, suppose min-sup as 5%, min-conf as 30%, mining result is as below,

$$430 \Rightarrow 02(\text{sup} = 31.06\%, \text{conf} = 52.39\%) \quad (70.1)$$

That is $RSI(1) = \text{Weaker}$, $RS(2) = \text{Weak} \Rightarrow$ stock on the very day "big fall" (sup = 13.06%, conf = 52.39%)

$$203 \Rightarrow 03(\text{sup} = 29.34\%, \text{conf} = 51.58\%) \quad (70.2)$$

That is $RSI(1) = \text{Weak}$, $RS(3) = \text{Weak} \Rightarrow$ stock on the very day "great change" (sup = 29.34%, conf = 61.68%)

$$100 \Rightarrow 20(\text{sup} = 5.69\%, \text{conf} = 37.03\%) \quad (70.3)$$

Table 70.1 RSI(1) and sup and conf

Rsi(1) ∈ stronger	Big fall	Small fall	change	Small rise	Big rise
Times	0	21	56	128	130
Sup	0	0.28%	3.09%	5.69%	5.71%
Conf	0	5.31%	13.7%	37.03%	37.08%

That is RSI(1) = stronger, => stock on the very day “big rise” (sup = 5.69%, conf = 37.03%)

$$333 \Rightarrow 01(\text{sup} = 23.39\%, \text{conf} = 37.03\%) \tag{70.4}$$

That is RSI(1) = Weak, RSI(2) = Weak, RS(3) = Weak => stock on the very day “small fall” (sup = 23.93%, conf = 56.67%)

$$020 \Rightarrow 20(\text{sup} = 5.23\%, \text{conf} = 38.12\%) \tag{70.5}$$

That is RSI(2) = Weaker => stock on the very day “small rise” (sup = 5.23%, conf = 38.12%)

$$043 \Rightarrow 02(\text{sup} = 19.55\%, \text{conf} = 52.17\%) \tag{70.6}$$

That is RSI(2) = Weaker, RS(3) = Weak => stock on the very day “fall” (sup = 19.55%, conf = 52.17%)

Make analysis on association rule between mining RSI and stock daily pricing limits; carry data statistics during the process of algorithm implementation; then we can find that RSI changes along with different operation which is based on prediction of stock trend. RSI (1) appears stronger situation 335 times; chart 1 is support and confidence between RSI and daily pricing limits in different attribute value,

It can be seen that the corresponding daily pricing limit support is small. But, in the situation of extremum qualification, we pay more attention on its prediction which is also called confidence. From chart 1, when RSI is in “stronger” field, stock price seldom fall and big fall never happen; it states that the stock trend is in good situation. From this, when RSI is “stronger”, the stock price will be higher, also with greater probability for rising Table 70.1.

70.4 Conclusion

Data mining is a new-emerging research field in recent years which has mixed many subjects, such as database technology, artificial intelligence, and statistics and so on; its ultimate aim is to find hidden information and knowledge from mass data. This article has carried research on association rule based on particle swarm optimization, proposed association rule mining frame based on self-adaptation swarm, applied particle swarm optimization on mass stock data, worked out data

algorithm design of this application; through dealing with primitive stocks, it can not only dig hidden rules behind deal data but also verify the efficiency of the algorithm, besides, it can dig association rules totally.

References

1. Li N, Qin Y, Sun D, Zou T (2008) Particle swarm optimization with Mutation operator [C]. In: Proceedings of international conference on machine learning and cybernetics. pp 2251–2256
2. Trelea L-C (2009) The particle swarm optimization algorithm convergence analysis and parametre selection. *Information processing letters* :317–325
3. Liang JH, Suganthan PN (2007) Dynamic Multi-swarm particle opyimizer. In: Proceedings of IEEE conference on swarm intelligence symposium, pp 124–129
4. Zeng J, Cui Z (2009) PSO algorithm. *Comput res dev* 41(8):1333–1338

Chapter 71

Research on a Novel Distributed Multi Agent System Plan Method

Wei jin Jiang, Qing Jiang, Hong Li and Yan Su

Abstract This paper presents a distributed multi-agent planning algorithm based on constraints propagating. In the algorithm, each agent plans itself's goals. The planning is determined by action steps and constraints. Use sequence constraints between action steps to assign the time sequence of action steps, and use equivalence constraints and unequivalence constraints between variables to assign action steps unentirely. Conflict detection and coordination between agent's planning are solved by multi-agent coordination, which is the distributed judgement of constraint consistency between agent's planning. The algorithm is reliable under the deterministic circumstance. Because in the algorithm only the actions, constraints, cause and effect chain, which are related to conflicts, are exchanged between agent, the algorithm has less communication and higher security.

Keywords Multi-agent system (MAS) · Planning algorithm · Constraint propagation · Distributed solution · Cooperation

W. Jiang (✉)

Institute of Computer Application, Hunan University of Commerce, Changsha, China
e-mail: lxh_yyy@163.com

Q. Jiang · H. Li · Y. Su

School of Art Layout, Hunan University of Commerce, Changsha, China
e-mail: qingjiang@gmail.com

H. Li

e-mail: hongli@126.com

Y. Su

e-mail: jiangwj_nudt@163.com

71.1 Introduction

Planning is the agent for the plan worked out to achieve the Goal. The merits and inferior position of planning algorithm direct reflection of the agent's intelligence and flexibility. In Multi-Agent System, the Planning algorithm of agent based on classical planning algorithm; also need to resolve the conflict between different Agent-Plan. In classical planning algorithm, only one enforcement action agent; from the initial status, traversing the state space, generate an action sequence to achieve its objectives. Are different is that there are number of agents and each of agents have their own Goals and Planning; The action of different agents will Conflict and Resulted in planning goals can not be achieved.

Ephrati said proposed Multi-Agent System which based on STRIPS [1]. In Ephrati's system, first, one center agent decompose the whole task into sub-tasks and allocate to other agents; each agent make a sub-planning for the sub-tasks and send it to center agent; last, these elements will be combined into an overall plan by the center agent [2]. In the way of CNP, one agent allocates sub-task by tender mechanism. In Jennings system, there is center organization Agent which has the ability to use the knowledge of other agents' assigned tasks and use to describe the intent of the joint collaboration between the agents [3]. However, the centralized planning system has some limitations: firstly, the central coordinating agent computing power and communication bandwidth limitations, System of the number of agents is not excessive; second, when a failure in the planning agent. Will lead to the failure of the entire mission, people started to look Distributed Planning.

Duryea and Lesser proposed Partial Global Plan (PGP), agents through exchange Partial Global Plant to decompose task and coordination dynamically; and in PGP, the programming and execution of action is alternately; it allows agent to execute action before reaching agreement [4]. This characteristic is very suitable for dynamic environments, but on the whole, considered to be unreliable, and in some areas will bring disastrous results [5]. This problem can be solved by build a comprehensive planning before executing action; research in this area include Conditional planning, Probabilistic Planning and Universal Planning, but they came back to a centralized planning algorithm model [6].

This chapter presents One-Distributed-Multi-Agent-Plan. In this system, each agent is planning to carry out their goals and bound by the plan and to determine action steps. Action steps with the routine sequence of steps bound to the time sequence Equivalence of constraints and variables used restraint is not equivalent to not entirely routine steps [7]; Programming is gradually adding and refining constraints on the planning process, algorithm based on the classic UCPOP Planning algorithm; through a special Multi-Agent arrange to check and clear up collision between Agent-Plan, then restriction their action; we call this process Constraint propagation.

71.2 Base on Constraint Propagations' Distributed-Multi-Agent-Plan

We expand Pednault's Action Description Language (ADL) to describe agent's action; ADL used the premise and effectiveness of the dual action group to describe the action. In order to show different agents' action and the cooperation of several agents in the Multi-Agent System. We join the concept of role in the action the description, describing action with Triples [8]. Saw from the angle of programming, the now of difference body different agent of agent cannot carry out different action, therefore, we used the action that the agent cannot carry out to describe an agent ability [9]. The programming still uses action and act up of stipulation to mean, but did to expand homogeneously in acting peace treaty to tie of meaning. After improving of act a mold piece and act step, agent ability and programming definition as follows.

Define 1 Action modules. Action modules OP is Triples (P, E, R), Among them, P is Prior condition, E Is a result, R is Participating the operative role gathers, Be played by agent. $\langle P, E, R \rangle$ Meaning action can all roles performances for participating in the appearance P from R, produce appearance E. If R has several roles, so that action is a cooperation action, P and Q can include variable. We use Pop to mean to act OP prior conditions, and use EOP- to show the result of action OP, All adopt this kind of to mean a method in this text.

Define 2 Action steps. Action steps OI is an instance of OP. we instantiation OP can get OI.

Define 3 Agent capacity. Agent have some capacity Ca, which can play a certain role to mean in a certain action mold the piece, the agent ability concentration means for : $CaS = \{Ca1, Ca2, \dots, Can\}$.

Define 4 Programming. Programming II is Triples $\langle SO, OO, CS \rangle$. SO is a set of Action steps [10], including agents 'self action and other agents' executive action; The other executive action of SO executed by other agent. Several agents together carry out of the consociation action have to have 1 to start, we unify to mean for from start the performance unite action, here recognize tacitly action medium of the first role for start. OO is the Order set constraints of SO and can be controlled directly on runtime. The Order constraints of different agents between actions can be realized by adding two simultaneous actions. CS is action steps' variables and constants or other variables' equivalence or in-equivalent set of SO, which can subdivide further into equivalent constraints Co(II) and in-equivalent constraints Nonco (II).

Define 5 Causal chains. causal chains described the action as a prerequisite for the establishment of another action; there is causal chains $\langle E, pu, U \rangle$ between E and U.

If and only if,
 $(E \prec U) \in OO$), namely, there is Order constraints between E and U;
 $\exists e_E \in E_E$, make $(eE \approx pu)$, here into $P_u \in P_u$, namely, there is effect of E is premise of;

$\forall T$ if and $(E \prec T), (T \prec U)$ and $\forall e_T \in E_T$, that $\neg(e_T \approx \neg pu)$

The calculate way supposes according to the following:

Beginning's starting the appearance of world is complete, as to it's describe is certain and limited;

BE caused by the agent action to all changes of world appearance, and can be accurate to describe;

The operative performance is certain;

The action is consistent, under namely any circumstance have no action would to increase self-contradict result ϕ and $\neg\phi$ in the meantime in the consistent world appearance;

Agents are cooperation.

Arithmetic uses a sub-function MGU (p, q), this function return a commonly Consistent displacement of atom-predication p and q. if not, return \perp . The form of consistent displacement is accidentally set $\{(u, v)\}$; Means variable U is equal to v in order to make sure that predication P is equal to Q. we can make the Equivalent binding constraint of variable as replacement set.

Planning algorithm from a space planning, including the initial state and the only goal of the state, then passed to constantly add new action plan to improve the planning and restraint. Until all the goals and sub-goals are achieve full [11]. With the goal of adding movement to move module integration will move into a template examples moves example. Meanwhile the action to judge whether China will pose a threat to some of the causal chain planning. New action is not only a threat to the agent's own planning, but also all the other agents in the planning system, Algorithm agent will be sent to other new moves agent. Each agent to detect whether a threat to their own plan: If a threat to. We are to be used to remove the constraints of conflict back to the planning agent. Deciding which agent to be bound by planning conflict resolution, before joining the new binding constraint should determine whether the original agreement. If the deadlock will be inconsistent and needs to choose other constraints, such consistency is bound to ensure the consistency algorithm [12]. If the action does not constitute a threat to the existing chain of causation or possibly through binding to remove the conflict. Then it can be added to their real plans, updating the plans set the causal chain. Set Planning process exists only in the causal chain; the move is intended to facilitate detection of conflict, which does not contain the results of the causal chain planning pool.

71.3 Arithmetic Distributed-Multi-Agent-Plan (OPS, CaS, I, G)

Input: All modules collective action OPS, the ability set of all-agent CaS, current environment state I, Goal G

Output: Programming II

Step 1. Notify all Agent-Plan started, and wait for all agent to enter programming state after executed currently action.

Step 2. initialization, up build programming $\Psi = \langle SO, OO, CS, CLS \rangle$, set $SO = \{\text{Init}, \text{Goal}\}$, $OO = \phi$, $CS = \phi$, $CLS = \phi$. Let Goal G decompose into sub-Goal gather GS, first set $GS = \phi$, then adding GSS $\langle g, \text{Goal} \rangle$ for each atom-predication where $g \in G$.

Step 3. Call recursion arithmetic

$$\phi = \text{Generate} - \text{Next} - \text{Step}(\Psi, GS)$$

Step 4. Programming takes effect.

Step 5. Message all Agent-Plan finish, all agent go on to execute their own programming.

Step 6. If succeed, return ϕ , else Out.

Arithmetic for create an action step:

Arithmetic *Generate - Next - Step*(Ψ, GS)

Input: current programming Ψ , current sub-Goal set GS

Output: a programming ϕ after adding a new action step.

Step 1. End condition. if GS is null, arithmetic succeed, return Ψ .

Step 2. Selecting sub-Goal. Select a sub-Goal $\langle g, A \rangle$ from GS, if there is cause chain $\langle A', \neg g, A \rangle$ in the LS quit for false;

Step 3. Selecting action. Select a existed action from existed set SO, or select a new action from OPS; set action H, and make $\exists e \in E_H$ and MGU $(e, g) \neq \perp$. If there is no a action according with condition quits for false.

Step 4. Programming takes effect, Set $\Psi' = \langle SO', OO', CS' CLS' \rangle$, there into: $SO' = SO \cup \{H\}$, $OO' = OO \cup \{H \prec A\}$, $CS' = CS \cup \text{MGU}(e, g)$, $CLS' = CLS \cup \{\langle H, g, A \rangle\}$ make $CS' = CS - \{\langle g, A \rangle\}$.

Step 5. create sub-Goal, if $H \notin SO$, add $\langle p, H \rangle$ to CS' for each $p \in P_H \setminus CS'$.

Step 6. Check and clear up for collision. Add new cause chain $\langle H, g, A \rangle$ and action H to programming; (IF $H \notin SO$) send to all agent, and request returning cause chain and action which collision with it. Then through programming agent clear up collision, set $l = \langle E, pu, U \rangle$ as a cause chain in CLS, A is a action step C1 in programming, if after the restrict of variable CS' , $\exists e \in E_A$ still makes $\text{MGU}(e, \neg pu) \neq \perp$ and can $E \prec A \prec U$, follow show threaten-cause of action A's effect e is 1.

Upgrade. $U \prec A$.

Degrade. $A \prec E$.

Apart. Select restriction β in existed free variable and make $\text{MGU}(e, \neg pu) \neq \perp$.

Add repair. Add $\langle pu/CS', A \rangle$ to CS' .

Step 7. Restrict consistency judgment.

Step 8. Recursion call $\phi = \text{Generate} - \text{Next} - \text{Step}(\Psi, GS)$, IF true, return ϕ , else quit for false.

The arithmetic is dependable, as long as find out a programming as a result, can promise to is right. The credibility of the arithmetic can use induction to prove: (1) when all of programming of agent have not action in the system, Sure there is no

action conflict and constrained Inconsistency; (2) If before add an action, there is no action conflict and constrained Inconsistency in the system; Arithmetic ensures that after adding an action and related-restriction, and also not exist action conflict and Control inconformity in the system. So we use Chapman's Conflict-Eliminates-Method and YangQian's Control-Consistency-Judgment-Method. The credibility of this method had proved in literatures, as (1), (2) conclude can know that arithmetic is credible.

71.4 Conclusion

The arithmetic which this chapter presented is One-Distributed-Multi-Agent-Plan, which need not central coordinating agent to maintenance all programming, so It's system scope unlikely depends on the restrict of the central coordinating agent's compute ability. In arithmetic, It is exchanged only the action which related with conflict, cause chain and restrict. Compared with the whole programming, It has benefit of little traffic and high security.

Farther work mostly focuses on advancing the efficiency of programming result and optimizes the policy of task allocation etc. The algorithm may not be the optimal solution; through the amendment moves the selection strategy can reduce the steps needed to achieve the Goal of the operation. By introducing such tasks as distribution mechanism tender could be made action possible parallel. Those all can raise the efficiency of Multi-Agent System to finish task.

Acknowledgment This work is supported by the Natural Science Foundation of Hunan Province of China No. 10JJ5064

References

1. Ephrati E, Pollack M, Rosenschein JA (1995) Tractable heuristic that maximizes global utility through local plan combination. In: Proceedings of the 1st international conference on multi-agent systems. San Francisco, CA, USA, pp 94–101
2. Edwin G, Cox MT (2001) Resource coordination in single agent and multi agent systems [A]. In: 13th IEEE international conference on tools with artificial intelligence (ICTAI'01) [C], Dallas, Texas, USA, pp 18–24
3. Jennings NR (1995) Controlling cooperative problem solving in industrial multi-agent systems using joint intentions. *Artif Intell* 75(2):195–240
4. Rana OF, Winikoff M, Padgham L, Harland J (2002) Applying conflict management strategies in BDI agents for resource management in computational grids [A]. In: Oudshoorn. MJ (ed) Proceedings of the 25th Australasian computer science conference (ACSC2002) [C]. Conferences in research and practice in information technology, 4. ACS. Melbourne, Australia, pp 205–214
5. Etxioni O, Hanks S, Weld D, Draper D, Lesh N, Williamson M (2002) An approach for planning with incomplete information. In: Proceedings of the 3rd international conference on principles of knowledge representation and reasoning. Boston, MA, pp 115–125

6. Peot M, Smith D (2002) Conditional nonlinear planning. In: Proceedings of the 1st international conference on artificial intelligence planning systems. College Park, MD, pp 189–197
7. Blythe J, Veloso MM (1997) Analogical replay for efficient conditional planning. In: Proceedings of the 14th national conference on artificial intelligence. Providence, RI, pp 668–673
8. Nguyen TV (2005) Understanding trust: a study of interfirm trust dynamics in Vietnam [J]. *J World Bus* 40(2):203–221
9. Pavlou PA (2002) Institution based trust in inter organizational exchange relationships: the role of online B2B market places on trust formation [J]. *J Strateg Inf Syst* 11(3/4):215–243
10. Jiang W, Zhang L, Pu W (2009) Research on grid resource scheduling algorithm based on MAS cooperative bidding game [J]. *Sci China F* 52(8):1302–1320
11. Barber KS, Han DC (2000) Tse-Hsin Liu strategy selection-based meta-level reasoning for multi-agent problem-solving [A]. *AOSE [C]*, pp 269–283
12. Excellant-Toledo CB, Jennings NR (2004) The dynamic selection of coordination mechanisms [J]. *J Auton Agents Multi-Agent Syst* 9(1–2):55–85

Chapter 72

A Method of Object Detection Based on Improved Gaussian Mixture Model

Jiajia Sun, Yanan Lian and Yumin Tian

Abstract For the traditional Gaussian mixture model has the shortages of low convergence rate and being sensitive to sudden light change, a nearest neighbor n-frame improved algorithm based on chroma space is presented. On one hand, when luminance has an abrupt change, the objects can be extracted by using chroma information of every pixel instead of RGB, so that the disturbance for the detection resulting from sudden light change can be avoided effectively; on another hand, for accelerating the convergence rate and reflecting the dynamic changes of the background in time, the improved algorithm needs to record the matching of Gaussian distributions with measured values in recent n frames, and use the exponential function of matching times to update the weight of models. The results of experiments prove that the improved algorithm proposed in this paper has better convergence and robustness than that of the traditional one.

Keywords Moving object detection · Gaussian mixture model · Chroma · Nearest neighbor n-frame · Matching times

J. Sun (✉) · Y. Lian · Y. Tian
School of Computer Science and Technology, Xidian University,
Xi'an Shanxi 710126, China
e-mail: sunjiajia0427@163.com

Y. Lian
e-mail: lyn1213@126.com

Y. Tian
e-mail: ymtian@mail.xidian.edu.cn

72.1 Introduction

Moving object detection is an important research subject in the field of intelligent supervisor and control, and it is also a pivotal procedure in visual movement analysis. In real life, huge amounts of important visual information are contained in movements. Therefore, for sequence images, segmenting and extracting the moving and changing part from the background image is a basic work for the post process of moving object classification, tracking, behaviour understanding and so on.

Nowadays, Background Image Difference Method is commonly used in moving object detection, and it has become a hot topic in the computer vision field. For instance, Pfinder system [1] utilized the Single Gaussian Distribution model to establish the background model; Gaussian mixture model is an adaptive background modeling method which was proposed by Stauffer and Grimson [2], et al. and the main point is to build several Gaussian statistical models for each background; meanwhile, some improved algorithms have been proposed. The Ref. [3] reduce the complexity of computation by introducing restriction on luminance and color; Sofka [4] addressed the method of replacing the universal threshold with the segmentation threshold of the foreground and the background.

72.2 Improved Gaussian Mixture Model

Though many improved algorithm of Gaussian mixture model have been proposed so far, there are still several weaknesses of traditional Gaussian mixture model. In view of the problems below, a nearest neighbor n-frame improved algorithm based on chroma space is presented.

Low in convergence rate of parameter updating, so that the change of background cannot be reflected in time;

Being sensitive to the sudden change of the universal luminance. On some extreme occasions, the whole video frames are even treated as the foreground.

72.2.1 Improved Algorithm Based on Chroma Space

For colored images, when the luminance in the scene changes, the RGB value of the pixel will change at the same time. Traditional algorithm treats the changes in RGB as being caused by objects' movement. As a result, a wrong outcome can be obtained. To deal with this problem, when luminance has an abrupt change, the object detection work can be accomplished by using chroma information of every pixel instead of RGB. This paper presents a new idea of transforming color spaces so as to eliminate the noise caused by sudden light change and remain the key

information. This method cannot only avoid the disturbance to the result, but also the computation complexity.

Improved algorithm based on chroma space chooses HSB space to replace the RGB space, and avoids the influence to the final detection result caused by the changes in lamination. The relationship between them is shown as follows:

$$\begin{cases} H = 0.2990R + 0.5870G + 0.1140B \\ S = -0.1687R - 0.3313G + 0.5000B + 128 \\ B = 0.5000R - 0.4187G - 0.0813B + 128 \end{cases} \quad (72.1)$$

with the help of image processing software, we can conclude: for an image, the sudden light change would lead to the change of RGB, while the values of chroma and saturation remain the same. In addition, under the condition that luminance and saturation are certain, just adjusting the chroma value can represent most colors, while the color areas of saturation is limited. Therefore, in order to reduce the amount of calculation, we attempt to use chroma space model of each pixel to replace the RGB color space given that the detection result can be little influenced.

However, because of neglecting the luminance information and saturation information of the images, there are some limitations in this improved strategy. To guarantee the detection effectiveness of detection system and avoid the disturbance of luminance change, we can use this improved method to deal with the current frame only when the sudden light changes occur. The luminance can be thought having changed while the number of pixels which doesnot match the existing Gaussian models is bigger than 80% of the total pixels.

72.2.2 Fast Convergence Algorithm Based on Nearest N-Frame

In order to ensure the stability of the model, the weight updated factors of the traditional algorithm is always small and constant. Therefore the convergence rate is usually slow. To deal with this problem, we attempt to seek for a new way to accomplish the updating work, and introduce the idea of getting rid of the updated factors in our study. It is necessary to record the matching times of Gaussian distributions matching with measured values in recent n frames in the improved algorithm. Calculate with normalization method by using a exponential function of matching times, so as to increase the convergence rate and reflect the dynamic changes of the background in time.

A common phenomenon exists in general object detection system: part of the background will be constant for a long period, the weight of a certain Gaussian distribution will increase to be far larger than the other' because of matching with the pixels for times. If the constant background is suddenly changed, the detection result will be influenced, even be wrong. The reason is that when a Gaussian

Fig. 72.1 Original image of the 84th frame



Fig. 72.2 Traditional foreground of Frame 85



distribution with larger weight can no longer successfully reflect the background status, a long time is needed to train an appropriate background model.

If a certain Gaussian model with a small weight at some time matches repeatedly with the measured value later, it indicates the background changes and this model can reflect the background well. To increase the weight of the model rapidly and reflect the dynamic change of the background in time, the following methods are adopted in updating the weight:

Firstly, add a new variable $n_{i,t}$ for each pixel to record the matching number of the i th Gaussian model in the recent n frames at the time t .

Secondly, add a statistic $M_{i,t} = 2^{n_{i,t}}$ for each pixel. If the pixel of current frame matches with the i th Gaussian distribution at the time t , $n_{i,t}$ needs to be added by 1 and $M_{i,t} = 2M_{i,t-1}$;

At the same time, because n frames continue sliding forward, there must be a Gaussian distribution j whose matching number $n_{j,t}$ will reduce by 1 and accordingly $M_{j,t} = M_{j,t-1}/2$.

Finally, the weight of each Gaussian distribution can be obtained according to the normalizing expression as follows:

$$\omega_{i,t} = \frac{M_{i,t}}{\sum_{u=1}^K M_{u,t}} \quad (72.2)$$

72.2.3 Algorithm Implementation and Result Analysis

To test the effectiveness of the new algorithm, in the objective detection experiments, the learning rate is 0.005 and the background segmentation threshold is 0.7, the number of Gaussian mixture model is 3, the initialization weight value is

Fig. 72.3 Improved foreground of Frame 85



Fig. 72.4 Original image of the 45th frame



Fig. 72.5 Traditional background of Frame 55



0.005, the initialization variance is 30 and the nearest video frames is 20. (Fig. 72.1, 72.2).

The effectiveness of the improved algorithm based on chroma space can be tested by the image sequence whose size is 176×144 . A sudden change of luminance occurred when the video turn to the 84th frame. Figure 72.3 shows the moving objects of the 85th frame obtained by using the traditional Gaussian mixture model. Because of the sudden light changes, most pixels do not match with the existing Gaussian models, which leads to the whole frames are extracted as the foreground by the detection system; Fig. 72.4 shows the foreground of the 85th frame obtained by using the improved algorithm based on chroma space. Though the luminance changes, the improved system can extract the object accurately.

Considering the transformation relationship between the foreground and the background, two pieces of video can be used to test the effectiveness of the nearest neighbour n-frame improved algorithm from two points of view.

The first one is a piece of indoor video with transformation from background to foreground, and its size is 250×188 . In this video, the cup on the desk turns to move in frame 45. Figure 72.5 shows the 55th frame of the background image obtained by traditional Gaussian mixture model, there are still some shadows in the image because of the transformation process. Whereas, the improved Gaussian mixture model can eliminate the interruption from shadows and reflect the actual

Fig. 72.6 Improved background of Frame 55



Fig. 72.7 Original image of the 75th frame



Fig. 72.8 Traditional background of Frame 88



Fig. 72.9 Improved background of Frame 88



changes of background truly, the detection result of the 55th frame is shown in Fig. 72.6.

The other one is a piece of outdoor video with transformation from foreground to background. In Fig 72.7, the skiff stop moving in Frame 75, that is to say the skiff became a part of the background. Figures 72.8 and 72.9 respectively are the pictures before and after the algorithm being improved, we can find that the skiff in Fig. 72.8 is still obscure while it is clear when obtained by the improved algorithm. Finally, Table 72.1 shows the number of the video frames when the background models reach convergence after the transformation between the background and the foreground. According to the table, we can conclude, the convergence rate of the improved algorithm is faster than the traditional one.

Table 72.1 Compare of the convergence rate between the traditional algorithm and the improved one

Model	Traditional model	Improved model
From background to foreground	30 frames	10 frames
From foreground to background	35 frames	12frames

72.3 Conclusion

According to analysis of the basic framework, a nearest neighbor n-frame improved algorithm based on chroma space is presented. First, to overcome the shortcomings of the traditional algorithm in being sensitive to sudden light change, the chrominance information of each pixel is used in extracting the object; second, meeting the demand of a faster convergence rate, update the background model via the matching results of Gaussian distributions in the recent n frames; finally, compare and analyze the experimental results, and test the detection effect. By experiment, this method can reduce the negative influence from sudden light change effectively, reflect the dynamic change of the background in time and eliminate the shadows caused by transformation between the background and the foreground.

References

1. Wren CR, Azarba Yejan IA, Darrell A (1997) Pfunder real-time tracking of the human body. *IEEE Trans Pattern Anal Mach Intell* 19(7):780–785
2. Stauffer C, Wel G (2000) Learning patterns of activity using real-time tracking. *IEEE Trans Patterns Anal Mach Intell* 22(8):747–757
3. Li H, Yang J, Ren X, Wu R, (2008) An improved mixture of gaussians model for background subtraction, *ICSP2008 proceedings* 1380
4. Sofka M (2008) Commentary paper on “On stable dynamic background generation technique using Gaussian mixture models for robust object detection”, *IEEE fifth international conference on advanced video and signal based surveillance*, pp 50–51

Chapter 73

A Novel Dynamic Resource Allocation Model Based on MAS Coordination

Wei jin Jiang, Luo Zhong, Yuhui Xu and Yan Su

Abstract The chapter broke through the limits of the traditional optima methods in the dynamic environment and use MAS (Multi Agent System) method to do research on the resource coordination. Based on the coordination reasoning based on BDI, use utility as the subjective judgment of benefit when individual uses resources. The chapter built the BDI resource coordination reasoning model based on utility, and state the calculation, reasoning and reciprocal process of utility of agent. In the end, the experiment results are given. Research of the chapter is of reference value for the research and application of the management decision and decision support system.

Keywords Multi-agent system (MAS) · Resource allocation · Resource coordination · Coordination reasoning · Utility

73.1 Introduction

The resources coordination is various individual and its between the work goal reasonably assigns the resources in the social group [1], causes the resources which a physical ability reasonably uses it to grasp, then can achieve its individual goal in

W. Jiang (✉)

Institute of Computer Application, Hunan University of Commerce, Changsha, China
e-mail: lxh_yyy@163.com

W. Jiang · L. Zhong · Y. Su

School of Computer Science and Technology, Wuhan University of Technology,
Wuhan, China

e-mail: zhongluo@whut.edu.cn

Y. Xu

School of Art Layout, Hunan University of Commerce, Changsha, China

e-mail: xyh5866@163.com

a higher degree and/the community goal. In general, regarding the coordinated demand is as a result of the resources, the entity, the information distributed characteristic as well as between them the interdependence produces [2]. In the resources coordinated question, the individual plan is restricted resources, needs to carry on the appraisal, the adjustment to own plan, and possibly needs (including other individuals) to obtain from the outside may use the resources to be able better to achieve the goal.

Hazelrigg [3] believed, the decision-making namely cannot be changed resource distribution Generalized resources concept including manpower, physical resource, thus it may be known, the assignment resources is a count for much policy-making activity, the traditional resource distribution method generally may transform as mathematics plan question, but often receives the variable number invariable limit, moreover the operand centralized, the unsuitable modern computation debt is balanced, high fault-tolerant and so on the request. In comparison, (the multi-agent system) realizes the resources coordination distributional solution through MAS, not only may the very good solution above question, moreover may support the group decision-making well the situation. In the variable number dynamic change situation, can manifest MAS in particular the superiority which solves in the resources coordination question, it does not like mathematics plan method such need again modelling. Based on this, this chapter utilizes the MAS coordination method research resources coordination question, the discussion enhances the resource distribution efficiency the new method, (Belief–Desire–Intention) infers the model through the establishment based on effectiveness BDI, uses the multi-agent consultation to achieve the resources is coordinated, the case has confirmed this method validity.

73.2 MAS Resources Coordinating

This chapter believed that, the coordination is take the benefit as the center, belongs to selfish rational individual the cooperation question. Under such environment, through the design reasonable society interactive rule instruction individual benefit guidance, may pursue own benefit maximization in the individual in the foundation, effectively achieves the community goal. The individual depends on to the partial goal realization its plan resources demand whether does obtain satisfies, is different as a result of the partial goal importance, as well as may substitute the plan the existence, the different individual (weighs to the resources by resources effectiveness) usually is different by chance. However the Arrow not impossible theory to point out, even if is the community which the rational individual composes also possibly is non-rational, namely it order is by chance cannot be transmitted, creates the local interest and the interests of the whole not symbol, can solve this problem well through the payment monetary award method [4], finally forms Pareto the most superior plan (Pareto-optimality).

Says from the coordinated strategy, the coordinated strategy mainly has the consultation, the arbitration, the election and own adjustment and so on, these strategies respectively have the advantages [5], arbitration may safeguard obtains the feasible result, but the result quality mainly relies on factor and so on arbitration knowledgeability; consultative result quality relative better, and is very close with the general knowledge in resource distribution, the shortcoming is the price Philippines -agent quantity which brings by the correspondence belt is not n MAS carries on consultative time complex is $O(n^2)$; The election and the body adjust the effect is situated between first two between.

Says regarding participation resources coordination agent, but also some important question is the internal inference, in MAS, between the agent coordination and the inference behavior involves to the agent thought condition model. The Agent thought condition model uses for to describe the Agent thought attribute and between them the connection, as well as with active the and so on sensation, plan, coordination, cooperation relations [6], in the Bratman philosophy analysis [7] in the foundation, Jiang W, Zhang L, Pu W [8] proposed the Agent intention model, Pavlou PA [9] then its development are the BDI frame. The BDI theoretical model regards as Agent is the rational main body, through the faith (Belief), the desire (Desire), the intention (Intention) expresses Agent the rationality. We have established the Agent coordinated inference mechanism in the BDI frame foundation [10], studies based on the state-of-art coordinated inference.

The coordinated inference core manifests in the desire definite process and the intention choice, the definite process. By DEL, DIS, INT separately expressed the Agent individual faith, the desire, the intention set, among, $INT = \{x \mid x = (p_x, \pi_x, g_x)\}$, P_x is the realization intention premise collection, $P_x^r \in P_x$ of resources r demand, $\pi \in \Pi$ (a, t) is the realization intention behavior plan ($a \in ACT$, ACT is all atom behavior collection, Π expression power collection, t expressed time), g_x is a goal which the intention finally realizes. Agent interior phase transition process like Fig. 73.1 (Agent acting resources demand, resources tenderer any role, also possibly simultaneously acts two kinds of roles):

Take the resources demand as the example, its inference process as follows:

sets a target, assigns g , $g \in D$, it may by the correspondence, the sensation, internal inference obtain or the artificial hypothesis;

examines the body ability, seeks can realize the goal plan, if $\exists X, X \subseteq INT$, s, t . $x \in X$, $g_x = g$, then X for the possible Italian atlas, to be corresponding π_x ($x \in X$) is may realize the g plan;

seeks the insufficient resources, regarding P_x^r ($x \in X$) all resources r_i , through obtains its information intermediary with the resources intermediary correspondence, the resources produces r_i the lowest tenderer;

carries on the consultation on resources r_i , and renovates faith storehouse BEL; if $\exists X_p, X_p \subseteq X$, s, t . $x \in X_p$, $\forall P_x, P_x \in BEL$, in which contains the resources request to obtain satisfies, namely Prix BEL, then X_p is the reality possible Italy atlas;

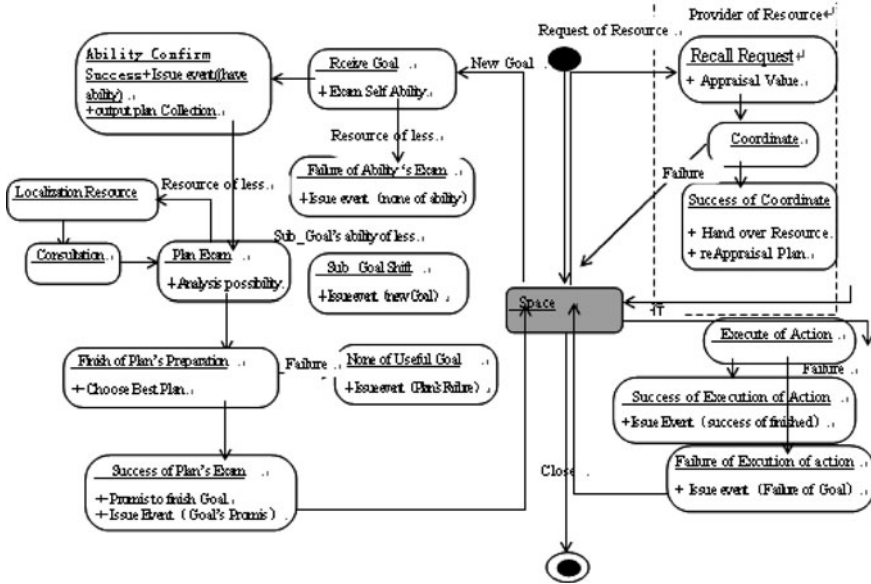


Fig. 73.1 Agent interior phase transition process

if $x_m \in X_p, \forall x \in X_p, F(x_m) \geq F(x)$, then takes x_m is the pledge intention, $F(\cdot)$ is the policy-making choice function, calculates effectiveness choice superior according to agent.

Not difficult to imagine, two kind of roles both need effectiveness change which causes to the resources transfer process to carry on the analysis computation.

73.3 Effectiveness Analyses

By effectiveness (Utility) expressed the individual to the resources by chance, defines as the anticipated income and the cost difference of, namely Utility = Benefit–Cost. In the central resource distribution, may use mathematics plan method for each duty assignment resources, causes the interests of the whole maximization, however, in the distributional resources coordination, the resources mixes must depend upon various individual the judgement as well as each other interactive, is dissimilar as a result of the individual current execution duty cost and the income, causes to lack the resources the individual possibly to have compared to another individual to the resources by chance more intense, very obvious, if the individual is by chance accommodating with the overall situation goal, then chooses by chance a stronger individual and causes the resources to it from other individual giving, is advantageous to promotes the overall situation goal optimization.

By chance the same goal concerns with the information. Because we take cooperate the environment as the supposition, therefore, the individual goal accommodating is very easy to be established, but regarding the individual faith difference, may be alternately regular through the formulation, the control resources shift causes benefit change flowing to dispels with the current capacity. For example, if two Agent has the resources exchange, extracts this letting with to change the letting benefit, then this kind of resources transfer must follow the benefit exchange.

Refers to person and so on Cox in the air operations question research [1], presently take a supply chain management in vehicles transportation dispatch question as the example; a transport company through undertakes to transport the cargo to make a profit, gets down governs various motorcades independent calculation and the turning in partial profits, generally speaking, the motorcade only considered the body the profit growth, has the possibility to harm other motorcades the benefit, but the company actually hoped all motorcades the profit all obtains the growth, reduces this kind of conflict as far as possible, because each motorcade completes the situation regarding the duty to affect company's profit.

In order to be advantageous for the description, supposition that, the transport company has two motorcades, separately needs to transport quantity n_i ($i = 1, 2$, similarly hereinafter) the cargo, is called duty T_i , each motorcade current may dispatch the vehicles have R_j to transport the strength equal freight vehicle. According to contract, if all cargos on time deliver, the transport company may obtain rewards for services rendered Pay_i ; If cannot completely on time deliver, rewards for services rendered in Pay_i to have to deduct is late cargo indemnity $Penalty_i$, does not permit the cancellation or the unlimited time detention duty, simultaneously the supposition distance enough is far, must embark immediately only then can arrive on time, namely:

$$\text{Benefit}_i = \text{Pay}_i \quad (73.1)$$

$$\text{Cost}_i = \text{Penalty}_i + \text{Fee}_i \quad (73.2)$$

$$\text{Utility}_i = \text{Pay}_i - \text{Penalty}_i - \text{Fee}_i \quad (73.3)$$

Among them, the indemnity according to is late cargo proportion P_i and compensates the paying rate Rate_i computation, $\text{Penalty}_i = \text{Pay}_i \times \text{Rate}_i \times P_i$, Fee_i is the duty T_i cartage expense, including fuel, vehicles loss and so on. Says basically as a result of cartage expense Fee_i regarding the duty is invariable, and does not have the influence regarding the question solution, for the simplification question in order to, may leave out it, namely:

$$\text{Utility}_i = \text{Pay}_i - \text{Penalty}_i \quad (73.4)$$

If transports the strength enough, the transport company generally canon time deliver the complete cargo, but in transports in the strength resources scarce situation, for example, the majority vehicles all carried out other tasks, was only left over the minority vehicles to use into carry out these two tasks, was unable completely to

Table 73.1 The effective change when the T₂ transfer resources gives T₁

T ₁ → T ₂	R ₂ > 1	R ₂ = 1
Δ Utility ₁	Pay ₁ × Rate ₁ × $\frac{1}{n_1}$ - Compensation _{1,2}	Pay ₁ (1 - Rate × $\frac{n_1-1}{n_1}$)Compensation _{1,2}
Δ Utility ₂	Compensation _{1,2} - Pay ₂ × Rate ₂ × $\frac{1}{n_2}$	Compensation _{1,2} - Pay ₂ × (1 - Rate × $\frac{n_1-1}{n_1}$)

achieve the contract requirement, then only could reduced the loss in the threshold under the principle to carry on the dispatch, this was very common in the actual situation. When between two motorcades dispatches the vehicles, obtains the vehicles the motorcade (to suppose is x) must to provide the vehicles a side (to suppose is y) pays makes up happy compensation x, y, this has made the elaboration in front. Compensation x, y regarding x said is the cost increase, but to y is the income increase. For example, if reassigns the vehicles resources from T₂ to give T₁, bilateral effectiveness change like Table 73.1. Each duty acts according to own Δ Utility to judge whether carries on the resources transfer, namely is profitable time only then can have the transaction, but compensates compensation causes the key which this kind of transaction produces, if too is high, then the resources deficient duty possibly can give up; if too is low, the resources owner obviously thought own duty are profitable but is not willing to sell the resources. General compensation consults by resources transfer both sides obtains, may express is

$$\begin{aligned}
 \text{Compensation}_{1,2} &= \text{ShareUtility} + \Delta \text{Cost}_2 \\
 &= \text{ShareRatio} \times \sum_{i=1}^2 \Delta \text{Utility}_i + \Delta \text{Cost}_2 \quad (73.5)
 \end{aligned}$$

Namely because the resources transfer causes the always effectiveness change by the resources transfer bilateral share, this conforms to cooperation both sides “the risk altogether to take on, the benefitbuys” principle. The ShareRatio value scope generally for [0, 1], enables both sides both to make a profit from the resources transfer.

On the other hand, if simultaneously has many duties to lack there sources, which duty the resources does assign for, after also must choose obtains the resources the income to grow Δ Utility highest,because this time resources owner can arrive higher Compensation. This also can obtain the appropriate result through the above effectiveness analysis.

Thus, the result was each duty effectiveness all obtained the growth,thus the total effectiveness also similarly obtained the growth, such result also similarly appeared under the dynamic environment, behind produced under a MAS environment to utilize this method the resources coordination experiment result.

73.4 Conclusion

The resources coordination is a kind of important policy-making activity. This chapter with MAS method research resources coordinated, is easy to understand and the application, has the load distribution, fault-tolerant, dynamic, the adaption

surface is broader than the traditional method. Expresses Agent by effectiveness to the resources use by chance subjective judgement, has established based on effectiveness BDI resources coordination inference model. Effectiveness computational method should act according to the concrete question determination.

Acknowledgment This work is supported by the Natural Science Foundation of Hunan Province of China No. 10JJ5064.

References

1. Edwin G, Cox MT (2001) Resource coordination in single agent and multi agent systems [A]. In: 13th IEEE international conference on tools with artificial intelligence (ICTAI'01) [C], Dallas, Texas, USA, pp 18–24
2. Lesser VR (1998) Reflections on the nature of multi-agent coordination and its implications for an agent architecture [J]. *Auton Agents Multi Agent Syst* 1(1):89–111
3. Hazelrigg GA (1996) *Systems engineering: an approach to information-based design* [M]. Prentice-Hall, New York
4. Rana OF, Winikoff M, Padgham L, Harland J (2002) Applying Conflict Management Strategies in BDI Agents for Resource Management in Computational Grids [A]. In: Oudshoorn MJ (ed) *Proceedings of 25th Australasian computer science conference (ACSC2002)* [C], Melbourne, Australia. *Conferences in research and practice in information technology*, vol 4, ACS, pp 205–214
5. Barber KS, Han DC, Liu TH (2000) Coordinating distributed decision making using reusable interaction specification [A]. In: *Design and applications of intelligent agents* [C]. *Lecture Notes in Artificial Intelligence*, vol LNAI1881. Springer, Berlin, pp 1–15
6. Hu S, Shi C (2000) An intention model for agent [J]. *J Softw* 11(10):965–970
7. Nguyen TV (2005) Understanding to trust: a study of interfirm trust dynamics in Vietnam [J]. *J World Bus* 40(2):203–221
8. Jiang W, Zhang L, Pu W (2009) Research on grid resource scheduling algorithm based on MAS cooperative bidding game [J]. *Sci China F* 52(8):1302–1320
9. Pavlou PA (2002) Institution based trust in inter organizational exchange relationships: the role of online B2B market places on trust for mation [J]. *J Strat Inf Syst* 11(3/4):215–243
10. Yi W, Xia H, Chen X (2004) Distributed coordination reasoning with BDI and knowledge level [J]. *Syst Eng* 7:93–98

Part VII
Education and Informatics

Chapter 74

Research on Comprehensive Evaluation of Classroom Teaching Quality Based on Multi-Element Connection Number

Liao Xinfei

Abstract Scientific evaluation to the quality of teaching is an indispensable part in the process of teaching management. But the measure of teacher's classroom teaching is fuzzy and difficult to be clearly decided. The index system on evaluating the classroom teaching quality is given; a new classroom teaching quality evaluation method is presented based on multi-element connection number by using the theory of set pair analysis. It overcomes shortages of original classroom teaching quantity evaluation work, such as subjective and arbitrary and offers scientific, objective, fair, accurate evaluations. A new method and thought is provided for classroom teaching quantity evaluation work.

Keywords Classroom teaching quality · Set pair analysis · Multi-element connection number · Experiment grade · Comprehensive evaluation

74.1 Introduction

Classroom teaching is the basic organizational form of teaching of universities; classroom teaching quality is directly related to the school's educational level, related to the quality of personnel training. Because, the classroom is the main way to acquire knowledge of students, how well students grasp a course, teaching quality is the key. The merits of teaching in higher education have a decisive role. On the comprehensive assessment of the quality of classroom teaching, not only

L. Xinfei (✉)
Office of Academic Affairs, Wenzhou Vocational and Technical College,
Wenzhou 325035, China
e-mail: qtj2011@gmail.com

help to improve the quality of teaching but also for training high quality personnel is also far-reaching significance. How is classroom teaching, its evaluation has some fuzziness, for the same class, different people to evaluate, may vary the evaluation results, and some people think that the teacher put it well, and some people think that the teacher speaks not. In addition, subjective terms are often used in the evaluation, such as “excellent”, “Good”, “medium”, etc. they are no significant quantity boundary, but also with certain fuzziness. How to evaluate teaching effect of teachers in the classroom, qualitative analysis or quantitative evaluation of a single factor was used in the past. These methods are often subjective one-sidedness, not accurate and comprehensive. For this object there is no clear extension, using traditional exact scoring method is not appropriate. In view of nonlinear characteristics and fuzziness in the evaluation process of classroom teaching quality, the fuzzy comprehensive evaluation for teaching quality of teachers is needed. Therefore, this chapter proposes a new evaluation method for classroom teaching quality—multi-element connection number evaluation method on the basis of reasonably determining evaluation index system and weight. In this chapter, the classroom teaching quality evaluation of theory courses by students is studied.

74.2 Concept of Set Pair Analysis and Multi-Element Connection Number

74.2.1 Concept of Set Pair Analysis

Set Pair Analysis (SPA) [1] proposed by Chinese academician Keqin Zhao in 1989, it is a brand new theory of certain and uncertain investigation method. The core of the thought is certain and uncertain can be seen as a system, describe businesses from the connection and transformation businesses, discrepancy and opposite side. The basic concept of set pair analysis is set pair and connection number. Set pair is a pair that the two sets have certain relationship. Set pair analysis analyzes the characters of the two sets by unity of opposites; the main mathematic tool is connection number. The characters of set pair can be described by connection number.

Definition 1 [1] In defined set A and B, they compose a set pair $H = (A, B)$, in the condition W, analysis the character of set pair H, suppose set pair H has number N of the total number of characters, S is the number of character that both sets own, P is the number of character that are contrary between to sets, the number $F = N - S - P$, which means the number of character is neither contrary, nor the same; we call $\mu(H, W)$ is a connection number under condition W.

$$\mu(H, W) = \frac{S}{N} + \frac{F}{N}i + \frac{P}{N}j \quad (74.1)$$

Let $a = \frac{S}{N}$, $b = \frac{F}{N}$, $c = \frac{P}{N}$, $\mu(H, W)$ can be mark as

$$\mu = a + bi + cj \tag{74.2}$$

where a, b, c is called connection component, a represent the amount of similarity of the two sets; b is the difference degree, it represent the discrepancy of the two sets; c is the degree of contrary, represent the contrary of the two sets; $a, b, c \in [0, 1]$ and is real number, and satisfy $a + b + c = 1$; i is coefficient of discrepancy, generally $i \in [-1, 1]$, select the value according to the needs (Sometimes i just play the role of markers); j is coefficient of contrary, it has fixed value -1 (Sometimes j just play the role of markers). It reflects the uncertainty of the business from similarity, discrepancy and contrary. In hence, formula (74.2) is also called the connection number of similarity, discrepancy and contrary or three element connection number; it has been applied in many sectors.

74.2.2 Concept of Multi-Element Connection Number

In practical applications, multi-element connection number will be used.

Definition 2 [2] Multi-element connection number is based on $\mu = a + bi + cj$, and extend bi , please see below formula

$$\mu = a + b_1i_1 + b_2i_2 + \dots + b_ni_n + cj \tag{74.3}$$

Generally, when $n = k$, it calls $k + 2$ element connection number; when $k \geq 2$, the connection number calls multi-element connection number. Below is a five element connection.

For a five-element connection number, the formula can be represent as below (74.4)

$$\mu = a + b_1i_1 + b_2i_2 + b_3i_3 + cj \tag{74.4}$$

In Eq. 74.4, $a, b_1, b_2, b_3, c \in [0, 1]$, and they are real number, also satisfy $a + b_1 + b_2 + b_3 + c = 1, i_1 \in [0, 1], i_2 \in [0, 0]$ is neutral marker, it is not interpreted as $i_2 = 0, i_3 \in [-1, 0], j = -1$

During investigation and work, it is easy to list the expression of five-element connection number. For example, 10 teachers make the overall quality evaluation for a student. If there are five excellent, two that well, one that good, one that qualified, one that completely failed, then the overall quality evaluation result of a student can be represented by following five-element connection number:

$$\mu = 0.5 + 0.2i_1 + 0.1i_2 + 0.1i_3 + 0.1j$$

74.3 Evaluation Process

Multi-element connection number evaluation method is a fuzzy comprehensive evaluation method, it can take consider into the impact of various factors associated with judged things, and then make a general assessment of things [3]. This evaluation method can weigh the factors at all levels, so that the information is not lost, really make the evaluation process involved the evaluation information of each person to play the appropriate role in order to achieve a comprehensive and fair evaluation.

74.3.1 Establish Evaluation Index System

Evaluation index system is a collection of indicators that affect the evaluation objects. Classroom teaching quality evaluation index system relate to the objective, scientific and accurate of evaluation. Evaluation of the classroom teaching quality of teachers is an extremely complex multi-factor, multi-level evaluation issues. Different people will make different evaluation to same evaluation index, different types of courses (such as theory courses, practical training courses, sports and arts courses) require different evaluation index system. There are many evaluation criteria for classroom teaching quality, index system [4] are not the same. This chapter comprehensive assessment a teacher's classroom teaching quality from the teaching content, teaching method, teaching attitude and teaching effectiveness, the specific content shown in Table 74.1.

74.3.2 Establish the Weight Set

Weight is quantitative representation of important degree of each index; its rationality directly affects the accuracy of evaluation result [6]. According to the relevant theories and expert experience, given a weight set of first-level-index W , weight sets of second-level-index W_A , W_B , W_C and W_D .

$$W = (0.3, 0.3, 0.2, 0.2);$$

$$W_A = (0.1, 0.3, 0.3, 0.15, 0.15);$$

$$W_B = (0.35, 0.25, 0.2, 0.2);$$

$$W_C = (0.5, 0.2, 0.1, 0.1, 0.1);$$

$$W_D = (0.5, 0.25, 0.25).$$

Table 74.1 Evaluation indexes of classroom teaching quality [5]

First-level-index	Second-level-index
Teaching content (A)	The teaching contents meet the syllabus (A ₁)
	Basic concepts, basic theory and basic methods are correct (A ₂)
	Lecture clarity, heavy and difficult prominent and language specification (A ₃)
	Reflect the latest trend of academic (A ₄)
	Theory with practice and focus on knowledge application ability of students (A ₅)
Teaching method (B)	Able to use a variety of new teaching methods and new teaching methods (B ₁)
	Able to conscientiously organizing teaching, teaching students according to their aptitude (B ₂)
	Lecture schedule speed is moderate, reasonable schedule of classes (B ₃)
	To encourage students to ask questions and personal views and discuss them (B ₄)
Teaching attitude (C)	This course teaching plan is clear, preparing lessons fully (C ₁)
	Serve as role models, and strictly teach, serious and responsible attitude towards work (C ₂)
	The class on time, involuntary transfer (stop) Class (C ₃)
	Extra counseling, answering questions, correcting homework seriously (C ₄)
	Focus on communication with students, concern for students in all aspects of development (C ₅)
Teaching effectiveness (D)	Students understand and master the basic knowledge of the course, the basic theory and basic skills (D ₁)
	The abilities of students to discover, analyze and solve problem has been enhanced (D ₂)
	The students' practical ability and creative ability got training (D ₃)

74.3.3 Establish the Evaluation Set

Evaluation set is the range of possible evaluation results; they can be vague or non-ambiguous [7]. Let teaching quality evaluation are divided into m grades, and then evaluation set $V = \{v_1, v_2, \dots, v_m\}$. Our school uses five levels, $V = \{v_1, v_2, v_3, v_4, v_5\} = \{\text{excellent, good, medium, pass, fail}\}$. According to the number of evaluation grades, we use the five-element connection number.

74.3.4 Establish the Evaluation Matrix

According to evaluation results for classroom teaching quality of teachers which is done by students, statistics appear frequency all levels v_i of each index (see Table 74.2). Teacher Tom's evaluation results as an example, a total of 100 students have participated in the evaluation.

Table 74.2 Frequency of all levels v_i

	A ₁	A ₂	A ₃	A ₄	A ₅	B ₁	B ₂	B ₃	B ₄	C ₁	C ₂	C ₃	C ₄	C ₅	D ₁	D ₂	D ₃
v_1	95	88	82	78	91	85	75	93	81	76	94	89	79	90	87	92	72
v_2	2	8	11	12	4	6	13	4	9	15	4	5	14	6	11	3	16
v_3	2	2	4	5	3	5	6	2	5	4	1	3	4	2	1	2	7
v_4	1	1	2	3	1	2	3	1	4	3	1	3	1	2	1	2	4
v_5	0	1	1	2	1	2	3	0	1	2	0	0	2	0	0	1	1

According to Table 74.2, after normalized, get the following single index evaluation matrix.

$$R_A = \begin{bmatrix} 95 & 2 & 2 & 1 & 0 \\ 88 & 8 & 2 & 1 & 1 \\ 82 & 11 & 4 & 2 & 1 \\ 78 & 12 & 5 & 3 & 2 \\ 91 & 4 & 3 & 1 & 1 \end{bmatrix} \xrightarrow{\text{normalized}} \begin{bmatrix} 0.950 & 0.020 & 0.020 & 0.010 & 0.000 \\ 0.880 & 0.080 & 0.020 & 0.010 & 0.010 \\ 0.820 & 0.110 & 0.040 & 0.020 & 0.010 \\ 0.780 & 0.120 & 0.050 & 0.030 & 0.020 \\ 0.910 & 0.040 & 0.030 & 0.010 & 0.010 \end{bmatrix}$$

$$R_B = \begin{bmatrix} 85 & 6 & 5 & 2 & 2 \\ 75 & 13 & 6 & 3 & 3 \\ 93 & 4 & 2 & 1 & 0 \\ 81 & 9 & 5 & 4 & 1 \end{bmatrix} \xrightarrow{\text{normalized}} \begin{bmatrix} 0.850 & 0.060 & 0.050 & 0.020 & 0.020 \\ 0.750 & 0.130 & 0.060 & 0.030 & 0.030 \\ 0.930 & 0.040 & 0.020 & 0.010 & 0.000 \\ 0.810 & 0.090 & 0.050 & 0.040 & 0.010 \end{bmatrix}$$

$$R_C = \begin{bmatrix} 76 & 15 & 4 & 3 & 2 \\ 94 & 4 & 1 & 1 & 0 \\ 89 & 5 & 3 & 3 & 0 \\ 79 & 14 & 4 & 1 & 2 \\ 90 & 6 & 2 & 2 & 0 \end{bmatrix} \xrightarrow{\text{normalized}} \begin{bmatrix} 0.760 & 0.150 & 0.040 & 0.030 & 0.020 \\ 0.940 & 0.040 & 0.010 & 0.010 & 0.000 \\ 0.890 & 0.050 & 0.030 & 0.030 & 0.000 \\ 0.790 & 0.140 & 0.040 & 0.010 & 0.020 \\ 0.900 & 0.060 & 0.020 & 0.020 & 0.000 \end{bmatrix}$$

$$R_D = \begin{bmatrix} 87 & 11 & 1 & 1 & 0 \\ 92 & 3 & 2 & 2 & 1 \\ 72 & 16 & 7 & 4 & 1 \end{bmatrix} \xrightarrow{\text{normalized}} \begin{bmatrix} 0.870 & 0.110 & 0.010 & 0.010 & 0.000 \\ 0.920 & 0.030 & 0.020 & 0.020 & 0.010 \\ 0.720 & 0.160 & 0.070 & 0.040 & 0.010 \end{bmatrix}$$

74.3.5 First Level Evaluation

First level evaluation vector $H_i = W_i \cdot R_i, i = A, B, C, D$, we can get

$$H_A = W_A \cdot R_A = (0.8585, 0.0830, 0.0320, 0.0160, 0.0105)$$

$$H_B = W_B \cdot R_B = (0.8330, 0.0795, 0.0465, 0.0245, 0.0165)$$

$$H_C = W_C \cdot R_C = (0.8260, 0.1080, 0.0310, 0.0230, 0.0120)$$

$$H_D = W_D \cdot R_D = (0.8450, 0.1025, 0.0275, 0.0200, 0.0050)$$

74.3.6 Second Level Evaluation

Firstly, establishing second level evaluation matrix R.

$$R = \begin{bmatrix} H_A \\ H_B \\ H_C \\ H_D \end{bmatrix} = \begin{bmatrix} 0.8585 & 0.0830 & 0.0320 & 0.0160 & 0.0105 \\ 0.8330 & 0.0795 & 0.0465 & 0.0245 & 0.0165 \\ 0.8260 & 0.1080 & 0.0310 & 0.0230 & 0.0120 \\ 0.8450 & 0.1025 & 0.0275 & 0.0200 & 0.0050 \end{bmatrix}$$

Then, calculating comprehensive evaluation multi-element connection number μ [8]:

$$\begin{aligned} \mu_{Tom} &= W \cdot R \cdot E = (0.3, 0.3, 0.2, 0.2) \cdot \begin{bmatrix} 0.8585 & 0.0830 & 0.0320 & 0.0160 & 0.0105 \\ 0.8330 & 0.0795 & 0.0465 & 0.0245 & 0.0165 \\ 0.8260 & 0.1080 & 0.0310 & 0.0230 & 0.0120 \\ 0.8450 & 0.1025 & 0.0275 & 0.0200 & 0.0050 \end{bmatrix} \begin{bmatrix} 1 \\ i_1 \\ i_2 \\ i_3 \\ j \end{bmatrix} \\ &= 0.8417 + 0.0909i_1 + 0.0353i_2 + 0.0208i_3 + 0.0115j \end{aligned}$$

74.3.7 An Analysis of Comprehensive Evaluation Multi-Element Connection Number μ

Suppose, comprehensive evaluation multi-element connection number of classroom teaching quality of Jack and John are $\mu_{Jack} = 0.3287 + 0.5482i_1 + 0.1082i_2 + 0.0083i_3 + 0.0066j$ and $\mu_{John} = 0.1432 + 0.1998i_1 + 0.4499i_2 + 0.1347i_3 + 0.0724j$.

1. Sort of evaluation results

Firstly, according to “Equipartition Principle”, let $i_1 = 0.5, i_2 = 0, i_3 = -0.5, j = -1$, calculating the key value $\tilde{\mu}$ of comprehensive evaluation multi-element connection number. The results are as follows. $\tilde{\mu}_{Tom} = 0.8652, \tilde{\mu}_{Jack} = 0.5920, \tilde{\mu}_{John} = 0.1033$.

Then, sorting $\tilde{\mu}$ values, $\tilde{\mu}_{Tom} \succ \tilde{\mu}_{Jack} \succ \tilde{\mu}_{John}$.

Finally, from sort results, we know that teaching quality of Tom is the best, Jack’s second, and John’s the worst.

2. Determine evaluation level

Firstly, according to “Equipartition Principle”, we determine five range of key value of five-element connection number, following as five range, $[0.6, 1]$, $[0.2, 0.6)$, $[-0.2, 0.2)$, $[-0.6, -0.2)$, $[-1, -0.6)$, corresponding excellent, good, medium, pass, fail 5 grades.

Then, according to range of $\tilde{\mu}$ value, determining evaluation level $\tilde{\mu}_{Tom} = 0.8652 \in [0.6, 1]$, we can know that evaluation level of teaching quality of Tom is excellent.

Similarly, we know that evaluation level of teaching quality of Jack and John is “good” and “medium”.

74.4 Conclusions

In this chapter, multi-element connection number evaluation method was introduced to comprehensive evaluation of classroom teaching quality, it overcomes shortcomings of unity and subjectivity in traditional method. At the same time, this evaluation model took into account many factors impact on the evaluation results, give full consideration to the impact of certainty and uncertainty factors. From the evaluation results, multi-element connection number evaluation method not only contains the characteristics of the fuzzy comprehensive method, but with the advantages of no loss of intermediate information and evaluation results objective and reliable, provides a new method and idea for more objective comprehensive evaluation.

References

1. Zhao K (2000) Set pair analysis and preliminary application. Zhejiang Science and Technology Press, Hangzhou (in Chinese)
2. Wang WS, Jin JL, Ding J et al (2009) A new approach to water resources system assessment—set pair analysis method. *Sci China Ser E Tech Sci* 52(10):3017–3023 (in Chinese)
3. Su MR, Yang ZF, Chen B (2009) Set pair analysis for urban ecosystem health assessment [J]. *Comm Nonlinear Sci Numer Simulat* 14:1773–1780
4. Song Z, Gao X (2002) A method of index weight setting in multi-criteria synthetical evaluation [J]. *J Yanshan Univ* 26(1):20–22 (in Chinese)
5. Peng S, Liu G, Xiao Y, Man J (2005) Study of computer fuzzy evaluation system on classroom teaching quality [J]. *Appl Res Comput* 46:46–48 (in Chinese)
6. Iraga Y (2007) Fuzzy evaluation of water quality classification [J]. *Ecol Indicat* 7(3):710–718
7. Biswas R (2005) An application of fuzzy sets in students’ evaluation [J]. *Fuzzy Sets Syst* 74(2):187–194
8. Yu G (1996) The model and application of instructional evaluation of identical discrepancy contrary [J]. *J Shaoxing Coll Arts Sci* 17(6):41–48 (in Chinese)

Chapter 75

Program Design of DEA Based on Windows System

Ma Zhanxin, Ma Shengyun and Ma Zhanying

Abstract A correct and efficient software system is a basic precondition and important guarantee to realize the application of data envelopment analysis (DEA) method. It is an indispensable chain to connect the method and its applications. Therefore, the solving method and the calculation step of C^2R model and BC^2 model are presented. Then the design method and the program of DEA software system on the two models is given.

Keywords Comprehensive evaluation · Multi-objective decision making · Data envelopment analysis · Non-dominated solution · Program design

75.1 Introduction

Data envelopment analysis (DEA) [1–4] is a kind of common and important analysis tools and research technique in the fields of economics, management and evaluation technique [5, 6]. For example, DEA is applied to evaluate the technique progress of enterprises, urban economic situation, financial institutions' efficiency,

M. Zhanxin (✉) · M. Zhanying
School of Economics and Management, Inner Mongolia University,
Hohhot 010021, China
e-mail: em_mazhanxin@imu.edu.cn

M. Zhanxin · M. Shengyun
School of Mathematical Science, Inner Mongolia University,
Hohhot 010021, China

M. Shengyun
School of Science, Inner Mongolia Agricultural University,
Hohhot 010018, China

public management, exploitation and utilization of mineral resource, electric power industry development’s efficiency, forecasting and early warning of system [7–11]. The successful application of DEA in the practice, more and more attracted people’s attention, DEA method obtained the development in many fields, nearly 10,000 chapters about DEA can be searched at home and abroad, especially in recent years, the application of DEA method presented rapid growth trend, and become a hot spot of management science. But in terms of software development, it is still exist the following questions: (1) linear programming software cannot be used to solve the DEA model completely (2) software development of DEA is lagging on the domestic (3) reliable software system of DEA is the fundamental guarantee to the wide application of DEA method. So it is very necessary to develop an efficient and reliable software system of DEA based on Windows and extended.

75.2 The Algorithm Design of C²R Model and BC² Model

Suppose there are n decision making units, where

$$\mathbf{x}_j = (x_{1j}, x_{2j}, \dots, x_{mj})^T$$

and

$$\mathbf{y}_j = (y_{1j}, y_{2j}, \dots, y_{sj})^T$$

represent the input data and output data of the j th decision making unit respectively,

$$\boldsymbol{\omega} = (\omega_1, \omega_2, \dots, \omega_m)^T$$

and

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_s)^T$$

represent the weights of input–output indicators. Let

$$(\mathbf{x}_0, \mathbf{y}_0) = (\mathbf{x}_{j_0}, \mathbf{y}_{j_0})$$

the basic model (P) of DEA is as follows:

$$(P) \begin{cases} \max & (\boldsymbol{\mu}^T \mathbf{y}_0 + \delta \mu_0) = V_P \\ \text{s.t.} & \boldsymbol{\omega}^T \mathbf{x}_j - \boldsymbol{\mu} \mathbf{y}_j - \delta \mu_0 \geq 0, \quad j = 1, 2, \dots, n \\ & \boldsymbol{\omega}^T \mathbf{x}_0 = 1 \\ & \boldsymbol{\omega} \geq 0, \quad \boldsymbol{\mu} \geq 0 \end{cases}$$

$$(D_\varepsilon) \left\{ \begin{array}{l} \min \quad [\theta - \varepsilon(\mathbf{e}^T \mathbf{s}^- + \mathbf{e}^T \mathbf{s}^+)] = V_D \\ \text{s.t.} \quad \sum_{j=1}^n \mathbf{x}_j \lambda_j + \mathbf{s}^- = \theta \mathbf{x}_0 \\ \sum_{j=1}^n \mathbf{y}_j \lambda_j - \mathbf{s}^+ = \mathbf{y}_0 \\ \delta \sum_{j=1}^n \lambda_j = \delta \\ \mathbf{s}^- \geq 0, \mathbf{s}^+ \geq 0, \lambda_j \geq 0, \quad j = 1, 2, \dots, n \end{array} \right.$$

when $\delta = 0$, model (P) is C^2R model [1], when $\delta = 1$, model (P) is BC^2 model [2].

Definition 1 [1] If linear programming (P) exists an optimal solution ω^0, μ^0, μ_0^0 satisfied with $V_p = 1$, then DMU_{j_0} is called weak DEA efficient. If programming (P) exists an optimal solution ω^0, μ^0, μ_0^0 satisfied with $\omega^0 > 0, \mu^0 > 0, V_p = 1$, then DMU_{j_0} is called DEA efficient.

For programming (L₁) and (L₂), we can obtain the following conclusion:

$$(L_1) \left\{ \begin{array}{l} \max(-\theta) = V_{D_1} \\ \text{s.t.} \quad \sum_{j=1}^n \mathbf{x}_j \lambda_j + \mathbf{s}^- = \theta \mathbf{x}_0 \\ \sum_{j=1}^n \mathbf{y}_j \lambda_j - \mathbf{s}^+ = \mathbf{y}_0 \\ \delta \sum_{j=1}^n \lambda_j = \delta \\ \mathbf{s}^- \geq 0, \mathbf{s}^+ \geq 0, \lambda_j \geq 0, \quad j = 1, 2, \dots, n \end{array} \right.$$

$$(L_2) \left\{ \begin{array}{l} \max(\mathbf{e}^T \mathbf{s}^- + \mathbf{e} \mathbf{s}^+) = V_{D_2} \\ \text{s.t.} \quad \sum_{j=1}^n \mathbf{x}_j \lambda_j + \mathbf{s}^- = \theta^* \mathbf{x}_0 \\ \sum_{j=1}^n \mathbf{y}_j \lambda_j - \mathbf{s}^+ = \mathbf{y}_0 \\ \delta \sum_{j=1}^n \lambda_j \\ \mathbf{s}^- \geq 0, \mathbf{s}^+ \geq 0, \lambda_j \geq 0, \quad j = 1, 2, \dots, n \end{array} \right.$$

Lemma 1 [3] Let ε is a non-Archimedean infinitesimal, and the optimal solution of linear programming (D_ε) is $\lambda^0, \mathbf{s}^{-0}, \mathbf{s}^{+0}, \theta^0$, (1) If $\theta^0 = 1$, then DMU_{j_0} is weak DEA efficient. (2) If $\theta^0 = 1$ and $\mathbf{s}^{-0} = 0, \mathbf{s}^{+0} = 0$, then DMU_{j_0} is DEA efficient.

Theorem 1 If $V_{D_1} = -\theta^*$ is the optimal solution of (L_1) and $s^{-*}, s^{+*}, \lambda_j^*, j = 1, 2, \dots, n$ is the optimal solution of (L_2) , then $\theta^*, s^{-*}, s^{+*}, \lambda_j^*, j = 1, 2, \dots, n$ is the optimal solution of (D_ε) .

Proof Because $\theta^*, s^{-*}, s^{+*}, \lambda_j^*, j = 1, 2, \dots, n$ is a feasible solution of (D_ε) . If it is not a optimal solution of (D_ε) , then there must exist a feasible solution $\bar{\theta}, \hat{s}^-, \hat{s}^+, \bar{\lambda}_j, j = 1, 2, \dots, n$ of (D_ε) , subject to $\theta^* - \varepsilon(\hat{e}^T s^{-*} + e^T s^{+*}) > \bar{\theta} - \varepsilon(\hat{e}^T \hat{s}^- + e^T \hat{s}^+)$.

Because $\bar{\theta}, \hat{s}^-, \hat{s}^+, \bar{\lambda}_j, j = 1, 2, \dots, n$ is a feasible solution of (D_ε) , obviously it also a feasible solution of (L_1) .

(1) If $\bar{\theta} = \theta^*$, then $\bar{\theta}, \hat{s}^-, \hat{s}^+, \bar{\lambda}_j, j = 1, 2, \dots, n$ is a feasible solution of (L_2) , and

$$\hat{e}^T s^{-*} + e^T s^{+*} \geq \hat{e}^T \hat{s}^- + e^T \hat{s}^+.$$

Thus

$$\theta^* - \varepsilon(\hat{e}^T s^{-*} + e^T s^{+*}) \geq \bar{\theta} - \varepsilon(\hat{e}^T \hat{s}^- + e^T \hat{s}^+),$$

contradiction!

(2) If $\bar{\theta} \neq \theta^*$, then $\bar{\theta} > \theta^*$. Supposed

$$(\hat{e}^T \hat{s}^- + e^T \hat{s}^+) - (\hat{e}^T s^{-*} + e^T s^{+*}) \leq 0,$$

then

$$\theta^* - \varepsilon(\hat{e}^T s^{-*} + e^T s^{+*}) \leq \bar{\theta} - \varepsilon(\hat{e}^T \hat{s}^- + e^T \hat{s}^+),$$

contradiction!

If

$$(\hat{e}^T \hat{s}^- + e^T \hat{s}^+) - (\hat{e}^T s^{-*} + e^T s^{+*}) > 0,$$

Let

$$\varepsilon = (\bar{\theta} - \theta^*) / (2(\hat{e}^T \hat{s}^- + e^T \hat{s}^+) - 2(\hat{e}^T s^{-*} + e^T s^{+*})),$$

It is obviously that

$$(\bar{\theta} - \theta^*) - \varepsilon((\hat{e}^T \hat{s}^- + e^T \hat{s}^+) - (\hat{e}^T s^{-*} + e^T s^{+*})) > 0.$$

It is contradict with

$$\theta^* - \varepsilon(\hat{e}^T s^{-*} + e^T s^{+*}) > \bar{\theta} - \varepsilon(\hat{e}^T \hat{s}^- + e^T \hat{s}^+).$$

The basic idea of algorithm design to judge DEA efficiency is as following:

(1) Solve the optimal solution of (L_1) . (2) Substitute θ^* of (L_2) with the one in the optimal solution of (L_2) . So we get the optimal solution of (D_ε) .

Table 75.1 The simplex tableau of linear programming (L1)

Iterations	Initial base	C_B	λ^T	θ	s^{+T}	s^{-T}	z^T	δz_δ	\mathbf{b}	α_i
			0_n^T	-1	0_s^T	0_m^T	$-\mathbf{M}e_s^T$	$-\delta M$		
0	s^-	0_m	$\mathbf{X}_{m \times n}$	$-\mathbf{x}_k$	$0_{m \times s}$	$\mathbf{I}_{m \times m}$	$0_{m \times s}$	0_m	0_m	
	\mathbf{z}	$-\mathbf{M}e_s$	$\mathbf{Y}_{s \times n}$	0_s	$-\mathbf{I}_{s \times s}$	$0_{s \times m}$	$\mathbf{I}_{s \times s}$	0_s	\mathbf{y}_k	
	δz_δ	$-\delta M$	δe_n^T	0	0_s^T	0_m^T	0_s^T	δ	δ	
	Test number r_j									

75.3 The Computer Program Design of C²R Model and BC² Model

75.3.1 The Structure and Program Design of DEA Software System

The DEA program in this chapter includes four parts: enter data, show data, calculate and output result. The data can be entered by an input window or a data file. At the same time, we can display the input data by a special window too. Finally, the calculating results can be showed by a window and the corresponding data can be saved to an output file. The algorithms of DEA model are as following:

Step 1: Let the order number k of decision making unit is equal to 0.

Step 2: Let $k = k + 1$, the label f of degradation is equal to 0. If $k > n$, then stop, or else, continue next step.

Step 3: Use big M method to construct the initial base of (L₁) and establish the initial simplex tableau is as follows:

Here $\mathbf{I}_{s \times s}, \mathbf{I}_{m \times m}$ and $\mathbf{I}_{s \times s}$ are unit matrixes, C_B is the objective function coefficient corresponding to the base variable. For convenience, let

$$\mathbf{d} = (d_1, d_2, \dots, d_{m+s+\delta}) = (0_m^T, -\mathbf{M}e_s^T, -\delta M)^T,$$

and the linear programming in Table 75.1 is

$$(P) \begin{cases} \max & \mathbf{c}^T \mathbf{x} = V_P \\ \text{s.t.} & \mathbf{A} \mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq 0 \end{cases}$$

Step 4: If there is a test number

$$r_j = c_j - \sum_{i=1}^{m+s+\delta} d_i a_{ij} > 0,$$

then go to step 5. Otherwise, If $(\mathbf{z}^T, \delta z_\delta) \neq 0$, then print “no feasible solution of linear programming (L₁)”, return to Step 2. If $(\mathbf{z}^T, \delta z_\delta) = 0$, the optimal solution of (L₁) is obtained and go to step 8.

Step 5: When $f = 0$, If

Table 75.2 The simplex tableau of linear programming (L2)

Iterations	Initial base	$\bar{\mathbf{C}}_B$	λ^T	\mathbf{s}^{+T}	\mathbf{s}^{-T}	\mathbf{z}^T	δz_δ	$\bar{\mathbf{b}}$	$\bar{\alpha}_i$
			0_n^T	\mathbf{e}_s^T	\mathbf{e}_m^T	$-\mathbf{M}\mathbf{e}_s^T$	$-\delta\mathbf{M}$		
0	\mathbf{s}^-	\mathbf{e}_m	$\mathbf{X}_{m \times n}$	$0_{m \times s}$	$\mathbf{I}_{m \times m}$	$0_{m \times s}$	0_m	$\theta^* \mathbf{x}_k$	
	\mathbf{z}	$-\mathbf{M}\mathbf{e}_s$	$\mathbf{Y}_{s \times n}$	$-\mathbf{I}_{s \times s}$	$0_{s \times m}$	$\mathbf{I}_{s \times s}$	0_s	\mathbf{y}_k	
	δz_δ	$-\delta\mathbf{M}$	$\delta\mathbf{e}_n^T$	0_s^T	0_m^T	0_s^T	δ	δ	
	\bar{r}_j								

$$r_{j_0} = \max\{r_j | r_j > 0\},$$

then take the j_0 th variable enter the base. When $f = 1$, if

$$j_0 = \min\{j | r_j > 0\},$$

then take the j_0 th variable enter the base.

Step 6: If $a_{ij_0} \leq 0$ for each i , then print "linear programming (L₁) has unbounded solution", return to step 2. Otherwise choose

$$\alpha = \min \left\{ \alpha_i \mid \alpha_i = \frac{b_i}{a_{ij_0}}, \quad a_{ij_0} > 0 \right\}.$$

If

$$i_0 = \min \left\{ i \mid \frac{b_i}{a_{ij_0}} = \alpha, \quad a_{ij_0} > 0 \right\},$$

then choose the i_0 th base variable to leave the base. If there are two or more minimum ratio is equal to α , then let $f = 1$.

Step 7: Take $a_{i_0j_0}$ as the pivot number, operate by using Gauss elimination, then get a new basic feasible solution. Let $d_{i_0} = c_{j_0}$, return to step 4.

Step 8: Let $f = 0$.

Supposed that $-\theta^*$ is the optimal value of (L₁), then establish the initial simplex tableau for linear programming (L₂) is as follows:

For convenience, let

$$\mathbf{d} = (d_1, d_2, \dots, d_{m+s+\delta}) = (\mathbf{e}_m^T, -\mathbf{M}\mathbf{e}_s^T, -\delta\mathbf{M})^T,$$

and note the linear programming in Table 75.2 as following:

$$(P_1) \begin{cases} \max \bar{\mathbf{c}}^T \mathbf{x} = V_{P_1} \\ \text{s.t. } \bar{\mathbf{A}} \mathbf{x} = \bar{\mathbf{b}} \\ \mathbf{x} \geq 0 \end{cases}$$

Step 9: If there is a test number

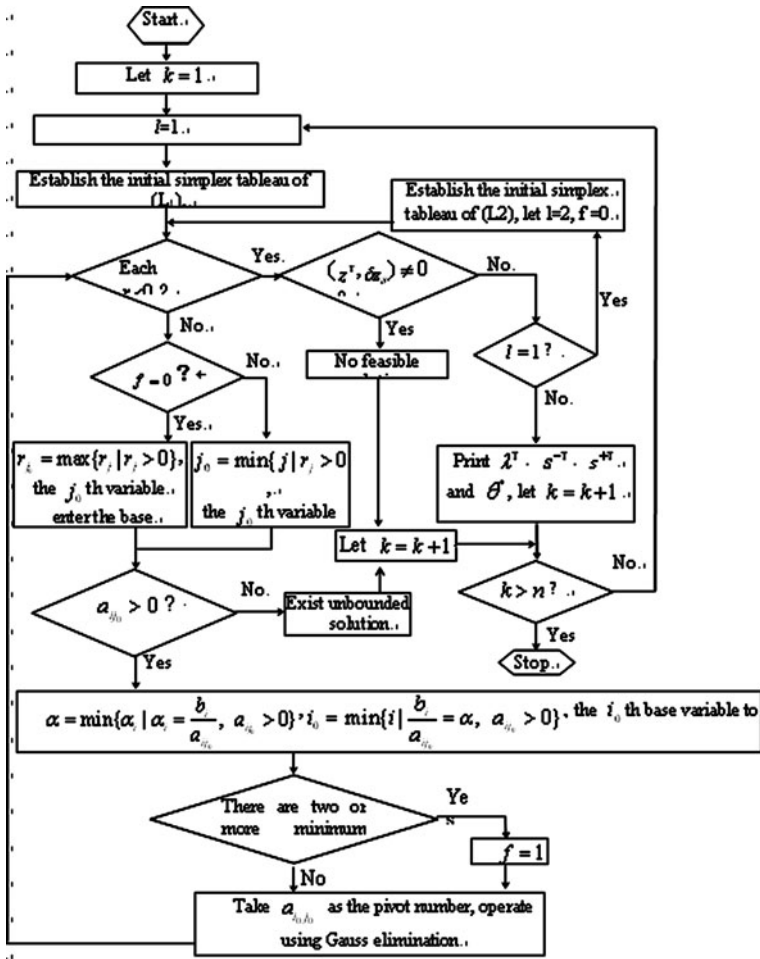


Fig. 75.1 The flow chart of DEA algorithm

$$\bar{r}_j = \bar{c}_j - \sum_{i=1}^{m+s+\delta} d_i \bar{a}_{ij} > 0,$$

then go to step 10. Otherwise, If $(z^T, \delta z_\delta) \neq 0$, then print “linear programming has no feasible solution”, return to step 2. If $((z^T, \delta z_\delta) = 0$, the optimal solution of (L_2) is obtained, go to step 13.

Step 10: When $f = 0$, if

$$\bar{r}_{j_0} = \max\{\bar{r}_j | \bar{r}_j > 0\},$$

then choose the j_0 th variable into the base. When



Fig. 75.2 Input data window

$$f = 1, \quad j_0 = \min\{j | \bar{r}_j > 0\},$$

then choose the j_0 th variable into the base.

Step 11: If $\bar{a}_{ij_0} \leq 0$ for each i , then print “linear programming (L2) has unbounded solution”, return to step 2. Otherwise, choose

$$\bar{\alpha} = \min \left\{ \bar{\alpha}_i \mid \bar{\alpha}_i = \frac{\bar{b}_i}{\bar{a}_{ij_0}}, \quad \bar{a}_{ij_0} > 0 \right\}.$$

If

$$i_0 = \min \left\{ i \mid \frac{\bar{b}_i}{\bar{a}_{ij_0}} = \bar{\alpha}, \quad \bar{a}_{ij_0} > 0 \right\},$$

then choose the i_0 th base variable to be taken from the base. If there are two or more minimum ratio is equal to $\bar{\alpha}$, then let $f = 1$.

Step 12: Take $\bar{a}_{i_0j_0}$ as the main element, operate a Gauss elimination, then get a new basic feasible solution, let $d_{i_0} = \bar{c}_{j_0}$, return to step 9.

Step 13: Print the optimal solution $\lambda^T, s^{-T}, s^{+T}$ and θ^* of (L₂), return to step 2.

The flow chart of algorithm to judge DEA efficiency is showed by Fig. 75.1.

75.3.2 The Design of the Input and Output System

Using EXCEL file as the type of input–output file can improve the accuracy of data entry and editing flexibility. We establish input.xls and output.xls file at first, and enters data through the program window (see Fig. 75.2) or open input.xls file to edit directly. The result output can be automatically completed, the results are displayed in the window in Fig. 75.3, and output to the file output.xls.

	L.1	L.2	L.3	L.4	L.5	L.6	D-1	D-2
单元 1	0	0.2	0.6	0	0	0	0	0
单元 2	0	1	0	0	0	0	0	0
单元 3	0	0	1	0	0	0	0	0
单元 4	0	0	0	1	0	0	0	0
单元 5	0	0	0	0	1	0	0	0
单元 6	0	0.259259259259259	1.00000000000000	0.325925925925926	0	0	0	0

Fig. 75.3 Calculation result display window

The DEA program has the following advantages: programming (L_1) and (L_2) have no non-Archimedes infinitesimal, so it reduces the impact of rounding errors. By using EXCEL software, we can edit input data and output results, calculate function and draw a diagram or make a table very easy. This can greatly reduce the workload and enhance the flexibility of data processing.

Acknowledgments Supported by National Natural Science Foundation of China (70961005, 70501012)

References

1. Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. *Eur J Oper Res* 6:429–444
2. Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Manag Sci* 30:1078–1092
3. Färe R, Grosskopf S (1985) A nonparametric cost approach to scale efficiency. *Scand J Econ* 87:594–604
4. Seiford LM, Thrall RM (1990) Recent development in DEA. The mathematical programming approach to frontier analysis. *J Econ* 46:7–38
5. Cooper WW, Seiford LM, Thanassoulis E et al (2004) DEA and its uses in different countries. *Eur J Oper Res* 154:337–344
6. Ma ZX (2002) Research on the data envelopment analysis method. *Syst Eng Electron* 24:42–46
7. Asmilda M, Paradib JC, Pastorc JT (2006) Centralized resource allocation BCC models. *Int J Manag Sci* (19):1–10
8. Lozano S, Villa G (2004) Centralized resource allocation using data envelopment analysis. *J Prod Anal* 22:143–161
9. Ma ZX, Zhang HJ, Cui XH (2007) Study on the combination efficiency of industrial enterprises. In: *Proceedings of international conference on management of technology*, Aussino Academic Publishing House, Australia, pp 225–230
10. Kao C, Hwang SN (2008) Efficiency decomposition in two-stage data envelopment analysis: an application to non-life insurance companies in Taiwan. *Eur J Oper Res* 185:418–429
11. Ma ZX, Yi R (2009) Research on the economic benefits of industrial enterprises in Western China. In: *Proceedings of international conference on management of technology*, Aussino Academic Publishing House, Australia, pp 19–23

Chapter 76

An Incremental and Autonomous Visual Learning Algorithm Based on Internally Motivated Q Learning

Qu Xinyu, Yao Minghai and Gu Qinlong

Abstract The traditional machine learning paradigm is commonly used for intelligent robot design, which causes problems of low learning initiative, lack of adaptability with uncertainty and bad expansibility of knowledge and ability. According to the new research direction called cognitive development learning, an incremental and autonomous visual learning algorithm based on internally motivated Q learning is proposed. The visual novelty is calculated by online PCA. The active learning and accumulation of knowledge is implemented in the form of updating PCA subspace, which is guided by internally motivated Q learning. By equipped with the proposed algorithm, robot makes next learning decision by judging the novelty between learned knowledge and what is seen now. Experimental results show that the algorithm has the ability of autonomous exploring and learning, actively guiding robot to learning new knowledge, acquiring knowledge and developing intelligence in an online and incremental manner.

Keywords Cognitive development · Internal motivation · Online PCA · Q learning

Q. Xinyu (✉) · Y. Minghai · G. Qinlong
College of Information Engineering, Zhejiang University of Technology,
Road Liuhe. 288, Hangzhou, China
e-mail: qxy2010@163.com

Y. Minghai
e-mail: qxy2010@163.com

G. Qinlong
e-mail: qxy2010@163.com

76.1 Introduction

One of the most attractive and challenging goal in artificial intelligence is to build a robot with human-like intelligence. Traditional machine learning paradigm is commonly used for intelligent robot, which has caused problems in the following aspects [1–3]. (i) In traditional machine learning paradigm, training and testing are divided into two separate phases, using a batch learning mechanism. As a result, robot cannot autonomously explore and learn while carrying out a task in an online, real-time and active manner like a human does. Once it comes to new learning samples, a repeated batch learning is inevitable. Besides, it is costly and time consuming for manually collecting huge amount of training samples rather than actively by robot itself. (ii) Traditional machine learning methods are usually based on the knowledge representation of human programmer, rather than by the robot's own perception, which cause it unable to predict uncertain situation and unable to improve the adaptability in face of uncertain environments and tasks. (iii) Traditional machine learning methods are mostly limited to learn simple and predefined tasks, which lead the robot unable to solve more complex tasks based on simple ones and achieve the human-like intelligence development.

Building robot with traditional machine learning paradigm is becoming more and more dissatisfying with complex environments and high level intelligence. Recent years researchers has proposed a new concept called cognitive development [3, 4] from research on interdisciplinary and mind mechanism. There are plenty of theoretical and practical problems under research with cognitive development learning method [1, 5, 6]. Brooks et al. [7] and Asada et al. [8] consider traditional robot as task-specific, off-line and environmental ideal. They proposed a computational model which simulates prefrontal cortex, perception and motor area in the brain, to solve complex problems in uncertain environment with high level intelligence. They emphasize the interaction way between robot and its environment to develop its intelligence. Kawamura et al. [2] proposed the concept of cognitive control, prefrontal cortex working memory computational model based on Actor-Critic network [9] and decision making method based on affection driven [10], according to the simplified mechanism of human brain. Weng et al. [11, 12] proposed candid covariance free incremental principle component analysis and incremental hierarchical discriminate regression tree to build perception and cognitive mapping, which let the knowledge and behavior develop. Pfeifer et al. [13] emphasize on the embodied cognitive concept, which means intelligence is developed by robot's interaction with its environment. Pierre-Yves Oudeyer et al. [14] proposed intrinsically motivation system based on creature autonomously explore behavior. They build error prediction learning machine to calculate the predicted error between states. Intelligence development is driven by this learned predicted error. Reference [15] researches autonomous mental development similarly with what Weng [16, 17] has done give surveys to the field of cognitive developmental robotics. The existing methods of cognitive developmental learning are still very preliminary and of many disadvantages, such as

dimensional disaster when it comes to huge amount of high dimensional perception, over fitting problem in using the neural network integrated with cognitive learning, still high computational cost in one single incremental step. There is still a great many works need to be done in the cognitive developmental learning method.

The goal of this paper is to propose a new and an effective cognitive development method for visual learning and recognition in intelligent mobile robot, which is expected to solve some problems in traditional machine learning paradigm. In this paper, an internally motivated of visual novelty based Q learning algorithm is proposed according to the concept of internal motivation in biology and psychology [18, 19], to guide robot autonomously explore and learn visual knowledge. The algorithm integrates the adaptive subspace online PCA which is our previous proposed method with Q learning. The visual novelty calculated by online PCA is as the reward function in Q learning. The knowledge active learning and accumulation is in the form of PCA subspace updating, which is guided by internally motivated Q learning. The robot is able to autonomously explore and learn visual knowledge online and actively adjust learning strategy by its own novelty with environment. Experimental results show that the algorithm has the ability of autonomous exploring and learning, actively guiding robot to learning new knowledge, acquiring knowledge and developing intelligence in an online and incremental manner.

76.2 Visual Novelty Based Internal Motivation

The appetite and the eager for live in biology, novelty, curiosity and boredom in psychology are all internal motivation, by which creatures keep themselves live on and develop their intelligence. Inspired by this fact, some computational models can be built to simulate the internal motivation. Then robot autonomous exploration and learning can be guided by internally motivated reinforcement learning and its intelligence can be developed with this method. In this paper, the internal motivation is defined by visual novelty, which is then used for designing the reward function in Q learning. The deference between traditional machine learning paradigm and the proposed learning paradigm is depicted in Fig. 76.1.

Assume that the agent perception is visual input, which is defined as image sequence on time steps, $S_v = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t, \mathbf{x}_{t+1}, \dots)$, where $\mathbf{x}_t = [x_{t1}, x_{t2}, \dots, x_{tm}]^T$ is the m dimensional image input on time step t . The visual novelty N_v is defined as the subtract of image input \mathbf{x}_t on time step t and predicted value $\hat{\mathbf{x}}_E$ from learned knowledge, which is

$$N_v = \|\mathbf{x}_t - \hat{\mathbf{x}}_E\| \quad (76.1)$$

where $\hat{\mathbf{x}}_E$ is obtained by predicting the input vector from learned knowledge. N_v is then used for designing the reward function in Q learning.

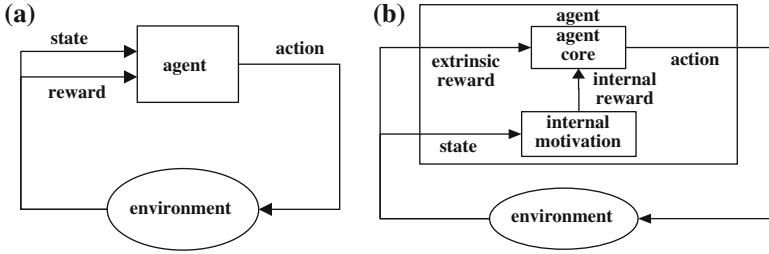


Fig. 76.1 Traditional machine learning paradigm and the proposed learning paradigm **a** traditional machine learning **b** the proposed learning method

However, the dimension of the input image is usually large, which cause highly computational cost and cannot be used in real time. PCA method is widely used in pattern recognition, with simple calculation procedure. Incremental and online manner is needed for robot because of the application of reinforcement learning, so traditional batch PCA should be placed by online PCA. The algorithm applies the adaptive subspace online PCA which is our previous proposed method to realize dimensional reduction online, which is proposed based on the improvement of Ref. [20]. Finally, the formula of visual novelty is presented.

Assume recent time step is t , Some variables are defined as follows: the current subspace \mathbf{U}_t , the current sample mean vector $\bar{\mathbf{x}}_t$, the current coefficient vectors \mathbf{A}_t , a new sample \mathbf{x}_t at time step t , its reconstruction $\hat{\mathbf{x}}_t$ and projection \mathbf{a}_t , the updated subspace \mathbf{U}_{t+1} , the updated mean vector $\bar{\mathbf{x}}_{t+1}$ and the updated coefficient vectors \mathbf{A}_{t+1} .

At time step t , when the new sample \mathbf{x}_t comes, the new subspace can be obtained by calculating PCA through low dimensional coefficient vectors instead of high dimensional reconstructions, since coefficient vectors and reconstructed images encompass the same visual variability, i.e., they are just represented in different coordinate frames. The new sample \mathbf{x}_t is projected onto the current subspace \mathbf{U}_t and the projection is obtained

$$\mathbf{a}_t = \mathbf{U}_t^T (\mathbf{x}_t - \bar{\mathbf{x}}_t) \tag{76.2}$$

Reconstruct the sample

$$\hat{\mathbf{x}}_t = \mathbf{U}_t \mathbf{a}_t + \bar{\mathbf{x}}_t \tag{76.3}$$

Compute the reconstruction error

$$\mathbf{r}_t = \mathbf{x}_t - \hat{\mathbf{x}}_t \tag{76.4}$$

And its Euclidean norm $\|\mathbf{r}_t\|$ which is orthogonal to $\mathbf{U}^{(n)}$, and the new basis is obtained

$$\check{\mathbf{U}}_t = \left(\mathbf{U}_t \frac{\mathbf{r}_t}{\|\mathbf{r}_t\|} \right) \tag{76.5}$$

Build a new coefficient vector matrix

$$\tilde{\mathbf{A}}_t = \begin{pmatrix} \mathbf{A}_t & \mathbf{a}_t \\ \mathbf{0} & \|\mathbf{r}_t\| \end{pmatrix} \quad (76.6)$$

In which \mathbf{A}_t is the coefficient vectors of previous $t - 1$ steps samples. The mean vector $\tilde{\boldsymbol{\mu}}_t$ and subspace $\tilde{\mathbf{U}}_t'$ is obtained by calculating PCA of $\tilde{\mathbf{A}}_t$. The updated coefficient vectors of at time step $t + 1$ is obtained by

$$\mathbf{A}_{t+1} = \tilde{\mathbf{U}}_t'^T (\tilde{\mathbf{A}}_t - \tilde{\boldsymbol{\mu}}_t \mathbf{1}) \quad (76.7)$$

The updated mean vector is

$$\bar{\mathbf{x}}_{t+1} = \bar{\mathbf{x}}_t + \tilde{\mathbf{U}}_t \tilde{\boldsymbol{\mu}}_t \quad (76.8)$$

And the updated subspace is obtained by

$$\mathbf{U}_{t+1} = \tilde{\mathbf{U}}_t \tilde{\mathbf{U}}_t' \quad (76.9)$$

New subspace is obtained by judging $\|\mathbf{r}_t\|$ and the adaptive updating manner are $\dim(\mathbf{U}_{t+1}) = \dim(\mathbf{U}_t) + 1$, $\dim(\mathbf{U}_{t+1}) = \dim(\mathbf{U}_t)$, and $\mathbf{U}_{t+1} = \mathbf{U}_t$. The updated \mathbf{A}_{t+1} , $\bar{\mathbf{x}}_{t+1}$ and \mathbf{U}_{t+1} is then calculated as the initial parameters for next step of online PCA. By adaptive subspace online PCA, high dimensional image perception is converted into low dimensional vector perception, with much computational cost reduced.

With the above online PCA, visual novelty is defined using $\|\mathbf{r}_t\|$

$$n_v = \|\mathbf{r}_t\| = \|\mathbf{x}_t - \hat{\mathbf{x}}_t\| = \|\mathbf{x}_t - (\mathbf{U}_t \mathbf{a}_t + \bar{\mathbf{x}}_t)\| \quad (76.10)$$

Visual novelty n_v at every time step t can be calculated through calculation of recent subspace \mathbf{U}_t in an incremental and online manner.

76.3 Internally Motivated Q Learning

Reinforcement learning [21] is an unsupervised online learning technology, It uses “trial and error” learning mechanism, through the perceived state of the environment and the uncertainty obtain reward value from the environment to learn the optimal behavior strategy of dynamic systems. Based on this, Reinforcement learning is widely used in intelligent control system [22, 23], and proved to be an effective computational method [24], which enhances intelligent system independent learning capability gradually by the experience, in which Q Learning algorithm [25] is the most widely used reinforcement learning algorithm.

The reward function in traditional Q learning is designed according to specific task manually by human programmer, such as the reward function designed for

navigation is for navigation only. Once the task is changed, the reward function should be designed again according to new task manually, which leads to the failure of solving non-task specific problems and autonomously incremental learning new knowledge in a cognitive developmental way with traditional machine learning method. The reward function is extended with internal motivation by integrated internal reward and external reward in Q learning, in which the external reward is as the same as traditional reward, such as +1 reward signal for a task success. The integrated reward is represented as

$$R = \zeta r^{in} + \eta r^{ex}, \quad \zeta, \eta \in [0, 1] \quad (76.11)$$

where r^{in} , r^{ex} internal motivation function and external motivation is function relatively, ζ , η is the weight of r^{in} , r^{ex} relatively. From formula (76.11) we can see that when $\zeta = 0$ and $\eta = 1$, reward is simplified to traditional reward function $R = r^{ex}$. The internal motivation function r^{in} is chosen as visual novelty n_v , $r^{in} = n_v$, so the reward function of internal motivation based Q learning is

$$R = \zeta n_v + \eta r^{ex} = \zeta \|\mathbf{r}_t\| + \eta r^{ex} = \zeta \|\mathbf{x}_t - \hat{\mathbf{x}}_t\| + \eta r^{ex} \quad (76.12)$$

The iterative formula of traditional Q learning

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[r + \gamma \max_{a'} Q(s', a') - Q(s, a) \right] \quad (76.13)$$

Is then modified to internal motivation based Q learning

$$\begin{aligned} Q(s_t, a_t) &\leftarrow Q(s_t, a_t) + \alpha \left[R + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right] \\ \Leftrightarrow Q(s_t, a_t) &\leftarrow Q(s_t, a_t) + \alpha \left[(\zeta \|\mathbf{x}_t - \hat{\mathbf{x}}_t\| + \eta r^{ex}) + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right] \end{aligned} \quad (76.14)$$

In which s_t is the agent's perception of recent time step, presented as

$$s_t = \mathbf{x}_t \quad (76.15)$$

a_t is the action taken by agent at recent time step, which is the action taken to actively percept and learn the visual input.

The flowchart of the proposed algorithm is depicted in Fig. 76.2.

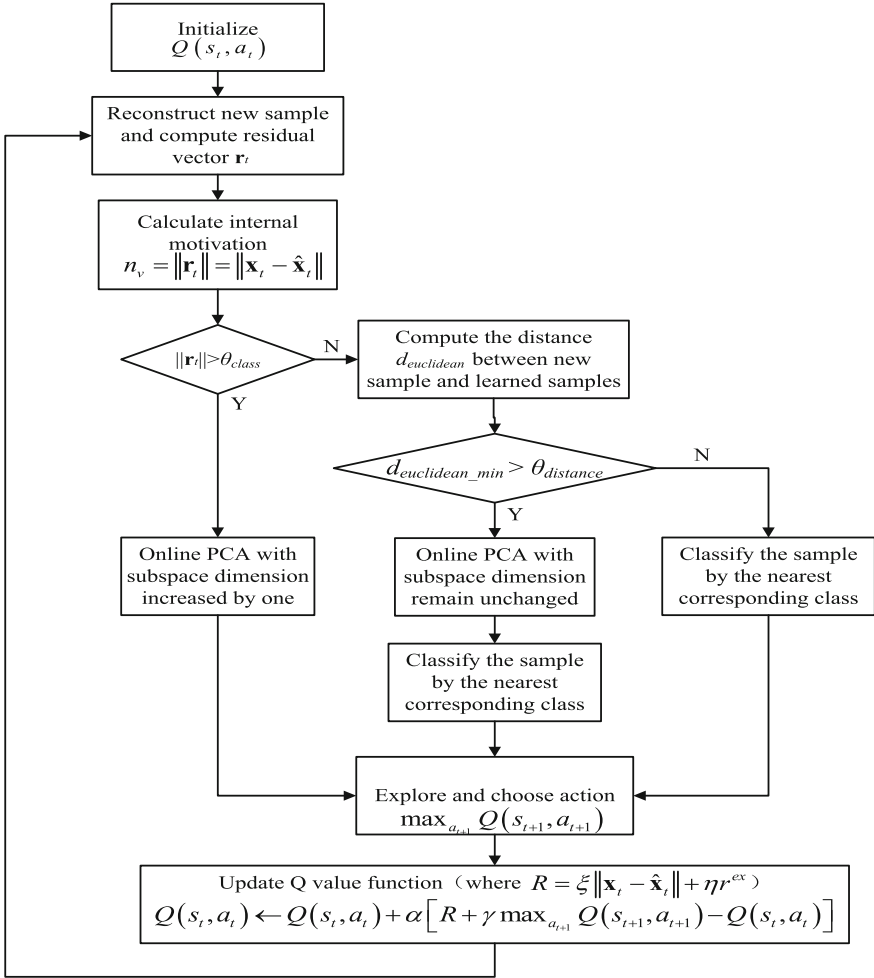


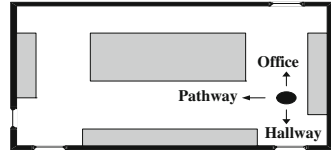
Fig. 76.2 The flowchart of the proposed algorithm

76.4 Experimental Result and Analysis

76.4.1 Experiment Design

Mobile robot visual scene learning and recognition experiments are designed to verify that the proposed algorithm has the ability of autonomous exploring and learning, actively guiding robot to learning new knowledge, acquiring knowledge and developing intelligence in a online and incremental manner. The hardware for experiment is Shanghai Grandar mobile robot AS-R, Sony video camera with two

Fig. 76.3 The experiment environment



degree of freedom and Osprey video capture card. Visual C++ 6.0 and OpenCV open source library for vision are used as the software platform for the proposed algorithm.

In the experiment, the proposed algorithm is running on the mobile robot, with Sony video camera with two degree of freedom and Osprey video capture card to acquire the image input. The image size is converted from 320×240 to 60×45 pixels to reduce the computational cost and obtain real time performance. At every time step, the image dimension acquired to the robot is 2,700.

The experiments are taken place in the laboratory, and three different scenes in the lab is learned by the robot from three views of one position to verify the autonomous and incremental property of the proposed method. The experiment environment is depicted in Fig. 76.3. The black round is present the robot, with facing to the pathway as the initial state. The robot is equipped with three actions: stay still, turn left by 90° and turn right by 90° , from which action the robot percepts different scenes such as pathway, hallway and office, respectively. Autonomous exploring and learning scenes through the choosing of the above three actions is verified by the designed experiment. The three scenes input in the robot by choosing three actions are depicted in Fig. 76.4. In the experiment the parameters of Q learning are $\alpha = 0.6$, $\gamma = 0.8$.

76.4.2 Experimental Result and Analysis

Experiment 1. The three scenes are explored and learned guided by the proposed algorithm with the robot. The trend of the exploration procedure is recorded and the discipline of the exploration with learning steps is analyzed. In the experiment, the robot explores and learns within 500 time steps, divided into five periods, in each of which there are 100 time steps. As depicted in Fig. 76.5, with the proposed algorithm, the visual novelty is high in face of the “pathway” scene in the initial state, as a result, “pathway” is chosen more frequently in the first several time steps than the other actions with the robot stand still facing to the pathway. With the “pathway” visual novelty decreasing, more and more exploration chances are received to the other two scenes until the visual novelty of the three scenes are nearly equal, which means that the exploration and learning of the three scenes are enough and the choosing chances of the three scenes are equal.

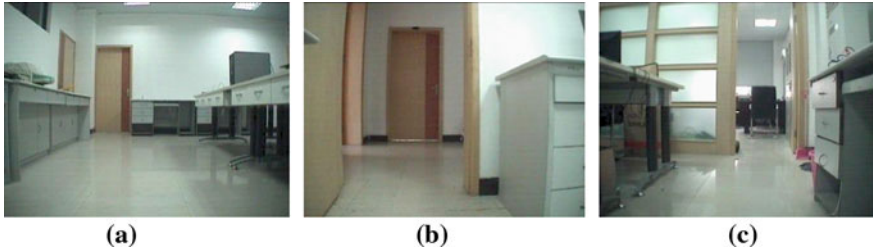


Fig. 76.4 The three scenes input in the robot a Pathway b Hallway c Office

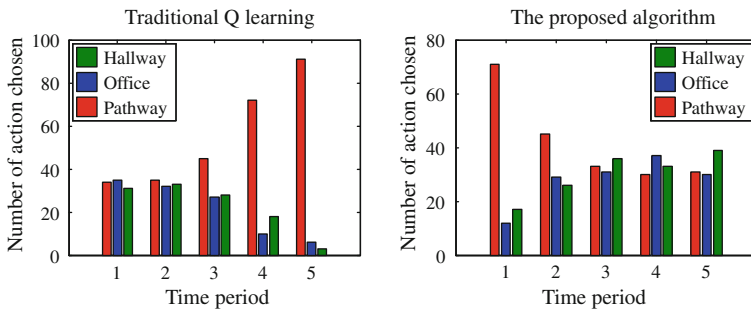


Fig. 76.5 Exploration and learning of traditional Q learning and the proposed algorithm

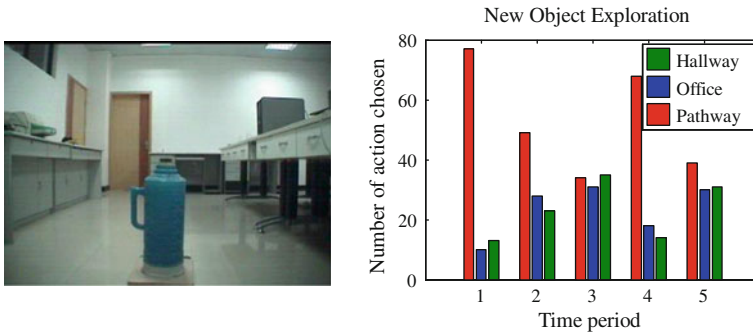


Fig. 76.6 New object exploration with the proposed algorithm

The proposed algorithm and the traditional Q learning is compared, in which the external reward signal is given by human. The maximum positive reward is given to “pathway” and the learning procedure is also recorded as depicted in Fig. 76.5. Traditional Q learning mechanism is based on external and manually

reward, by which the robot is unable to act based on its own judgment. Once the task is changed, the maximum positive reward may not be given to the “pathway”. The manner of manually designing the reward function cannot fulfill the needs of autonomous exploring, learning and intelligence development.

Experiment 2. A new object is placed in the “pathway” scene after enough learning is taken, and the learning procedure after placing the object is recorded. The object is placed at the 301st time step, as depicted in Fig. 76.6, the learning procedure is also recorded as depicted in Fig. 76.6. When the object has been placed in the pathway, the visual novelty of “pathway” increases to a higher level, because it is different with the previous learned three scenes and is novel again to the robot. The algorithm then guides the robot learning “pathway” again, as a result, the choose of “pathway” obviously increases in the 4th period until learning enough, then the choose of three scenes become equal again.

76.5 Conclusion

According to the new research direction called cognitive development learning, an incremental and autonomous visual learning algorithm based on internally motivated Q learning is proposed. The visual novelty is calculated by online PCA. The active learning and accumulation of knowledge is implemented in the form of updating PCA subspace, which is guided by internally motivated Q learning. By equipped with the proposed algorithm, robot makes next learning decision by judging the novelty between learned knowledge and what is seen now. Experimental results show that the algorithm has the ability of autonomous exploring and learning, actively guiding robot to learning new knowledge, acquiring knowledge and developing intelligence in a online and incremental fashion. How to design a more general internal motivation is one of a future direction. Besides, seeking for more effective way to store and search knowledge in a developmental framework is another direction for us.

Acknowledgments We thank the support of China National Science Foundation: No.61070113 and the Zhejiang Science Project Foundation for University students: No. 2010R403071.

References

1. Vernon D, Metta G, Sandini G (2007) A survey of artificial cognitive systems: implications for the autonomous development of mental capabilities in computational agents. *IEEE Trans Evol Comput* 11(2):151–180
2. Kazuhiko K, Stephen G (2006) From intelligent control to cognitive control. In: ISORA 2006 11th international symposium on robotics and applications (ISORA), pp 1–9
3. Weng J, McClelland J, Pentland A, Sporns O, Stockman I, Sur M, Thelen E (2001) Autonomous mental development by robots and animals. *Science* 291(5504):599–600

4. Asada M, Hosoda K, Kuniyoshi Y, Ishiguro H, Inui T, Yoshikawa Y, Ogino M, Yoshida C (2009) Cognitive developmental robotics: a survey. *IEEE Trans Auton Mental Dev* 1(1):12–34
5. Jochen JS, Heiko W (2006) Recent trends in online learning for cognitive robot. In: *ESANN'2006 Proceedings—European symposium on artificial neural networks*, pp 77–87
6. Rachid A, Raja C, Aurelie C, Sara F, Matthieu H, Vincent M, Emrah AS (2006) Towards human-aware cognitive robots. *Am Assoc Artif Intell* 19(6):91–95
7. Brooks R, Breazeal C, Irie R, Kemp C, Marjanovic M, Scassellati B, Williamson M (1998) Alternative essences of intelligence. In: *Proceedings of the American association of artificial intelligence (AAAI)* 42(3–4):146–163
8. Asada M, MacDorman K, Ishiguro H, Kuniyoshi Y (2001) Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robot Auton Sys* 37:185–193
9. Kazuhiko K, Stephen MG, Palis R, Erdem E, Joseph F (2008) Hall implementation of cognitive control for a humanoid robot
10. Gordon SM, Kawamura K, Wilkes DM (2010) Neuromorphically inspired appraisal-based decision making in a cognitive robot. *IEEE Trans Auton Mental Dev* 2(1):17–39
11. Weng J (2004) Developmental robotics theory and experiments. *Int J Humanoid Robot* 1(2):199–235
12. Weng J (2002) A theory for mentally developing robot. In: *Proceedings of the IEEE second international conference on development and learning (ICDL 2002)*, June 12–15, MIT, Cambridge, pp 131–140
13. Pfeifer R, Bongard JC (2006) *How the body shapes the way we think: a new view of intelligence*. MIT Press, Cambridge
14. Oudeyer P-Y, Kaplan F, Hafner V (2007) Intrinsic motivation systems for autonomous mental development. *IEEE Trans Evol Comput*, 11
15. Gao Y, Dong-yue C, Li-ming Z (2005) An exploration of autonomous developing robot with real time vision learning. *J Fudan Univ* 44(6):964–970
16. Qu X, Yao M (2009) Cognitive robotics: a survey. *chinese intelligence and automation conference*, Nanjing
17. Hua-long Y, Chang-ming Z, Hai-bo L, Guo-chang G, Jing S (2007) A survey on developmental robotics. *Caai Trans Intell Sys* 2(4):34–39
18. Deci E, Ryan R (1985) *Intrinsic motivation and self-determination in human behavior*. Plenum, New York
19. Schultz W, Dayan P, Montague P (1997) A neural substrate of prediction and reward. *Science* 275:1593–1599
20. Skocaj D, Leonardis A (2003) Weighted and robust incremental method for subspace learning. In: *Proceedings of the IEEE international conference on computer vision*. Nice, France, pp 1494–1501
21. Mitchell TM (2004) *Machine learning*. Machine Press, Beijing, pp 263–280
22. Xiao NF, Nahavandi S (2002) A reinforcement learning approach for robot control in an unknown environment. In: *Proceedings of the IEEE international conference on industrial technology*. Bangkok II, pp 1096–1099
23. Yichi W, Usher JM (2005) Application of reinforcement learning for agent-based production scheduling. *Eng Appl Artif Intell* 18:73–82
24. Gao Y (2006) *Progress of reinforcement learning research machine learning and its application*. Tsinghua University Press, Beijing, pp 116–134
25. Watkins C, Dayan P (1992) Q-Learning. *Mach Learn* 8:279–292

Chapter 77

Design and Implement of Information System for Water Management Based on SaaS

Lilong Jiang, Junjie Chen, Zhijie Lu and Junjie Li

Abstract According to the problems of the high cost in the development and maintenance of the information system for water management, the information system for water management based on Software as a Service was designed and implemented. With the information system, the Water Supplies Department can make customary configuration, sell water and do data analysis and so on. This paper describes the design and implement of SaaS in the information system for water management. It can serve to reduce the operating cost of the Water Supplies Department and finally promote the informatization of the water industry.

Keywords SaaS · Configurability · Information system for water management · XML

77.1 Introduction

Recently, with the development of information technology, the information system for water management is more and more applied into the water industry, which improves the efficiency of water management greatly [1]. But the traditional information system for water management is for some water supplies department, the cost of purchasing, maintaining and updating software is relatively high. What's more, the technical workers must be employed by the water supplies department to maintain the software. Software as a Service (SaaS) arises as a new software application mode with the development of internet and the mature of

L. Jiang (✉) · J. Chen · Z. Lu · J. Li
Northeastern University, Complex Building 118, Shenyang,
Liaoning, China
e-mail: jianglilong0225@gmail.com

Table 77.1 SaaS maturity model

	Configurability	Multi-tenant-efficiency	Scalability
Level 1	×	×	×
Level 2	√	×	×
Level 3	√	√	×
Level 4	√	√	√

application software [2, 3]. In the SaaS model, the users don't have to purchase software and hardware, construct the IT room and employ technical workers any longer while they can make use of the system by online software rental. This system is an information system for water management based on SaaS. The water supplies department can use the system through internet and order the necessary software services by actual demand. As a result, the running cost of the water supplies department is reduced greatly.

77.2 Introduction of SaaS

SaaS is a new commercial model which software is deployed as managed service and provides service through the internet. The client accesses to services by means of rental. And according to their actual needs they can order relative services which is charged by the content and the duration of the service. The service provider provides upgradable, multi-tenant and configurable services, which simplifies the deployment and maintenance of software. According to configurability, multi-tenant-efficiency and scalability, the SaaS applications maturity can be expressed using a model with four distinct levels, which is given in Table 77.1.

Level 1 and Level 2 Maturity Model can cause high cost of hardware and running, for they design a separate instance for each tenant. Although in terms of architecture, Level 4 Maturity Model with configurability, multi-tenant-efficiency and scalability is the most ideal application architecture, this level needs Load Balance Level and multi-application server and it also involves the complex problems of data horizontal split, a high demand for hardware and software and so on. So the Level 3 Maturity Model is the real SaaS application basic architecture and it is more suitable for the application of Information System for Water Management.

77.3 Design and Implement of Level 3 Maturity Model

77.3.1 Multi-Tenant Data Architecture

There are three approaches to manage Multi-Tenant Data [4, 5], which are "Separate Databases", "Shared Database, Separate Schemas" and "Shared Database, Shared Schema". The comparison of all three approaches is showed in Table 77.2.

Table 77.2 Comparison of three approaches to the management of Multi-Tenant Data

Approach	Isolation level	Share level	Security	Cost
Separate database	High	Low	High	High
Shared database	Middle	Middle	Middle	Middle
Separate schemas				
Shared database	Low	High	Low	Low
Shared schema				

In order to reduce the cost of hardware and maintenance, the information system for water management will adopt the approach of “Shared database, Shared schema”. And this will make the system provide service for more customers with less server resources.

77.3.2 Custom Data Model

Custom Data Model is made up of entity-predefined meta data, entity instance’s data and tenant’s extended meta data. Meta data consists of DataNodes table and DataProperties table, which both of them constitute the data dictionary. DataNodes connecting with each other makes up a tree by parentId and DataProperties table describes the properties of DataNode. The definitions of entity and field both correspond the DataNodes. NodeType is used to distinguish Entity whose value of nodetype is 1 from Field whose value of nodetype is 2. Every row in DataNodes has a TenantId. DataNodes whose tenantId is NULL represent the system predefined meta data while others mean the users’ extended meta datas. It is the same in DataProperties. The relationship digram of database is showed in Fig. 77.1.

Taking into the consideration that different tenant has different price model, in the information system for water management; the price model was created. The price type entity consists of two sub DataNodes which are price DataNode representing the price of the price type and price type representing the type of price. Tenants can extend the predefined price type entity, for example adding pollution price DataNode.

77.3.3 Configurable Invoice

Tenants in the system need to print invoice. But the invoice format differs in different tenants and the common Crystal Report can’t print precisely. The system describes the invoice format with XML, therefore the precise printing and configurable invoice were solved. The XML definition is showed in Fig. 77.2.

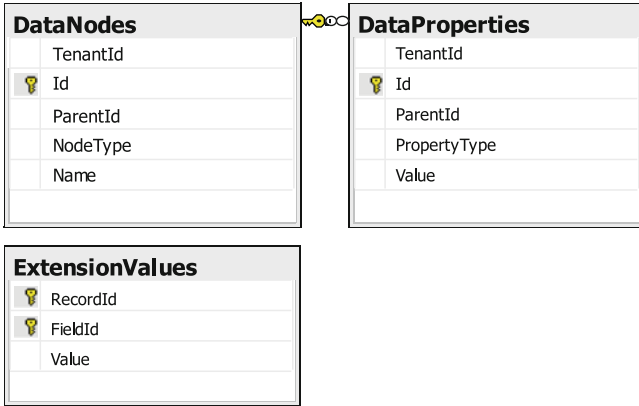


Fig. 77.1 Relationship diagram of database

```

<root>
  <pagesetting>
    <Landscape>>false</Landscape>
    <paperkind>A4</paperkind>
    <paperwidth>200</paperwidth>
    <paperheight>100</paperheight>
    <pageleft>10</pageleft>
    <pageright>10</pageright>
    <pagetop>10</pagetop>
    <pagebottom>10</pagebottom>
  </pagesetting>
  <reporttable>
    <text field="CName" value="Customer Name" x="20" y="20"
    fontsize="10" />
    <text field="NULL" value="Org Name" x="120" y="20"
    fontsize="10">Hefei Water Company </text>
    <text field="Price" value="Price" x="20" y="70" fontsize="10" />
    .....
  </reporttable>
</root>

```

Fig. 77.2 XML definition for invoice template

Pagesettings configuration section defines the page settings of the invoice while reporttable configuration section defines the printing items of the invoice. The printing item of the invoice is defined by Text configuration section. The field attribute of Text is responsible for the binding database field while value attribute marks the item and points out the meaning of the item. If field is NULL, it means this item’s content is fixed and it doesn’t need to fill up. X and Y attributes are the coordinate used to locate the printing content. Fontsize defines the font size of the content.

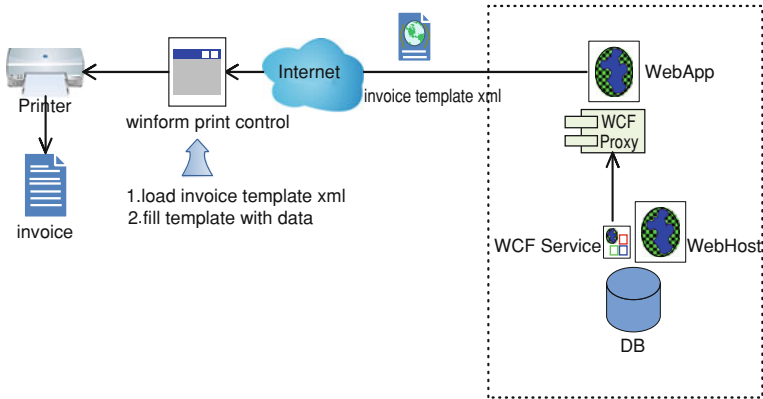


Fig. 77.3 Process of printing invoice

Figure 77.3 illustrates the process of invoice printing. The tenants customize their own invoice template and then the template is saved in the database. When printing invoice, the invoice template will be taken out of the database. On clicking the printing button, the water-selling data will fill up the template and the template will be sent to the winform printing control embedded in the asp.net page which aims to parse the XML and accomplish the task of printing invoice.

77.3.4 Configurable User Interface

The configurable user interface is made up of two parts which are the configurable system menu and the configurable page elements.

77.3.4.1 Configurable System Menu

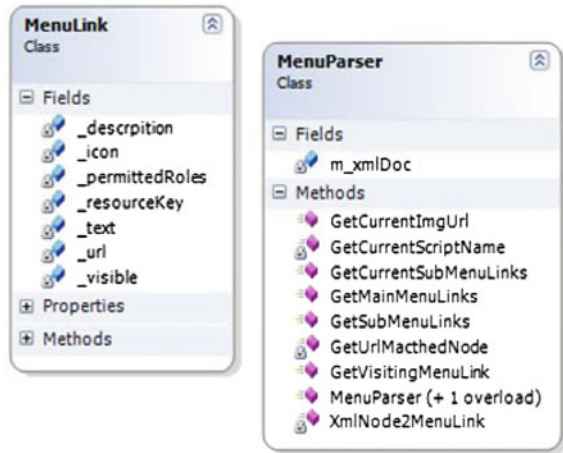
In system, each tenant owes one menu and each sub menu item corresponds one atomic function. The class diagram of class MenuLink and its corresponding parse class MenuParse in showed in Fig. 77.4.

The menu information of each tenant is stored in XML. The XML definition is showed in Fig. 77.5.

77.3.4.2 Configurable Page Elements

As the same with the system menu [6], the content on every function page is also responsible for the interaction between the user and the system. So different tenant will have different demands, whatever the demand are the number, location and showing sequence of the page elements.

Fig. 77.4 The class diagrams of MenuLink and MenuParser



```

<ArrayOfMainMenu xmlns="">
  <MainMenu Text="Water Sale" Url="~/SaleWater/WaterSaling.aspx"
    Roles="Administrator,Salesman" Icon="~/Images/task.gif" >
    <SubMenu>
      <SubMenu Text="Sell Water" Url="~/SaleWater/WaterSaling.aspx"
        Icon="~/Images/admin.gif" Roles="Administrator" Depends="Print
        Invoice"/>
      <SubMenu Text="Print Report" Url="~/Rpt/DateSale.aspx"
        Icon="~/Images/print.gif" Roles="Administrator" />
      <SubMenu Text="Print Invoice" Url="~/SaleWater/InvoicePrint.aspx"
        Icon="~/Images/print.gif" Roles="Administrator" Visible="false"/>
    </SubMenu>
  </MainMenu>
  ...
</ArrayOfMainMenu >

```

Fig. 77.5 XML definition for tenant’s menu

The implement of configurable page elements is mainly through adopting the technologies of Themes and WebPart in .NET. Tenants can replace their themes or skins, add, modify and delete their own webparts.

77.3.5 Configurable Functions

SaaS stresses “use on demand” and “pay per use” [7]. This is a good solution for the differences in the different tenants’ demands and different demands in different times for the same tenant. The function packages in the system are showed in Fig. 77.6.

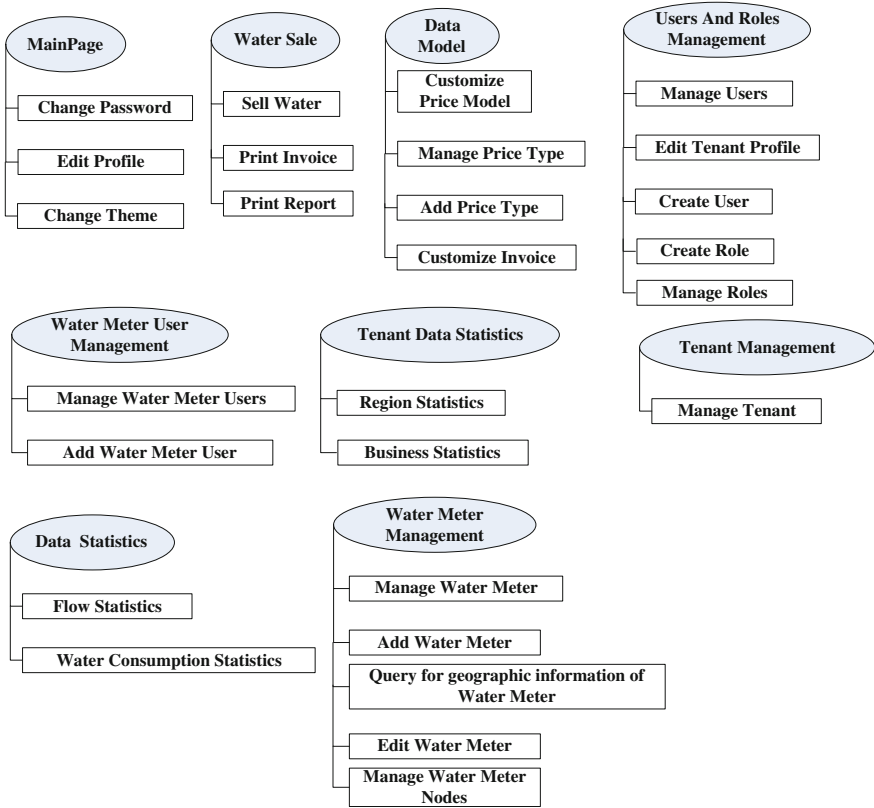


Fig. 77.6 Function Packages

The corresponding relation between function package and menus is stored in XML file showed in Fig. 77.5. Each MainMenu label corresponds one function package and each SubMenu label corresponds one atomic function. The dependence relation between atomic functions is embodies by the Depends attribute in SubMenu. Since each tenant has his own XML file, the configurable functions are implemented. What's more, through Roles attribute in MainMenu and SubMenu, Tenants can create different roles with different rights.

77.4 Conclusions

This chapter introduces basic contents about SaaS, designs and implements the information system for water management based on SaaS. This system can serve to reduce the cost of the water supplies department's purchasing and maintaining

software. Moreover it can enhance the construction of information. Research on SaaS can be referenced by other similar systems.

References

1. Ye W (2009) The software revolution in internet times-design of SaaS architecture. PHEI, Beijing (in Chinese)
2. Meng Z (2007) Application development of SQL server 2005. TsingHua University, Beijing (in Chinese)
3. Software+Services. <http://msdn.microsoft.com/en-us/library/bb977472.aspx>
4. Wen N (2010) The research of GSP medicine management system s security based on SaaS. CNKI:CDMD:2.2010.166049 (in Chinese)
5. Zhao B (2009) The application research of SaaS maturity model level 3 in buy-out model. CNKI:CDMD:2.2009.167654 (in Chinese)
6. Mietzner R, Leymann F, Papazoglou MP (2008) Defining composite configurable SaaS application packages using SCA, variability descriptors and multi-tenancy patterns. 2008 3rd international conference on internet and web applications and services, pp 156–161
7. Byanjankar S, Winoto P, Paik W (2010) In: Proceedings of the 2010 sixth international conference on networked computing and advanced information management (NCM 2010), pp 440–445

Chapter 78

Pulse-Based Analysis on Teaching Quality Evaluation

Aijun Zhang, Yujia Ge and Biwei Li

Abstract Teaching evaluation is essentially a kind of value judgments on teaching activities and their effects, but currently the teaching quality assessment is usually mainly carried out through the assessment made by experts or the university itself. Student's opinion is rarely considered and the evaluation of students has only minor impact, so it can not fully reflect the teaching effects and the final evaluation outcome lacks practical effects. This chapter presents a teaching evaluation method based on pulse information analysis, which takes the learning status of students into consideration to achieve objective and quantitative evaluation on teaching effectiveness. The pulse data of students in the class is measured and recorded to extract pulse parameters, and according to the relationship between learning status and emotion, the emotional model is further established. Through analyzing and processing the pulse characteristics parameters and emotional model, the model can reflect the learning status in real time, which is beneficial to teachers in summarizing teaching experience, identifying problems and improving the teaching effectiveness.

Keywords Teaching evaluation · Pulse parameters · Emotion model · Cognition

A. Zhang (✉) · Y. Ge · B. Li
College of Computer and Information Engineering, Zhejiang GongShang University,
Hangzhou 310018, China
e-mail: zaj@mail.zjgsu.edu.cn

Y. Ge
e-mail: yge@mail.zjgsu.edu.cn

B. Li
e-mail: libiwei@yahoo.com

78.1 Introduction

Since teaching evaluation is essentially a value judgment on teaching activities and their effects. To ensure objectivity, teaching evaluation should be based on some objective criteria through various measurement and related data collection to objectively measure and scientifically determine the teaching activities and their effects. Currently, teaching quality evaluation is mainly carried out through the assessment made by experts or the university itself, which has many deficiencies. First, the evaluation is mainly made through personal evaluation, which is much influenced by various subjective factors, and can not completely reflect the quality of teaching. Secondly, since the evaluation is mainly made by experts, student's opinion is rarely considered and the evaluation of students has only minor impact. Thirdly, the evaluation process provides only the results of the evaluation and can not timely and effectively report the problems of the teaching process, which makes the final results lacking of actual efficiency.

Students are the learning subjects, the learning conduct of students should be the core of the teaching evaluation. Whether students in the classroom are thinking positively or responding negatively, enthusiastically or indifferently, these mental state and emotional status clearly demonstrate the status of student learning and truly reflect the difference between good and bad teaching effects. Though human actions and expressions could be observed to judge the mood, how to record and measure person's emotional and mental state is still a hard nut to crack.

Pulse condition is the reflection of the overall conditions of body's Yin and Yang, and Qi and blood, which could not only reveal the body's pathological information, but also reflect personal emotional status such as joy, anger, sorrow, happiness, thought and so on, and the variation of mental activities.

78.2 Emotion, Pulse and Cognition

78.2.1 *Emotion and Pulse*

Emotion is the human attitude and experience on objective things and the corresponding behavioral responses [1], an important part of human mental activity. Human beings have thoughts and feelings, and the reality always has some specific meaning for humans, one way or another. When body's various sensory perceptions perceive the external stimuli, a series of nerve impulses would be sparked, and these nerve impulses will be transmitted to the cortex center of the brain. According to the degree of satisfaction of personal wishes and needs because of the real objects, different mood will be engendered.

The traditional Chinese medicine considers that emotions are closely related to the functions of five zang-organs. Suwen: primary conversation documented that the functioning and activities of the five zang-organs would produce the normal

spiritual emotions and activities, showing different emotions such as joy, anger, sorrow and happiness. On the other hand, emotions can also affect organ activities. Lingshu: spiritual pivot pointed out that emotional activities of sad, sorrow, worry and anxiety will turn mind into turmoil, and the mind in turmoil will shake zang-organs and further affect the operations of all meridians, causing different revelations of the pulses. Modern medicine thinks that pulse changes are mainly related to cardiovascular function, emotion and blood functions. The emotional changes can cause changes in cardiovascular function, for clinical practice the corresponding changes of blood pressure, heart rate, and pulse can be observed.

The psychological and physiological changes are closely related, since every psychological activity will result in a corresponding psychological pulse, such as the people with psychological laziness will have slow and loose pulse and person with lucid thinking will have the smooth pulse. Lingshu: spiritual pivot put forward the idea that pulse supports spirit, while one of the meanings of spirit in traditional Chinese medicine refers to a person's spiritual awareness activities. Therefore, the concept of pulse supports spirit contains rich information about spirit, consciousness and thinking activities.

In Suwen et al. discussed that all human mental activity and behavior would be reflected in the pulse. No matter under various psychological conditions there is changes because of panic, fear, anger, fatigue, rest or motion, as long as there exists mental activity, the pulse will always change.

Chen initially summarized the relationship between seven sentiments and the pulse, and documented in Three Categories of Pathogenic Factors and Prescriptions that joy results in loose pulse, anger in excited, worry in sluggish, thinking in knotted, sorrow in tense, fear in deep, alarmed in activated”.

In short, human mental activity will be revealed by psychological appearances and corresponding pulse, and pulse information will reflect mental activity and psychological state. Through pulse measurement, current mental activity can be inferred to as one of the seven sentiments. For example, if most students in the classroom are in the thinking state, then it means that the teaching content has attracted and aroused students' learning interests, and encouraged students to think actively and try to solve the problem. If it is in happy state, then there should be an active classroom atmosphere, and students are enthusiastic, but for anxious state, then the classroom atmosphere is dull, students are in low spirits and active participation in the classroom is not high. Various emotional statuses could timely reflect the state of teaching and the performance of student learning in the classroom.

78.2.2 Emotion and Cognition

In the non-intellectual factors that influence learning, emotion has relatively more obvious promoting or inhibiting role in the human cognitive activity. Cognitive processes are always accompanied by cognitive experience, while cognitive experience requires the involvement of emotions and emotion affects the quality

and efficiency of the cognitive processes. Everyone has a variety of emotions, no matter the type of emotion, once specific emotion is generated, it will affect the whole cognitive process, making the whole process infected with emotional coloration. When emotion is positive, the cognitive process is also positive; while negative emotion results in negative cognitive process. Psychologists believe that emotions and feelings are very important factors in the learning process, and are closely linked to cognitive processes. Negative emotions (such as stress, depression, and anger) would narrow the perception and dull thinking, which is detrimental to learning; while positive emotion (such as pleasure and satisfaction) could wide the perception and sharpen thinking so as to promote learning [2]. Brain science research shows that the influence of emotion on human learning can not be underestimated [3, 4]. Many experimental studies have shown that positive emotion can promote the accomplishment of cognitive tasks [5] and boost particularly the cognition flexibility [6–10].

Modern educational psychology research has shown that students' learning process is not only a process of acquiring knowledge, but also a process of discovery, analysis and problem-solving. Since emotion dominates the cognitive effects of learning, teaching process should become a pleasurable and positive emotional experience. Emotion has the characteristics of motivation, so good or bad emotion will affect the interest of students, and further influence the students' learning status and learning behavior. If the knowledge content and teaching methods could arouse learning interests or bring happiness and a sense of accomplishment to the students, then students will focus their attention and interests to the study, thinking positively and exploring, analyzing and solving problems, thus the students would not only be satisfied with the teaching contents. Positive emotions, such as happiness, satisfaction and a sense of achievement, initiatively promote student learning. When in a good emotion, students will become agile thinking, open-minded, cognitive efficiency. Conversely, negative emotion, such as failure, pain, frustration and fear will hinder students from learning and students will consider learning as a burden and an errand with much pain, so they are hard to concentrate and are not willing to think and learn, resulting in low cognition efficiency.

Emotional fluctuations often occur, and pulse measured during class learning could infer the emotional fluctuations. With emotional model and pulse characteristics parameters, the emotion change during the learning process could be used to conjecture the cognition status of students, so a quantitative method could be established to evaluate teaching effectiveness.

78.3 Pulse Characteristics Parameters

Twenty-eight kinds of pulses of traditional Chinese medicine is identified by pulse position, pulse rate, pulse force and pulse presentation. For pulse depth (superficial or deep), the pulse position is different; for pulse latency, the pulse rate is different;

for the pulse size, the pulse shape is different; and for pulse empty or replete, the pulse force is different. The pulse components can be reflected and elaborated by some physical characteristics, such as pulse depth, speed rate, strength, rhythm, morphology, etc. of pulse wave.

The characteristics of clinical pulse could be summarized as pulse position, pulse rate, pulse force, pulse width, pulse length, pulse rhythm, and pulse flow [11]. Since this study mainly considers the pulse characteristics associated with learning and emotion, with no intention to the identify of all the pulses, so only part of the characteristic parameters including pulse position, pulse rate, pulse force and pulse presentation are taken into consideration.

78.3.1 Pulse Position

Pulse position represents the depth of the pulse, and the corresponding clinical pulse characteristics are respectively superficial, medium and deep. When the class ambience is vivid and students are enthusiastic, the willingness to show off results in floating pulse. But when the atmosphere of the classroom is dull and students are not interested in the contents, the reluctance to show off results in deep pulse. Pulse position can be measured by pressure sensor, and the range of pressure P1 is between 25 and 150 g. Based on the characteristics of superficial or deep pulse, the identification parameter of pulse position is set to the pulse pressure P1.

78.3.2 Pulse Rate

Pulse rate is the speed of pulse, and the corresponding clinical pulse characteristics are respectively slow, normal and rapid. The pulse rate of less than 60 times per minute is defined as slow pulse, pulse rate between 60 and 70 times per minute is the normal pulse, and the pulse rate greater than 90 times per minute is rapid pulse. When students in the classroom is actively exploring and thinking, in a hurry to solve the problem, the feelings is tense and the pulse is beating faster, then the pulse would be rapid pulse. If students are not interested in learning, indifferent and mentally lethargic, then the pulse would be slow and scattered. Pulse rate could be measured by the parameters of pulse period C1 and the number of pulse cycles per minute Nm.

78.3.3 Pulse Presentation

Pulse force is the intensity of the pulse, and the corresponding clinical pulse characteristics are respectively feeble, normal and full. Pulse stream represents the

degree of pulse flow, corresponding respectively to the clinical pulse of slippery pulse and sluggish pulse. The unified identification of pulse force and pulse flow could be pulse presentation. For slippery pulse, the rise and fall of the pulse curve are speedy, and the wave is quite significant as if flows are fluent; while for sluggish pulse, the rise and fall of the pulse curve are slow, and the peak of the pulse and the descending branch were small in degrees. When the students are acquiring knowledge and the mood is relaxed and happy, then the blood flow is fluent and the pulse rises and falls significantly. If students do not quite understand the content or have difficulties in the learning process, then the blood stagnates and the pulse rises and falls more sluggish. The 11 parameters to be calculated include [11]: Main Amplitude height hsp (PP'), tidal amplitude height hef (KK'), weighted amplitude height hfg (LL'), starting point of tidal wave hee (EE'), the wave rise time $T1$, pulsation cycle $C1$, the area of pulse image A and the $rep = hee/hsp$, $rfp = hff/hsp$, $res = hef/hsp$ and $rfs = hfg/hsp$.

78.4 Teaching Evaluation Model

78.4.1 Emotional Model

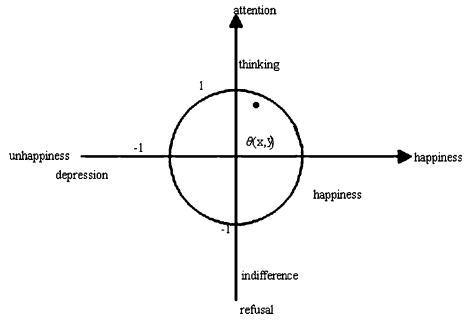
Experimental psychology research shows that if students can well understand and accept the learning content, then they will be in a happy mood, but if they can not understand the learning content or have obstacles in the learning process, or do not like the current teaching methods, or could not keep up with the teaching schedule, then they will be pessimistic and are not interested in learning.

For categorizing emotion, there is no widely acceptable theory. In ancient China, seven emotions of happiness, anger, worry, anxiety, grief, fear, panic was proposed. U.S. psychologist Plutchik proposed eight basic emotions: grief, fear, surprise, acceptance, ecstasy, rage, vigilance. For the teaching evaluation, not every kind of emotion should be analyzed, and only learning-related would be chosen. Student emotions are grouped into four categories: pleasant (understand and accept the learning content), worry (could not understand learning content), thinking (very interested in learning), indifferent (no interests in learning). From student's emotion state, the current cognitive state could be inferred, and classroom teaching quality could be evaluated. If the students are in poor cognitive status, then the situation could be known and analyzed, which would be beneficial to teachers in adjusting the teaching content, schedule or teaching methods in a timely manner.

Two aspects of emotional state and the corresponding pulse changes are selected, namely the degrees of interests and pleasure:

Degrees of interests. When a student is interested in the learning content, he is willing to express himself, ruminate actively and hope to solve the problem quickly, then the spirit is highly concentrated and nerve tension is increased. All these psychological changes would lead to changes of blood, reflected as pulse beats in a

Fig. 78.1 Two-dimensional emotion model



state of superficial, strong and fast. When a student is not interested in the learning content, the attitude of indifference and the unwillingness to learn would make him psychologically reject the study, and then pulse is sunken, feeble and slow.

Degrees of pleasure. When the students understand and accept the learning content, or have solved difficult problems in the learning process, aspirations is realized and needs are met, then a sense of accomplishment and satisfaction would make them in a happy mood, and feelings are relaxed, with no more anxiety and tension. The pulse is deep and scattered, pulse rate is of unsmooth speed and pulse flow is fluctuating. If students do not acquire the learning content, or have encountered obstacles, then they feel disappointed, worried and depressed, the pulse would be feeble and sluggish and the pulse rhythm is irregular with weak pulse presentation.

In order to clearly reflect emotional states of students, a two-dimensional emotional model based on the theory of emotional dimension and two-dimensional emotional space diagram [12] is proposed, as shown in Fig. 78.1. The two dimensions of the emotional model represent happiness-unhappiness on one dimension and attention-refusal on another dimension. All normal emotions can have the corresponding position within a circle of radius of one, which have happiness-unhappiness as the horizontal axis and attention-refusal as the longitudinal axis. The points outside the circle represent abnormal emotion out of control. The origin of coordinates denotes the state of calm. Happiness and depression represent contrary emotions, defined respectively as 1 and -1 on the horizontal axis. Thinking and indifference are directly opposite emotions, represented as 1 and -1 on the longitudinal axis.

Any kind of emotion can be represented by a two-dimensional vector (x, y) . The module of vector r has an intensity range of $[0, 1]$. If a specific vector has the intensity beyond the range, then it means that the feeling is out of control. Vector angle θ represents the tendency degree of the vector to these four basic emotions. Which:

$$r = \sqrt{x^2 + y^2}, \theta = \begin{cases} \arccos(x/\sqrt{x^2 + y^2}), & y > 0 \\ \pi + \arccos(x/\sqrt{x^2 + y^2}), & y < 0 \end{cases} \quad (78.1)$$

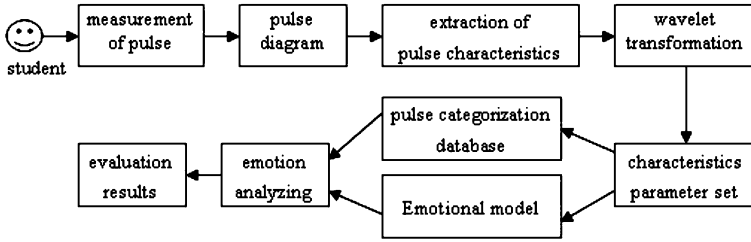


Fig. 78.2 Model of teaching evaluation

78.4.2 Teaching Evaluation Model

Teaching evaluation model is shown in Fig. 78.2. During classroom teaching evaluation, the pulse of students is recorded and measured to extract pulse parameters. Wavelet transformation is adopted to reduce noise, and then the pulse characteristics parameter set is established. According to the analysis based on emotional model and the pulse characteristics parameters set, the type of emotion is further determined and the student learning status is identified, achieving objective and quantitative evaluation of teaching effectiveness.

For example, suppose e is any point in the emotional model, $e = (x, y)$, where x represents the degree of understanding of the learning content, and y represents the level of interests in learning. When x is positive, it indicates that learning is within the state of understanding, and $x = 1$ indicates that students fully understand the learning content. The happy emotion and lively classroom atmosphere indicates effective teaching; when x value is negative, and then it indicates that students do not understand the learning content and is fearful of not being able to keep pace. With students in sad state, teachers need to adjust the learning content or slow down the progress of teaching. When y is positive, indicating that students are interested in active thinking and highly motivated; and when y is of negative value, then indicating that students are not interested in learning, tired and lazy, thus teachers need to change teaching methods to attract the interest and attention of students.

78.5 Conclusions

Students are the primary subject of evaluation, and the results of pulse analysis should not replace all of the teaching quality evaluation. Examinations and tests are also methods available to evaluate teaching quality, but it has time lag, and with different reference objects, the evaluation results will have much differences.

The pulse evaluation is an evaluation that focuses on process, which utilizes the pulse information to reflect the real state of student learning. It is beneficial to discover advanced teaching methods and experiences, and provides experimental evidence for innovating teaching methods and improves teaching for poor learning state. The pulse changes are physiological, so it can avoid the interference of personal factors on evaluation, resulting in authentic, credible, effective and directive evaluation, and achieving the objective quantitative evaluation.

References

1. Dan-Ling P (2001) General psychology[M]. Beijing Normal University Publishing House, Beijing
2. You-zhi W, Ouyang L (2003) Foundation of psychology principle and application [M]. Capital University of Economics and Business Press, Beijing
3. Ding D, Zhang W, Shi M (2004) Modern brain science and education[J]. J Shanxi Norm Univ (Nat Sci Ed) 32(9):189–193
4. Dolan RJ (2002) Emotion, cognition, and behavior[J]. Science 8(298):1191–1194
5. Sui X, Gao S, Wang J (2010) Research progress about the influence of emotion on cognition[J]. J Liaoning Norm Univ (Soc Sci Ed) 33(2):55–58
6. Fredrickson BL (2001) The role of positive emotions in positive psychology. The broaden-and-build theory of positive emotions. Am Psychol Special Issue 56:218–226
7. Hill EL (2004) Evaluating the theory of executive dysfunction in autism. Dev Rev 24:89–233
8. Gesine D (2006) How positive affect modulates cognitive control: the costs and benefits of reduced maintenance capability. Brain Cogn 60:11–19
9. Shan-shan C (2008) An experiment study investigating the modulation of personality and emotion on cognitive flexibility[D]. ShanDong Normal University, Jinan
10. Wang Y (2008) The effects of positive emotions on task switching[J]. Acta Psychol Sinica 40(4):301–306
11. Fei Z (2003) Contemporary sphygmology in traditional Chinese medicine [M]. People's Medical Publishing house, Beijing
12. Meng X, Wang Z, Wang L (2007) Emotional model based teaching assistant system [J]. Appl Res Comput 24(4):74–76

Chapter 79

French Learning Online Platform: French Livehand

Mei Ying Lu, Chun Lai Chai and Yu Shen

Abstract Popularization of the computer made network learning–E-learning more popular. As French learners increased rapidly in recent years, in order to meet those needs, French livehand this French learning online platform was designed. The platform used Struts2, Java, Spring, ExtJS, etc. technique to build. It offered a more simple and quicker approach to learning French. People can add new words to improve vocabulary library and use the test system to test its French level. Besides, people can use this platform to look up words, translation sentences, searching grammar, scene system, and so on.

Keywords French learning · Online platform · ExtJS · Struts2 · Spring

79.1 Introduction

Nowadays, with the rapid development of Internet, everyone all around the world, including children and the disabled, knowing and using computer. And under this background, E-learning brought new study passion. Lots of schools began to use

M. Y. Lu (✉) · C. L. Chai
College of Computer and Information Engineering, Zhejiang Gongshang,
University, Hangzhou 310018, China
e-mail: lmyld@yahoo.cn

C. L. Chai
e-mail: ccl@mail.zjgsu.edu.cn

Y. Shen
College of Foreign Language, Zhejiang Gongshang, University,
Hangzhou 310018, China
e-mail: ailleenshen@163.com

Internet to provide educational services. Such as Yale University, transcribing some courses and uploading to the official website. In that way, learners around the world can study these knowledge without geographical limitation. Many kinds of Internet platforms started to offer some materials, including video, text data, etc.

French is the first written language in UN and six other working languages. It is widely used in International communication and activities. It is not only the official language of France, but also used in more than 40 countries. As a language only superior to English, French has a large reputation in the international society. However, the development of French in China is still lag behind.

For the past few years, French learning in China was becoming more and more common, French learners were increased gradually. So we decided to make full use of the existing computer technology to open up the French learning online platform, to provide a new learning platform to French learners.

79.2 Design and Implementation

Platform using the distributed structure, that is to say web server, database server separation structure. Web server deployed web applications, deploying database in database server, depositing all the relevant data, all these did good to web security management. Among these, database used SQL Server, the languages including: Java, Delphi, HTML, and used ExtJS, Struts2 [1], Spring [2], Hibernate [3] to develop and design system framework. Platform resources included words, vocabulary, some proverbs, idioms and grammar, etc. As platform transmitted through HTTP protocol, users could use the learning platform in any networking machines, and would not be cut off by firewall.

The following is the description of related technologies:

79.2.1 *ExtJS*

ExtJS could be used to develop RIA or rich clients AJAX application [4]. It was written by JavaScript, mainly used to create front user interface. It was a front Ajax frame work. So it could be used in Net, Java, Php, etc. these kinds of development in the application of language development. ExtJS was based on YUI in the beginning and developed by JackSlocum. According to consult JavaSwing mechanism to organization visualization components. No matter from the applications of CSS styles in UI screen, to anomalies processing in data analytical, could it be regard as a rare JavaScript client technology products.

The UI component model and development of Ext were born out of Yahoo component gallery YUI and Swing in Java platform. It helped developer shield a large number of treatments in IE. Relatively, EXT was easier for developer to exploitation.

79.2.2 Struts2

The design idea of WebWork was the core of Struts2 [5]. It used a Filter Dispatcher as the CPU. It helped separating Action and Servlet API.

Simple processes of Struts2:

- (1) The browser sent request;
- (2) According to struts.xml documents, CPU find the corresponding request of Action;
- (3) Interceptor chain of WebWork used request application commonly and automatically, for example: WorkFlow, Validation, etc.;
- (4) If deployed Method parameters in struts, xml documents, then invoking Method ways of corresponding Action, or invoking common Execute way to deal with request;
- (5) Response the results of corresponding methods in Action to IE.

Through the analysis of demand, platform located as a French learning platform, including knowledge, interactive, practicability and interesting, and open to public. The platform not only had the function that general language software had, but had special function such as scene system, grammar system and user management system. Among these, entered users could manage new words, recite words and do exercise (See Fig. 79.1).

79.3 Main Functions

When users input particular URL in the web browser address bar. Users entered homepage first. In here, users could press corresponding column to study.

79.3.1 Intelligent Word-Searching System

System background had more than 300,000 records, including most of French words. Besides, word-searching system also offered autosuggest. Users just need enter one word or letter; it would pop up drop down list. It provided users convenience. Moreover, the system had historical words searching function. Users used previous or next button to search the words users had searched.

79.3.2 Translation System

Users input text which need to translation, the system would base on different sentence to operate in backend database. If the sentence had added in database, the system would offer the corresponding result, or invoking Google API.

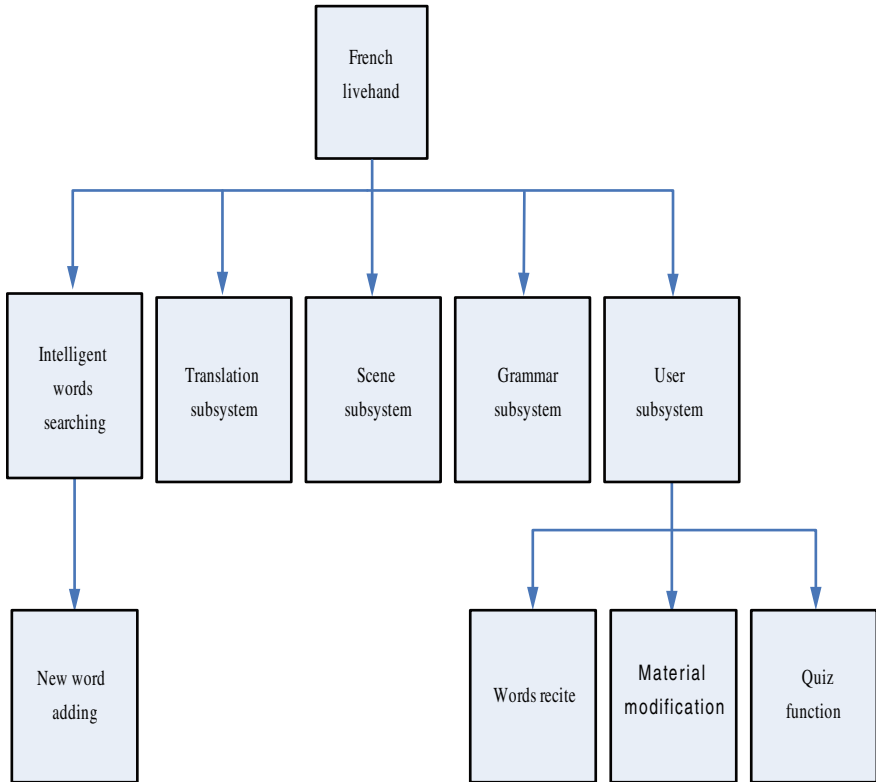


Fig. 79.1 Structure of system function

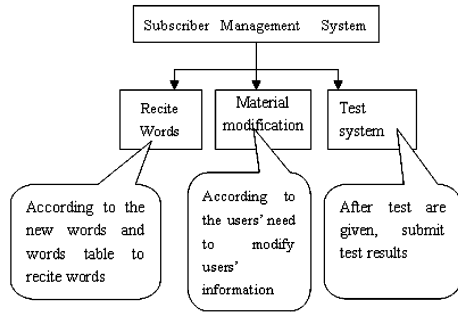
79.3.3 Scene System

Learning French could not just stay in remembering words and grammar application. It is very important to use different sentence in different occasions. Aiming at this, the platform developed “scene system”. For instance, users did not understand or master “self-introduction” very well, he could click self-introduction in scene system, he would get sentence and words that could use in that occasion.

79.3.4 Grammar System

During the progress of language learning, grammar was very important. The grammar system offered grammar help to learners. Users could click grammar

Fig. 79.2 User management function diagram



system to learn corresponding grammar. Such as French words had negative and positive, users could click this system and input question.

79.3.5 User Management System

In this system, users could base personal needs alter personal information. It still could manage new words, recite words and do exercise.

- (1) Users could add new words and perfect lexicon. After landing, users could modify;
- (2) Click “recite words” button, the system would give words, letting users recite;
- (3) Users could enter the system to have a test at any time to view their French level. After finishing the text, the system would prompt correct answer. Test system could serve users to resize the learning methods and formulate a new plan (Fig. 79.2).

79.4 Conclusion

Although the platform was a platform under study and perfect continuously, the system had showed the advantage of language learning.

The platform had three characteristics:

- (1) Based on the open Google API, put Chinese traditional culture, such as proverb, idioms, etc. to the translation system. To make the translation more accurate and suit for Chinese or French culture.
- (2) In the new words management, users could alter or perfect the words he added; through recite words, users could enlarge vocabulary; through Internet, users could do exercise at any time.

(3) Another feature was scene system and grammar system. Platform perfected different kinds of scene and grammar.

This chapter described a creative development of French online learning platform. And analysis its features roundedly. Platform aimed to help users to learn and use French better, promoting their French study. Platform still needed further development now, in the following time, we would perfect it further.

References

1. Donald B, Chad MD, Stanlick S (2008) Struts2 in action. Manning Publications
2. Craig W (2011) Spring in action. Manning Publications
3. Christian B, Gavin K (2004) Hibernate in action. Manning Publications
4. Wang S, He Y, Liu Z (2009) Personalized web based english learning system using artificial neural networks [J]. ICCSE 1263–1268
5. Sun X, Zhao F, Luo X (2009) Application and education of computer for special people [J]. ICCSE 1400–1403

Chapter 80

Application of Text Data Mining to Education in Long-Distance

Jianjie Song and Hean Liu

Abstract Based on Web technology and Text Data Mining, focusing on Personalized service of Long-distance Education, this chapter aims to apply Content Mining Algorithm into Personalized recommendation of learning content, use the simulation results to verify the effectiveness of the fusion algorithm, and apply the tested algorithm to the construction of higher vocational teaching website.

Keywords Text data mining · Long-distance education · Content mining algorithm

80.1 Introduction

If a physical link is recommended from two aspects: novelty and information. Being away from the website current user is visiting should be a priority target. The physical link path length is determined by the topology of a directed graph. Each node of directed graph represents a site in the corresponding page URL [1]. If there exists a physical link from page X to page Y, there is a directed edge from corresponding node X to node Y there exists. The path distance of two web sites (i.e., u_1 and u_2) with physical link is defined as: a directed graph on the site, from u_1 to u_2 the minimum access path length.

J. Song
Hunan vocational college of science and Technology, Changsha 410004,
Hunan, China

H. Liu (✉)
Hunan City University, Yiyang 413000, Hunan, China
e-mail: liuheanlaoshi@sina.com

Assuming sliding window size W is 3 [2], the operating sequence of current sliding window visitors is $W = \langle A, B, C \rangle$, according to W and $|W|$. When we visit Pathset sequence database, firstly we focus on the search four the top three as A, B, C sequence, and put the last figure of sequence meeting the requirements into the recommended set. If the element in recommended set is greater than 1, such as the recommended including the following elements of $\{D, E, F, M\}$, it is advisable to choose the longest distance as the recommended webpage while concentration of greater than 1 if the recommended concentration of elements, then the study page C links to recommended concentration of the physical path distance, choose from the largest of the recommended page. Assuming M meets the requirements page, M is recommended as the next visiting page. When a user visited the M , the user access operations sequence in new sliding window changes into a $\langle B, C, M \rangle$, then completing a recommended operation. If the four items do not meet the requirements of the sequence of focus, then search the three items begin with B, C sequence, and the last figure satisfying the requirement is added to recommended set. The other operation is the same the same as above. If the concentration does not meet the sequences requirements, then search the two items set begin with C until you find series satisfying the requirements, and other operations is the same as above procedure.

80.2 Vector Space Model

The basic concept of vector space model is as follows [3]:

- (1) Documentation: refers to an article or a part of the text or fragment.
- (2) Feature items: the contents of any document or fragment to be simple as it contains the basic morpheme units (characters, words, phrases, or phrases, etc.) posed by the collection, these basic characteristics of language units are collectively referred to as, namely the set of documents with term list can be expressed as $D = (t_1, t_2, \dots, t_i, \dots, t_n)$, where it is the first i -feature items, $1 \leq i \leq n$.
- (3) Characteristics of the weights of items: one item for the D -document containing the n ($t_1, t_2, \dots, t_i, \dots, t_n$), t_i feature items are often given a certain weight w_i , and they were importance in the document, namely: $D = (w_1, w_2, w_i, \dots, w_n)$. Similarly, the user information requirement can also be expressed with the vector form.
- (4) The vector space model: given a document $D = (\langle t_1, w_1 \rangle, \langle t_2, w_2 \rangle, \dots, \langle t_i, w_i \rangle, \langle t_n, w_n \rangle)$, because t_i is repeated in the document and must also have priorities and the relationship have some difficulty. In order to simplify the analysis, it may temporarily not be considered in order of t_i in the document and requested the t_i can not be repeated. Then the t_1, t_2, \dots, t_n can be regarded as an n -dimensional coordinate system, $w_1, w_2, w_i, \dots, w_n$ as the corresponding coordinate value, the $D = (w_1, w_2, w_i, \dots, w_n)$ can be seen as n -dimensional space

(feature items document space, i.e. TD space) in a vector, we call $(w_1, w_2, w_i, \dots, w_n)$, namely the document vector D .

- (5) Similarity: it is used to measure the related degree between documents or between the user's information needs (content). This method use the similarity information retrieval or information filtering, you must first be able to document the characteristics of individual items weighted and operation of collection, and then calculated the transmission document vector space with the information needs of users of the similarity between vectors, and finally provides users with a set of documents in descending order by similarity list.

80.3 The Recommendation of the Corresponding Web Page Hyperlink Method

To achieve the navigation, you must understand the interest of objects so as to targeted. First of all, it should be recommended hyperlink consistent with user interests; Second, it must a hyperlink on a Web page and user interest in the match. We use the vector space model to achieve this match. Application in the text content analysis, we give the representation of the characteristics of web pages, the page p can be expressed as a k -dimensional vector, which feature items that the page p in f_i weight. Because each corresponds to a Web page hyperlink, the hyperlink to each corresponding to a k -dimensional vector p . Suppose that a user browsed the web in the current m -hyperlinks, we interested users to calculate the k -vector and the m -dimensional vector of similarity, that is, trials, set a threshold α , according to the set threshold Value [4], we have the first three with the largest similarity to users interested in hyperlink recommendation to the user, completing the recommended action.

80.4 The Recommended Method Based on Cooperation

Based on cooperation, the recommended approach is also known as collaborative filtering, a person's interest is not isolated, it is a relative concern in the interest of a group. Under normal circumstances the information received by people around the crowds is a particular result. Based on the above factors, we can group similar information to evaluate through their recommendations to other groups of users. Under normal circumstances, the use of groups of people can be divided into two types of active and passive, active people can make full use of the initiative to provide feedback, which feedback will be applied to filter non-active population, its drawback is that information resources must be considered characteristics and can not find information of interest to the user, when the system uses the early, less education information resources, the use of proximity between objects is not easy to be investigated through the evaluation.

Information customization module, information needs analysis module, similar to the matching module together form the system based on collaborative filtering. Custom modules and information needs of information analysis module of these two methods and we said, before, like content filtering technology service system, so without in-depth analysis here, mainly for the third similarity matching module to expand the analysis. This module is the first to use clustering methods, by contrast it uses a user profile object clustering, you must first use the clustering method used. Clustering can then be mapped to the user profile concept hierarchy in the multidimensional space vector form several separate feature vector, and then calculate the distance method of vector space or vector space model approach to calculate the similarity between user profiles degree, you can arrive at a similar target group. Similar to the target group can also be mentioned in the article, we apply to user groups and the clustering algorithm to obtain. According to different rating and scoring documents, this target group users were the results of the situation to get a list of recommendations that can then be recommended list and information resources. In the past, analyzing the matching calculation, the user will receive the recommended information. Recommended based on content and cooperation the two methods, we can put him in combination. First, content-based methods have the user interest model, this model shows that the contents of each user for the level of interest, similar to the interested users of its feature vectors, so the user based on feature vectors are classified by content, known content class. However, to be able to recommend to the user information of interest, we must first consider the unity of the user evaluation will be divided into two category, known as co-class. Purpose of doing hope to use the evaluation to use objects and not within the given recommendation. Two effects must be considered in an integrated similarity to the user information corresponding to recommend. According to the evaluation of the user, dynamically adjust the user types and the adaptation of various parameters in order to improve the recommendation accuracy.

Denote by μ the overall average rating [5, 6]. A baseline estimate for an unknown rating r_{ui} is denoted by b_{ui} and accounts for the user and item effects:

$$b_{ui} = \mu + b_u + b_i \quad (80.1)$$

The parameters b_u and b_i indicate the observed deviations of user u and item i , respectively, from the average.

In order to estimate b_u and b_i one can solve the least squares problem:

$$\min_{b^*} \sum_{(u,i) \in \kappa} (r_{ui} - \mu - b_u - b_i)^2 + \lambda_1 \left(\sum_u b_u^2 + \sum_i b_i^2 \right) \quad (80.2)$$

Here, the first term $\sum_{(u,i) \in \kappa} (r_{ui} - \mu - b_u - b_i)^2$ strives to find b_u 's and b_i 's that fit the given ratings. The regularizing term— $\lambda_1 (\sum_u b_u^2 + \sum_i b_i^2)$ —avoids overfitting by penalizing the magnitudes of the parameters.

An easier, yet somewhat less accurate way to estimate the parameters is by decoupling the calculation of the b_i 's from the calculation of the b_u 's. First, for each item i we set:

$$b_i = \frac{\sum u : (u, i) \in \kappa(r_{ui} - \mu)}{\lambda_2 + |u|(u, i) \in \kappa} \quad (80.3)$$

Then, for each user u we set:

$$b_u = \frac{\sum i : (u, i) \in \kappa(r_{ui} - \mu - b_i)}{\lambda_3 + |i|(u, i) \in \kappa} \quad (80.4)$$

Averages are shrunk towards zero by using the regularization parameters, $\lambda_2 \cdot \lambda_3$, which are determined by cross validation. Typical values on the Netflix dataset are: $\lambda_2 = 25$, $\lambda_3 = 10$.

Central to most item-item approaches is a similarity measure between items. Frequently, it is based on the Pearson correlation coefficient ρ_{ij} , which measures the tendency of users to rate items i and j similarly. Since many ratings are unknown, it is expected that some items share only a handful of common raters. Computation of the correlation coefficient is based only on the common user support. Accordingly, similarities based on a greater user support are more reliable. An appropriate similarity measure, denoted by s_{ij} , would be a shrunk correlation coefficient:

$$s_{ij} \stackrel{\text{def}}{=} \frac{n_{ij}}{n_{ij} + \lambda_4} \rho_{ij} \quad (80.5)$$

The variable n_{ij} denotes the number of users that rated both i and j . A typical value for λ_4 is 100. Notice that the literature suggests additional alternatives for a similarity measure.

This set of k neighbors is denoted by $S^k(i; u)$. The predicted value of r_{ui} is taken as a weighted average of the ratings of neighboring items, while adjusting for user and item effects through the baseline estimates:

$$\begin{aligned} \hat{r}_{ui} &= b_{ui} + \frac{\sum j \in S^k(i; u) s_{ij} (r_{uj} - b_{uj})}{\sum j \in S^k(i; u) s_{ij}} \\ &= b_{ui} + \sum_{j \in S^k(i; u)} \theta_{ij}^u (r_{uj} - b_{uj}) \\ &= \mu + b_u + b_i + |R(u)|^{-1/2} \sum_{j \in R(u)} (r_{uj} - b_{uj}) q_i^T x_i + |N(u)|^{-1/2} \sum_{j \in N(u)} q_i^T y_j \\ &= \mu + b_u + b_i + q_i^T \left(|R(u)|^{-1/2} \sum_{j \in R(u)} (r_{uj} - b_{uj}) x_j + |N(u)|^{-1/2} \sum_{j \in N(u)} y_j \right) \end{aligned} \quad (80.6)$$

Model parameters are learnt by gradient descent optimization of the associated squared error function. Our experiments with the Netflix data show that prediction accuracy is indeed better than that of each individual model. For example, with 100 factors the obtained RMSE is 0.8966, while with 200 factors the obtained RMSE is 0.8953.

80.5 Conclusion

According to tests, we may draw the conclusion that we can make improvements in web site design by this algorithm. The main measures are as followings:

We can take the optimization of the WEB site linkage structure into account from two respects. First, WEB log files can be realized for the users to become more convenient to use the resources of the website, and to strengthen the close link between pages by adapting the relevance of them. Second, if the location that a user links actually is lower than expected location, we can apply through deeper web log files to find the application and we can optimize web site pages through the establishment navigation between practical and expected users.

We can improve the site by modifying some property of pages. These methods may include the following three aspects. First of all, the probability of any hypertext links being selected in a page, depend on the number of hypertext links there contain in a page. And if a web page the page contains a lot of hypertext links, the relative probability of the link being selected will be reduced. Secondly, compared with those hypertext links behind, those before them will be easily selected. Therefore, the position is proved to be very important. Besides, under equal conditions, the regional size is also another important factor of being selected. Lastly, another factor of being selected is related to the clearness of contents and the availability of recognition of the words in a hypertext. If they convey a clear and clean meaning between the words in a hypertext and the link page, then the probability of being selected will be larger.

References

1. Huang XJ, Croft WB (2009) A unified relevance model for opinion retrieval[C]. In: Proceeding of the 18th ACM conference on information and knowledge management, HongKong, ACM, pp 947–956
2. Kim SM, Hovy E (2010) Determining the sentiment of opinions[C]. In: Proceedings COLING-04, Geneva, Association for Computational Linguistics, 1, pp 267–1376
3. Bo P, Lillian L, Shivakumar V (2009) Thumbs upon sentiment classification using machine learning techniques, presented at the 2002 conference on empirical methods in natural language processing (EMNLP'2009), pp 79–86
4. Gelan Y, Xue X, Gang Y, Jianming Z (2010) Semi-supervised classification by local coordination lecture notes in computer science, vol 6444, Neural information processing. Models and applications, pp 517–524

5. Gelan Y, Xue X, Gang Y, Jianming Z (2010) Research of local approximation in semi-supervised manifold learning. *J Inf Comput Sci* 7(13):2681–2688
6. Borges J, Levene M (1999) Data mining of user navigation patterns. In: *Proceedings of the workshop on web usage analysis and user profiling (WEBKDD'99)*

Part VIII
Fuzzy System and Control

Chapter 81

Adaptive Disturbance Rejection Control of Linear Time Varying System

Dangjun Zhao, Zheng Wang, Yongji Wang and Weibing Hu

Abstract A novel adaptive disturbance rejection control scheme for a linear time varying (LTV) system from the perspective of differential algebraic framework is proposed. A numerical differentiator is used to obtain the derivative estimates from the system output, which contain overall dynamics of the system. Combining a local modeling technique and conventional proportional integral differential controller, the proposed control scheme perfectly accommodates disturbances and measurement noises. The convincing simulations validate the proposed control scheme well.

Keywords Linear time varying system · Numerical differentiator · Adaptive control · Disturbance rejection

81.1 Introduction

The case of LTV systems is important since one or some parameters of the real physical systems are time varying. Further more, the control of nonlinear system is, usually, accomplished by linearizing this system around a given trajectory

D. Zhao (✉) · Y. Wang
Department of Control Science and Engineering, Huazhong University
of Science and Technology, Wuhan, China
e-mail: zhao.abe@gmail.com

D. Zhao · W. Hu
School of Electrical and Information, Wuhan Institute of Technology,
Wuhan, China

Z. Wang
College of Electrical and Information Engineering, Naval University
of Engineering, Wuhan, China

which renders an LTV system [1]. Researchers have made a great number of contributions on linear time varying (LTV) systems [2] from 1960s'. In the literature [3] and its related literatures, a number of adaptive control schemes for LTV systems has been proposed. Most of these adaptive control schemes stemmed from the mature control theories of linear time invariant (LTI) systems, and have been used in engineering successfully. In this chapter, we propose a new adaptive control law for the LTV system via a differential algebraic observer, which is constructed by a new differentiator. The closed-loop error dynamics are the nature of LTI, and all signals in the closed-loop system are uniformly ultimately bounded (UUB). The main advantage of the proposed method lies in the excellent performance in the presence of disturbances and measurement noise.

81.2 Preliminary

We briefly present the method of numerical differentiation, which is proposed by Fliess and Mboup. Further information can be found in [4-7]. Consider an analytical real-valued signal $x(t)$, which has a truncated Taylor expansion $x_N(t) = \sum_{k=1}^N c_k t^k / k!$ at $t = 0$ without loss of generality. The expansion satisfies $d^{N+1}x_N(t)/dt^{N+1} = 0$, which is transformed into s domain, we therefore obtain

$$s^{N+1}x_N(s) = s^N x_N(0) + s^{N-1}x^{(1)}(0) + \dots + x^{(n)}(0) \tag{81.1}$$

Multiply both sides of Eq. 81.1 by operator $\Pi_k^{N,n} = \frac{s^{n+k}}{ds^{n+k}} \frac{1}{s} \frac{d^{N-n}}{ds^{N-n}}$ a direct estimation of $x^{(n)}(0)$ can be acquired as

$$x_N^{(n)}(0) = s^{v+n+k+1} \frac{(-1)^{n+k}}{(n+k)!(N-n)!} \frac{1}{s^v} \prod_k^{N,n} (s^{N+1}y(s)) \tag{81.2}$$

where $v = N + 1 + \mu, \mu \geq 0$. Let $N = n$, and read Eq. 81.2 in time domain thereby [7]

$$\tilde{x}^{(n)}(0) = x_N^{(n)}(0) = \frac{\gamma(n, k, \mu)}{(-T)^n} \int_0^t \frac{d^n}{d\tau^n} \{ \tau^{k+n} (1 - \tau)^{\mu+n} \} y(\tau) d\tau \tag{81.3}$$

where $\gamma(n, k, \mu) = (\mu + k + 2n + 1)! / [(\mu + n)!(k + n)!]$. The boundedness of the derivative estimate above is demonstrated in the following lemma.

Lemma 1 For $0 < t < \varepsilon$, by using Eq. 81.3 the estimate error $\|e_{x^{(n)}}\| = \|\tilde{x}^{(n)}(0) - x^{(n)}(0)\| < \delta$ with δ is a sufficiently small positive constant.

Proof The estimate error consists of two parts due to truncated error $R_N(t) = O(t^{N+1})$ and measurement noise $n(t)$. For truncated error when $t \rightarrow 0$ or $N \rightarrow +\infty$, the term of $O(t^{N+1})$ becomes negligible. For measurement noise, we have a reasonable assumption that $n(t) \in L_2$ is bounded fluctuated function of t with higher frequency. For $0 < t < \varepsilon$, there exist positive constant δ_{R_N} , δ_N and δ_h such that $\|R_N(t)\| < \delta_{R_N}$ and $\|n(t)\| < \delta_N$, meanwhile, $\|h(t)\| = \left\| \frac{\gamma(n,k,\mu)}{(-T)^\mu} \frac{d^\mu}{dt^\mu} \{t^{k+n}(1-t)^{\mu+n}\} \right\| \leq \delta_h$. Rewrite Eq. 81.3 as

$$\begin{aligned} x^{(n)}(0) &= \int_0^t h(\tau)[x_N(\tau) + R_N(\tau) + n(\tau)]d\tau \\ &= x_N^{(n)}(0) + \int_0^t h(\tau)R_N(\tau)d\tau + \int_0^t h(\tau)n(\tau)d\tau \end{aligned}$$

Then

$$\begin{aligned} \|e_{x^n}\| &= \|\tilde{x}^{(n)}(0) - x^{(n)}(0)\| \\ &= \left\| \int_0^t h(\tau)R_N(\tau)d\tau + \int_0^t h(\tau)n(\tau)d\tau \right\| \leq \left\| \int_0^t h(\tau)R_N(\tau)d\tau \right\| \\ &\quad + \left\| \int_0^t h(\tau)n(\tau)d\tau \right\| \leq \delta_h(\delta_{R_N} + \delta_N)t \\ &= \delta \end{aligned}$$

with $\delta > 0$.

Remark 1 Derivative estimation given by Eq. 81.3 are not of asymptotic nature [5]. One hand, from the proof above, there has $\delta \rightarrow 0$ when $t \rightarrow 0$, thus, as long as t is small enough, δ will be a sufficiently small positive constant. On the other hand, the differentiator functions as a low-pass filter, which will attenuate those fast fluctuated noises. However, the performance of noise rejection will degrade when $t \rightarrow 0$. Thus the choice of time window t is a compromise result.

Remark 2 The derivative estimate at time 0 is obtained from Eq. 81.3 is based on the observation of $y(t)$ on the time interval $I_{0+}^t = [0, t]$, and this is not causal. In order to obtain a causal estimate, we replace $y(\tau)$ by $-y(t - \tau)$ in Eq. 81.3 henceforth a causal estimate $\tilde{x}^{(n)}(t)$ based on the observation on the time interval $I_{t-}^t = [0, t]$. We can simply move the estimate from t to any $T \geq 0$ by a heaviside function [7].

81.3 Main Results

Problem Statement. Consider the tracking control of a uniformly controllable and uniformly observable single input single output (SISO) LTV system, which is n th order system and characterized by

$$\left\{ \begin{aligned} \begin{bmatrix} \dot{x}_1 \\ \vdots \\ \dot{x}_{n-1} \\ \dot{x}_n \end{bmatrix} &= \begin{bmatrix} 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \\ a_0(t) & a_1(t) & \dots & a_{n-1}(t) \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ b \end{bmatrix} u + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ d(t) \end{bmatrix} \triangleq \begin{cases} \dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{B}u + \mathbf{D}(t) \\ y = \mathbf{C}\mathbf{x} + n \end{cases} \\ y = x_1 + n \end{aligned} \right. \tag{81.4}$$

where $\mathbf{x} \in R^n, y \in R, u \in R; \mathbf{A}(t)$ is a time dependent matrix with corresponding dimensions, $B \in R^n, C \in R^n; d$ is the external disturbances, w_2 is the measurement noise. The definitions of uniform controllability and uniform observability for LTV system can be found in [2]. For the convenience of analysis we have are following assumptions

Assumption 1 $\mathbf{A}(t)$ and $\mathbf{D}(t)$ are continuous and uniformly bounded such that $\|\mathbf{A}(t)\| \leq M_A$ with $M_A > 0$ and $\|\mathbf{D}(t)\| \leq M_D$ with $M_D > 0$ for all $t > 0$.

Assumption 2 $\|\dot{\mathbf{A}}(t)\| \leq \delta_A$ with $\delta_A > 0$ and $\|\dot{\mathbf{D}}(t)\| \leq \delta_D$ with $\delta_D > 0$ for all $t > 0$.

It is to note that $\|\bullet\|$ is defined as a spectral norm of a matrix in here and the following paper.

Local Modeling. We rewrite Eq. 81.4 as an equivalent form

$$\left\{ \begin{aligned} x^{(n)} &= \sum_{i=0}^{n-1} a_i(t)x^{(i)} + bu + d = f(\mathbf{x}; t) + bu + d \\ y &= x + n \end{aligned} \right. \tag{81.5}$$

where $\mathbf{x} = [x \ \dot{x} \ \dots \ x^{(n-1)}]^T = [x_1 \ x_2 \ \dots \ x_n]$. Let $F(t) = f(\mathbf{x}; t) + d$, then we have $F(t) = x^{(n)} - bu$. Thank to the sampling technique, we can model F at a time instant k as $F_k = x_k^{(n)} - bu_k$ to avoid the algebraic loop, where $(\bullet)_k$ stands for the value of (\bullet) at time instant k . By using Eq. 81.3 the n th order derivative of x can be obtained from the observation of output y . Thus, the local model of F can be written as

$$\tilde{F}_k = \tilde{y}_k^{(n)} - bu_{k-1} \tag{81.6}$$

Since \tilde{y}_k^n can be estimated well even in noisy environment, \tilde{F}_k consists of the overall dynamics of the system at time instant k , including the external disturbance d .

Control Law. For the tracking problem of Eq. 81.4, let the desired trajectory of the output y_d be smooth enough and differentiable, then $\mathbf{x}_d = [y_d \ \dot{y}_d \ \dots \ y_d^{(n-1)}]$. The output y and its finite order derivatives can be estimated by Eq. 81.3, then we have $\tilde{\mathbf{x}} = [\tilde{y} \ \dot{\tilde{y}} \ \dots \ \tilde{y}^{(n-1)}]$. On the basis of the local model, we propose our disturbance rejection controller as the following

$$u = \frac{1}{b} \left[-\tilde{F}_k + x_d^{(n)} - \mathbf{k}^T (\tilde{\mathbf{x}} - \mathbf{x}_d) \right] \tag{81.7}$$

where \tilde{F}_k is given by Eq. 81.6, $\mathbf{k} \in R^n$ is the vector of designed parameters, and $x_d^{(n)} = y_d^{(n)}$. The stability of our control scheme will be discussed in the following subsection.

Stability. The following stability theorem regarding to the control law defined by Eq. 81.7 is stated.

Theorem 1 Consider the system governed by Eq. 81.4 and consider assumptions 1 and 2 are satisfied. If the control law is provided by Eq. 81.7 and the differential algebraic observer is given by Eq. 81.6, then all signals in the closed-loop system are UUB.

Proof Substitute Eq. 81.7 into Eq. 81.4 resulting in the closed-loop system $x^{(n)} - x_d^{(n)} = F(\mathbf{x}; t) - \tilde{F}_k - \mathbf{k}^T (\tilde{\mathbf{x}} - \mathbf{x}_d)$. Let $\mathbf{e} = \mathbf{x} - \mathbf{x}_d$ be the tracking error, $\mathbf{\varepsilon}_x = \tilde{\mathbf{x}} - \mathbf{x}$ the estimate error of state, and $\varepsilon_F = \tilde{F}_k - F(t)$ the model error of our local model. Rewrite the closed-loop equation as

$$\dot{\mathbf{e}} = \mathbf{\Lambda}_c \mathbf{e} + \mathbf{b}_c [\varepsilon_F + \mathbf{k}^T \mathbf{\varepsilon}_x] \tag{81.8}$$

where $\mathbf{\Lambda}_c = \begin{bmatrix} 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \\ -k_n & -k_{n-1} & \dots & -k_1 \end{bmatrix}$, $\mathbf{b}_c = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$.

Let $\varepsilon = \varepsilon_F + \mathbf{k}^T \mathbf{\varepsilon}_x$. We first prove ε is bounded. From Lemma 1, there exists a $\check{\xi}_{y^{(n)}} > 0$ which make $\|\tilde{y}^{(n)} - y^{(n)}\| < \check{\xi}_{y^{(n)}}$. Then the estimate error of x satisfies $\|\varepsilon_x\| = \left\| \begin{bmatrix} \check{\xi}_y & \check{\xi}_{\dot{y}} & \dots & \check{\xi}_{y^{(n-1)}} \end{bmatrix}^T \right\| < M_{\varepsilon_x}$. Similarly, we have $\|\mathbf{\varepsilon}_{\dot{x}}\| = \|\tilde{\dot{x}} - \dot{x}\| < M_{\varepsilon_{\dot{x}}}$. Here M_{ε_x} and $M_{\varepsilon_{\dot{x}}}$ are small positive constants, respectively.

From the assumption 1 and 2, $F(t)$ is continuous and bounded, meanwhile its derivative respect to time is bounded, thus $\forall t, \forall k, \exists t_2$ such that $\|F(t) - F(k)\| = \|\dot{F}(t_2)(t - k)\| \leq \delta_{\dot{F}} \|t - k\|$. Consequently,

$$\begin{aligned} \|\varepsilon_F\| &= \|\tilde{F}_k - F(t)\| = \|\tilde{F}_k - (F_k + \dot{F}(t_2)(t - k))\| \leq \|\tilde{F}_k - F_k\| + \|\dot{F}(t_2)(t - k)\| \\ &\leq \|\tilde{\dot{x}}_k - bu_{k-1} - (\dot{x} - bu_k)\| + \delta_{\dot{F}} \|t - k\| \leq \|\tilde{\dot{x}}_k - \dot{x}\| + \|b(u_k - u_{k-1})\| \\ &\quad + \delta_{\dot{F}} \|t - k\| \leq M_{\varepsilon_{\dot{x}}} + \delta_u + \delta_A T_s + \delta_{w_1} T_s = M_F \end{aligned} \tag{81.9}$$

Here we assume $\|u_k - u_{k-1}\| \leq \delta_u$ with $\delta_u > 0$, which is reasonable for the system. We henceforth have $\|\varepsilon_F + \mathbf{k}^T \mathbf{e}_x\| \leq \|-K e_x\| + \|e_F\| \leq kM_{e_x} + M_F \leq M$ with $M > 0$, i.e., the term $\varepsilon = \varepsilon_F + \mathbf{k}^T \mathbf{e}_x$ is bounded. For the closed-loop system (81.8), the completely solution is

$$\mathbf{e}(t) = \exp(-\Lambda_c t) \mathbf{e}(t_0) + \int_{t_0}^t \exp[-\Lambda_c(t - \tau)] \mathbf{b}_c \varepsilon(\tau) d\tau \quad (81.10)$$

If $\text{Re}(\lambda_i) > 0$, λ_i stand for the eigenvalues of the matrix Λ_c , there exist finite positive constant ρ (in fact $\rho = \max\|\lambda_i\|$) such that the transition matrices $\|\exp(-\Lambda_c t)\| \leq \exp(-\rho t)$.

Hence, the solution of the closed-loop error dynamics satisfies

$$\begin{aligned} \|\mathbf{e}(t)\| &= \left\| \exp(-\Lambda_c t) \mathbf{e}(t_0) + \int_{t_0}^t \exp[-\Lambda_c(t - \tau)] \mathbf{b}_c \varepsilon(\tau) d\tau \right\| \\ &\leq \|\exp(-\Lambda_c t)\| \|\mathbf{e}(t_0)\| + M \left\| \int_{t_0}^t \exp[-\Lambda_c(t - \tau)] d\tau \right\| \leq \exp(-\rho t) \|\mathbf{e}(t_0)\| \\ &\quad + M \int_{t_0}^t \exp(-\rho t) d\tau \leq \exp(-\rho t) \|\mathbf{e}(t_0)\| + [M \exp(-\rho t_0) \\ &\quad - M \exp(-\rho t)] / \rho \leq \|\mathbf{e}(t_0)\| + M \exp(-\rho t_0) / \rho < \infty \end{aligned} \quad (81.11)$$

According to the definition of UUB, the error \mathbf{e} will eventually converge into a hyper ball including the origin, and it is UUB.

81.4 Illustrative Example

In order to validate our control scheme, a simple example is presented here. Consider an LTV SISO system with standard form of observability

$$\begin{cases} \dot{x} = \begin{bmatrix} 0 & 1 \\ -(1 + 0.5 \cos t) & -(1 - 0.5 \sin t) \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u + \begin{bmatrix} 0 \\ \text{sign}(\sin 0.5t) \end{bmatrix} \\ \dot{y} = [1 \quad 0] x + w \end{cases} \quad (81.12)$$

where $w = N(0, 0.005)$ is the measurement noise. According to the output Eq. 81.9, there have $\tilde{x} = [\tilde{y} \quad \tilde{\dot{y}}]^T$ and $\tilde{\dot{x}} = [\tilde{\dot{y}} \quad \tilde{\ddot{y}}]^T$. By using the control law of Eq. 81.7, we design the controller as the following form

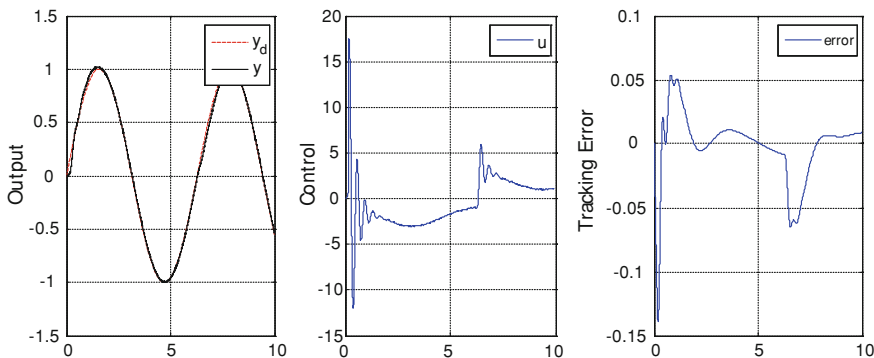


Fig. 81.1 History of output, control and tracking error

$$u = \ddot{y}_d - (\tilde{\ddot{y}} - u_{k-1}) - k_1(\tilde{\dot{y}} - \dot{y}_d) - k_0(\tilde{y} - y_d) \tag{81.13}$$

where $y_d = \sin t$ is the desired output trajectory. The design parameters k_1 and k_0 are chosen so as to render the closed-loop characteristic polynomial into a Hurwitz polynomial with desirable roots. In such case, the desired closed-loop equation is set as $s^2 + 2\zeta\omega_n s + \omega_n^2 = 0$, consequently $k_1 = 2\zeta\omega_n$ and $k_0 = \omega_n^2$.

According to Eq. 81.3 and choosing $N = n, k = 2$ and $\mu = 2$, the estimation of y, \dot{y} and \ddot{y} can be obtained by $\tilde{y} = 30 \int_0^1 p_0(\tau)y(T - T\tau)d\tau$, and $\tilde{\dot{y}} = -140 \int_0^1 p_1(\tau)y(T - T\tau)/T d\tau$ and $\tilde{\ddot{y}} = 630 \int_0^1 p_2(\tau)y(T - T\tau)/T^2 d\tau$, where polynomial $p_i(t)$ respectively are $p_0(t) = t^2 - 2t^3 + t^4$, $p_1(t) = 3t^2 - 12t^3 + 15t^4 - 6t^5$ and $p_2(t) = 12t^2 - 80t^3 + 180t^4 - 168t^5 + 56t^6$

Simulation experiment is conducted with Matlab, the sampling period $T_s = 0.001$ s. The controller parameters set as $\zeta = 1, \omega_n = 5$. Figure 81.1 reveals the proposed method accommodating the external disturbance and system parameters' variation, even in the noisy environment.

81.5 Conclusion

This Chapter has presented a novel control scheme via a differential algebraic framework. An online numerical differentiation technique was introduced for the derivatives estimate, from which a local model of LTV system was established. By using a PID controller, we obtain a closed-loop dynamics of the tracking error, with the nature of linear time invariant. The numerical simulations validate the proposed control scheme is efficient in the control of LTV system, even in the presence of external disturbances and measurement noises.

References

1. Marinescu B (2010) Output feedback pole placement for linear time-varying systems with application to the control of nonlinear systems. *Automatica* 46:1524–1530
2. Huang R (2007) Output feedback tracking control of nonlinear time-varying systems by trajectory linearization. Russ college of engineering and technology. PhD Thesis, Ohio University
3. Marino R, Tomei P (2003) Adaptive control of linear time-varying systems. *Automatica* 39:651–659
4. Fliess M, Sira-Ram H (2004) Control via state estimations of some nonlinear systems. In: Proceedings 6th IFAC symposium on nonlinear control systems(NOLCOS 2004), Stuttgart,Germany
5. Fliess M, Join CE, Sira-Ramírez H (2008) Non-linear estimation is easy. *IJMIC* 4:1–6
6. Mboup M, Join CE, Fliess M (2007) A revised look at numerical differentiation with an application to nonlinear feedback control. In: 2007 mediterranean conference on control and automation Athens, Greece
7. Mboup M, Join CE, Fliess M (2009) Numerical differentiation with annihilators in noisy environment. *Numer Algorithms* 50(4):1–27

Chapter 82

Nonlinear Predictive Functional Control Based on Support Vector Machine

Zhang Xin-fang, Wang Cheng-li, Su Xian-hua and Wang Qiu-jin

Abstract A novel nonlinear predictive functional control (NPFC) algorithm is proposed based on SVM after simple discussion of support vector machine for regression. A predictive model is established by nonlinear system identification with SVM based on linear kernel function. An explicit control law is obtained through the predictive functional control mechanism with both one base function (step function) and two base functions (step and ramp functions). Simulation results show that predictive functional control based on SVM can efficiently deal with nonlinear system. And the NPFC for nonlinear system has light online computational burden, good reference tracking and efficient disturbance rejection.

Keywords Predictive functional control · Support vector machine · Nonlinear system · Identification and control

82.1 Introduction

Predictive functional control (PFC) was proposed by Richalet and Kuntze at late eighties of last century. PFC remains the advantages of classical model based predictive control, produces a structured manipulated variable, which is different

Z. Xin-fang (✉) · W. Cheng-li · S. Xian-hua
State Nuclear Electric Power Planning Design & Research Institute,
100094 Beijing, China
e-mail: zhangxinfang@snpdri.com

W. Qiu-jin
North China Power Engineering Co., LTD of China Power Engineering
Consulting Group, 100120 Beijinh, China
e-mail: wangqj@ncpe.com.cn

from other predictive controls, and effectively reduces computational complexity. PFC was initially proposed for linear system and needs to establish relatively accurate mathematical model. But, most real industrial processes are dynamic and complex. If the process has high nonlinearity, due to the complexity of nonlinear system, it is not possible to develop nonlinear system identification techniques by straightforward applying the linear predictive control theory.

Reference [1] developed nonlinear predictive functional control algorithm based on artificial neural network model and got PFC parameters by adopting Quasi—Newton method. But, control law by using Quasi—Newton method possibly falls in local extremum [2]. Till now, the artificial neural networks have some deficiencies in modeling process, such as structure selection and the local optimal etc. Support vector machine (SVM) is now considered as a power tool to do classification and regression of functions. Reference [3] derived the predictive control law by a new stochastic search optimization algorithm [4], in which nonlinear plants are modeled on a support vector machine, however, the optimization algorithm has disadvantages such as slow convergence and complex computation. Reference [5] presented one step model-based predictive control based on SVM.

In this chapter, a novel NPFC algorithm based on support vector regression is proposed. A predictive model is established by nonlinear system identification with SVM based on linear kernel function. An explicit control law is obtained through the predictive functional control mechanism. The paper is organized as follows. Support vector machine for regression is introduced in Sect. 82.2. In Sect. 82.3, the SVM based on linear kernel function is used as predictive model, and NPFC algorithm is deduced. In Sect. 82.4, the simulation results of nonlinear process are presented. Conclusions are presented in Sect. 82.5.

82.2 Support Vector Machine for Regression

The SVM is derived from statistical learning theory. It determines support vectors and weights by minimizing an upper bound of generalization error. The structure of the SVM is described by the sum of the times of weight and kernel functions. A kernel function projects the data into a high dimensional feature space to increase the computational power of the linear machine. Suppose we have given data $(x_k, y_k)_{k=1}^l$, where $x_k \in R^n$ are input data, $y_k \in R$ are output data, l is the number of data. In the paper, SVM is applied for the purpose of regression.

$$f(x) = \omega^T \varphi(x_k) + b \quad \omega \in R^{nh}, b \in R \quad (82.1)$$

Here, nonlinear function $\varphi(\cdot) : R^n \rightarrow R^{nh}$ projects the data into a high dimensional feature space. According to Vapnik's structural risk minimization theory, the function is obtained by minimizing the following regularized risk functional R_{reg} , which is a combination of the model complexity and the empirical risk, for given error bound ε .

$$R_{reg} = \frac{1}{2} \|\omega\|^2 + C \cdot \frac{1}{l} \sum_{i=1}^l |y - f(x)|_\varepsilon \tag{82.2}$$

Here, $\|\omega\|^2$ is a term which characterizes the model complexity, $C > 0$ is a constant determining the trade-off and the insensitive loss function $|y - f(x)|_\varepsilon$ is given by

$$|y - f(x)|_\varepsilon = \max\{0, |y - f(x)| - \varepsilon\} \tag{82.3}$$

The regression function is obtained as follows:

$$\begin{cases} \omega = \sum_{i=1}^{nsv} S_i \varphi(x_i) \\ f(x) = \sum_{i=1}^{nsv} S_i k(x_i, x) + b \end{cases} \tag{82.4}$$

Where, $S_i = \alpha_i - \alpha_i^*$ is support vector, $nsv < l$ is the quantity of support vector, kernel function $k(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$ satisfies the Mercer condition. In the chapter linear kernel function is adopted $k(x_i \cdot x) = x_i \cdot x$, b is obtained by Karush–Kuhn–Tucker (KKT) conditions.

82.3 PFC Based on SVM

PFC has three basic principles of predictive control: internal model, online iterative optimization and error compensation.

82.3.1 Nonlinear Predictive Model Based on SVM

Nonlinear dynamic system can be described as NNARMRX:

$$y(k + 1) = f([y(k), \dots, y(k - n + 1), u(k), \dots, u(k - m + 1)]) \tag{82.5}$$

where, f is unknown nonlinear function.

Predictive model adopts support vector machine forward structure [5], Using the following equation to fit Eq. (82.5).

$$\hat{y}(k + 1) = \hat{f}(I_0) = \hat{f}([y(k), \dots, y(k - N + 1), u(k), \dots, u(k - M + 1)]) \tag{82.6}$$

$I_0 = [y(k), y(k - 1), \dots, y(k - N + 1), u(k), u(k - 1), \dots, u(k - M + 1)]$. The order N, M does not always equals to n, m , mostly depending on fitting accuracy and generalization [6].

Considering the training data $[I_{0,j}, y_j] j = 1 \dots l$, here, $y_j = y(j + 1)$, support vectors I'_0 and $S_i i = 1, \dots, nsv$, b can be obtained by SVM. The predictive model:

$$\hat{y}(k + 1) = \sum_{i=1}^{nsv} S_i(I'_{0,i} \cdot I_{0,k}) + b \tag{82.7}$$

Considering

$$I'_{0,i} \cdot I_{0,k} = I'_{0,1}(0)y(k) + I'_{0,1}(1)y(k - 1) + \dots + I'_{0,1}(N + M)u(k - M + 1) \tag{82.8}$$

Set $I_1 = [y(k), y(k - 1), \dots, y(k - n + 1), 0, u(k - 1), \dots, u(k - M + 1)]$,

$$I'_{0,1} \cdot I_{0,k} = I'_{0,1}(N + 1)u(k) + (I'_{0,1} \cdot I_1) \tag{82.9}$$

$$I'_{0,i} \cdot I_{0,k} = I'_{0,i}(N + 1)u(k) + (I'_{0,i} \cdot I_1) \tag{82.10}$$

$$\hat{y}(k + 1) = \sum_{i=1}^{nsv} S_i I'_{0,i}(N + 1)u(k) + \sum_{i=1}^{nsv} S_i (I'_{0,i} \cdot I_1) + b \tag{82.11}$$

When the prediction horizon $H = 2$, two step predictive output of SVM model:

$$\hat{y}(k + 2) = \hat{f}([\hat{y}(k + 1), y(k), \dots, y(k - N + 2)u(k + 1), u(k), u(k - 1), \dots, u(k - M + 2)]) \tag{82.12}$$

Algorithm of PFC pays great attention to structured manipulated variable to ensure high performance of controlled plant, which is different from other predictive controls. The future manipulated variable is structured as a linear combination of a pre-specified set of functions called base functions. The choice of these functions depends on the nature of the process and the setpoint. Generally a step, ramp, parabola, etc. are used.

$$u(k + j) = \sum_{i=1}^N \mu_i f_i(j) \quad j = 0, \dots, H - 1 \tag{82.13}$$

Where μ_i are linear combination parameters, H is prediction horizon. If one base function, that is, step function is used:

$$u(k + j) = u(k) \quad j = 0, \dots, H - 1 \tag{82.14}$$

Set $I_2 = [0, y(k), \dots, y(k - n + 2), 0, 0, u(k - 1), \dots, u(k - M + 2)]$, According to aforementioned one step predictive, combining with Eqs.(82.14) and Eqs.(82.12) can be deduced into Eqs.(82.15).

$$\begin{aligned}
\hat{y}(k+2) &= C_1 u(k) + C_2 + C_3 \\
C_1 &= \sum_{i=1}^{nsv} S_i (I'_{0,i}(N+1) + I'_{0,i}(N+2)) + \sum_{i=1}^{nsv} \sum_{j=1}^{nsv} S_i I'_{0,i}(1) S_j I'_{0,j}, \\
C_2 &= \sum_{i=1}^{nsv} \sum_{j=1}^{nsv} S_i I'_{0,i}(1) S_j I'_{0,j} I_1 + \sum_{i=1}^{nsv} S_i I'_{0,i} I_2, \\
C_3 &= b \left[1 + \sum_{i=1}^{nsv} S_i I'_{0,i}(1) \right]
\end{aligned} \tag{82.15}$$

Theoretically, we can get $\hat{y}(k+H)$ mathematical expression at any prediction horizon H , however, the expression becomes very complexity with the adding prediction horizon due to the system complexity. And it is difficult to get the unified expression by math induction. Reference [6] points out that control signal sequences to be optimized should be short to the best of it's abilities at every sample time for predictive control of nonlinear system. And a short predictive control sequences is insensitive to system model error and enhances robustness of controller. Thus, the chapter uses two step predictive output of SVM model.

82.3.2 Reference Trajectory

Expected output of control system is determined by reference trajectory. Reference trajectory has many forms, a simple exponential trajectory can be used when the setpoint has step value.

$$y_r(k+H) = \beta^H y(K) + (1 - \beta^H) y_{sp} \tag{82.16}$$

Where, $\beta = \exp(-T_s/T_r)$, T_r is time constant, y_r is reference trajectory, y_{sp} is setpoint, T_s is sample time, y is process output.

82.3.3 Modeling Error Compensation

Either state additive perturbations or structural perturbations (model mismatch) affect the model output which is always different from the process output. Here, the following expression is adopted.

$$y_P(k+i) = \hat{y}(k+i) + h(y(k) - \hat{y}(k)) \quad i = 1, \dots, H \tag{82.17}$$

Where, h is error compensation coefficient.

82.3.4 PFC Optimal Control law for Step Setpoint

According to single value predictive control theory, set $H_1 = H_2 = H$, thus, the optimal criterion is

$$J_P = \min(y_r(k + H) - y_P(k + H))^2$$

$$y_r(k + H) = y_P(k + H) = \hat{y}(k + H) + h(y(k) - \hat{y}(k)) \tag{82.18}$$

Set $H = 2$, combining with Eqs. (82.14–82.16) and (82.18), manipulated value at time k is obtained.

$$u(k) = (\beta^2 y(k) + (1 - \beta^2) y_{sp} - h[y(k) - \hat{y}(k)] - C_2 - C_3) / C_1 \tag{82.19}$$

Where C_1, C_2, C_3 are the same as Eq. (82.15).

82.3.5 PFC Optimal Control law for Ramp Setpoint

Two base functions, that is, step and ramp function, should be chosen when setpoint includes ramp signal [6]. And Eq. (82.14) can be converted into (82.20).

$$u(k + i) = \lambda_1 + \lambda_2 i \quad (i = 0, 1, \dots, H - 1). \tag{82.20}$$

Where, λ_1, λ_2 are coefficients, H is prediction horizon. $H = 1$, output of predictive model:

$$\hat{y}(k + 1) = \sum_{i=1}^{nsv} S_i I'_{0,i}(N + 1) \lambda_1 + \sum_{i=1}^{nsv} S_i (I'_{0,i} \cdot I_1) + b \tag{82.21}$$

$H = 2$, the following expression can be obtained from Eq. (82.12) combining with (82.20).

$$\hat{y}(k + 2) = C'_1 \lambda_1 + C'_2 \lambda_2 + C'_3$$

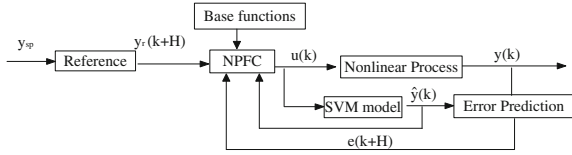
$$C'_1 = \sum_{i=1}^{nsv} S_i (I'_{0,i}(N + 1) + I'_{0,i}(N + 2)) + \sum_{i=1}^{nsv} \sum_{j=1}^{nsv} S_i I'_{0,i}(1) S_j I'_{0,j},$$

$$C'_2 = \sum_{i=1}^{nsv} S_i I'_{0,i}(N + 1)$$

$$C'_3 = \sum_{i=1}^{nsv} \sum_{j=1}^{nsv} S_i I'_{0,i}(1) S_j I'_{0,j} I_1 + \sum_{i=1}^{nsv} S_i I'_{0,i} I_2 + b [1 + \sum_{i=1}^{nsv} S_i I'_{0,i}(1)] \tag{84.22}$$

Theoretically, we can get $\hat{y}(k + H)$ mathematical expression at any prediction horizon H , the expression has the following form.

Fig. 82.1 PFC based on SVM



$$\hat{y}(k + H) = g_1(H)\lambda_1 + g_2(H)\lambda_2 + g_3(H) \tag{82.23}$$

Where, $g_1(H), g_2(H), g_3(H)$ are functions of $y(k), \dots, y(k - N + 1), u(k), \dots, u(k - M + 1)$. The chapter omits the deduction process due to complexity.

$$J_P = \min[(y_P(k + H_1) - y_r(k + H_1))^2 + (y_P(k + H_2) - y_r(k + H_2))^2] \tag{82.24}$$

Combining with (82.16), (82.17), the following equation group can be solved.

$$\frac{\partial J_P}{\partial \lambda_1} = 0 \text{ and } \frac{\partial J_P}{\partial \lambda_2} = 0 \tag{82.25}$$

$$u(k) = \lambda_1 = \frac{\gamma_1 + \gamma_2 - \gamma_3}{d_1} \tag{82.26}$$

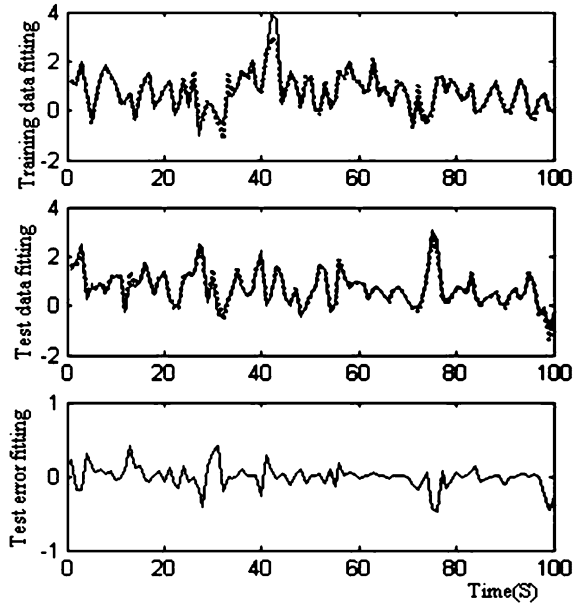
$$\begin{aligned} \gamma_1 &= g_2(H_1)g_3(H_2) - g_2(H_2)g_3(H_1) \\ \gamma_2 &= h \cdot (g_2(H_1) - g_2(H_2))(y(k) - \hat{y}(k)) \\ \gamma_3 &= g_2(H_1)y_r(k + H_2) - g_2(H_2)y_r(k + H_1) \\ d_1 &= g_1(H_1)g_2(H_2) - g_2(H_1)g_1(H_2) \end{aligned}$$

82.3.6 NPFC Control Algorithm Based on SVM

PFC based on SVM is shown in Fig. 82.1. The control algorithm is concluded as following:

- (1) Select parameters $C, \varepsilon, N, M, y_{sp}, T_r, T_s$.
- (2) S_i and $b, i = 1 \dots nsv$, are obtained by SVM nonlinear system identification using training data.
- (3) Test data verifies generalization of SVM.
- (4) At sample point k , expected setpoint $y_r(k + H)$ is calculated from reference trajectory Eq. (82.16).
- (5) Model output $\hat{y}(k + H)$ is obtained by SVM predictive model (82.15) or (82.23).
- (6) The revised predictive output $y_P(k + H)$ is obtained by Eq. (82.17).

Fig. 82.2 Training and test of SVM



- (7) The optimal control input $u(k)$ is calculated from Eq. (82.19) or (82.26), and $u(k)$ is executed.
- (8) set $k = k + 1$, and go to (4) step.

82.4 Simulation Study

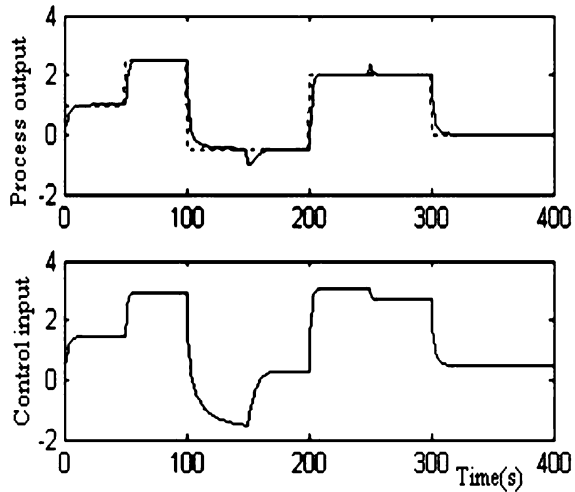
In the chapter, simulation study of nonlinear system from Ref. [5] is presented by using Matlab software, and assuming that $u(0), y(0)$ of nonlinear system are zero initial states.

Nonlinear system is described as following [5]

$$y(k + 1) = 0.4u(k) + 0.2y(k) + 0.1u(k)u(k - 1) \tag{82.27}$$

System sample time is 1 s, $N = M = 5$, one hundred training data and one hundred test data are obtained. The parameters of SVM are $C = 10^6, \varepsilon = 0.001$. Figure 82.2. shows fitting curves of training data, test data and fitting error of test data. In Fig. 82.2, the solid line is output of nonlinear system, the dotted line is output of SVM model. The SVM can accurately present dynamics of nonlinear system. Controller Eq. (82.19) is realized after SVM identification of nonlinear system. The sampling time is $T_s = 1$, time constant is $T_r = 5$. Figure 82.3 shows the simulation results, the dotted line is reference input, the solid line is process

Fig. 82.3 NPFC of nonlinear system



output. The results indicate that system output can quickly track reference trajectory. When the process is stable, step disturbance $\Delta = -0.5$ is applied into the process at $k = 150$ s, and $\Delta = 0.3$ at $k = 250$ s separately, the satisfied result still can be gotten as shown in Fig. 82.3. It indicates that system has good disturbance rejection.

82.5 Conclusions

The paper establishes a predictive model by nonlinear system identification with SVM based on linear kernel function, and then an explicit control law is obtained through the predictive functional control mechanism with both one base function (step function) and two base functions (step and ramp functions). Simulation results show that predictive functional control based on SVM can efficiently deal with nonlinear system. And the NPFC for nonlinear system has light online computational burden, good reference tracking and efficient disturbance rejection. The future manipulated variable is structured as a linear combination of a pre-specified set of base functions, and only a few parameters to be calculated online. Simulation results show that predictive functional control based on SVM can efficiently deal with nonlinear system. The NPFC for nonlinear system has light online computational burden, good reference tracking and efficient disturbance rejection.

References

1. Quanling Z, Shuqing W (2001) Nonlinear predictive control based on a neural network. *J Zhejiang Univ (Eng Sci)* 35(5):497–501
2. Shin SC, Park SB (1998) GA-based predictive control for nonlinear process. *Elect Lett* 34(20):1980–1981
3. Vapnik VN (1995) *The nature of statistical learning theory* [M]. Springer, New York
4. Chan WC et al (2001) On the modeling of nonlinear dynamic systems using support vector neural networks. *Eng Appl Artif Intel* 14:104–113
5. Zhong W, Pi D, Sun Y (2004) Study on model predictive control based on support machine. *Congress of fifth world intelligent control and automation*. vol 1, Hangzhou, China, pp 607-610
6. Zhang H, Han Z, Li C (2003) Support vector machine based nonlinear model predictive control. *Syst Eng Electron* 25(3):330–334

Chapter 83

Analysis and Circuit Simulation of a New Four-Dimensional Lorenz Time-Delay Chaotic System

Cui Zhiyong, Fan Zhongkui and Cao Panjing

Abstract A novel chaotic four-dimensional time-delay system based on Lorenz system was presented by linear feedback expansion. After the introduction of time delays, dynamical properties of the system have been further studied via Routh stability criterion. The oscillator circuit of the switchable time-delay chaotic system was designed by using Multisim. In computer simulations, experiment results by Multisim consistent with those of the continuous system in Matlab, which show that the new time-delay system has abundant dynamics behaviors.

Keywords Time-delay system · Routh stability criterion · Circuit simulation · Lyapunov–Krasovskii functional

83.1 Introduction

A chaotic system has been a subject of active research over the past four decades, whose signals displays broadband, non-periodic and unpredictable [1, 2]. Time-delay chaos systems can exhibit more complex hyper-chaotic behaviors, and many positive Lyapunov exponents could be generated even if a low dimensional time-delay system. Their properties offered by time-delay chaotic signals are relatively

C. Zhiyong (✉) · F. Zhongkui
Nanchang Campus, Jiangxi University of Science and Technology,
Nanchang 330013, China
e-mail: zhiyong49@hotmail.com

C. Panjing
Zhengzhou University, Zhengzhou 450000, China

suitable for signals used in the fields of electronic engineering, radars and information processing; in particular the secure communication systems.

During the last 4–5 years, designing analog circuit with discrete components to generate chaotic signals has been considerable interest in research for its potential usage. Time-delay chaos system generated by linear feedback expansion has more than one positive Lyapunov exponent, which means it possess complicated topological structures, those could better meet the practical applications need.

This chapter constructs a new time-delay chaos model by the introduction of linear time-delay perturbation in a kind of advanced Lorenz system. Subsequently the dynamical stability of the system is achieved based on the theory of Routh–Hurwitz criterion. Importantly, normal chaotic systems have been studied by analog circuits in relative papers [3, 4], and there are rarely reported design time-delay chaotic by circuit simulation in view of the complex hyper chaotic ones. In this paper, the circuit experiment of the new system through embedding a time-delay unit is described, and implementations are adequately supported by Multisim simulation results.

83.2 Description of System Model

To construct a four-dimensional chaotic system, we introduce a linear perturbation into the system (83.1) which is described follows. Where $x(t), y(t), z(t), w(t)$ are state variables of system, a, b, c, d and h are the control parameters. Obviously, there are two nonlinear equations in the fourthly system.

$$\begin{cases} \dot{x}(t) = a(y(t) - x(t)) \\ \dot{y}(t) = bx(t) + cy(t) - x(t)z(t) + w(t) \\ \dot{z}(t) = x(t)y(t) - dz(t) \\ \dot{w}(t) = -hx(t) \end{cases} \quad (83.1)$$

83.2.1 Simulation of the Constructed Time-delay System

When the constant parameters are selected as $a = 35, b = 7, c = 12, d = 3, h = 5$, the new system could be brought to chaotic even hyper chaotic. Time delay perturbation is added into the novel fourthly Eq. (83.1):

$$\begin{cases} \dot{x}(t) = a(y(t) - x(t)) + px(t - \tau) \\ \dot{y}(t) = bx(t) + cy(t) - x(t)z(t) + w(t) \\ \dot{z}(t) = x(t)y(t) - dz(t) \\ \dot{w}(t) = -hx(t) \end{cases} \quad (83.2)$$

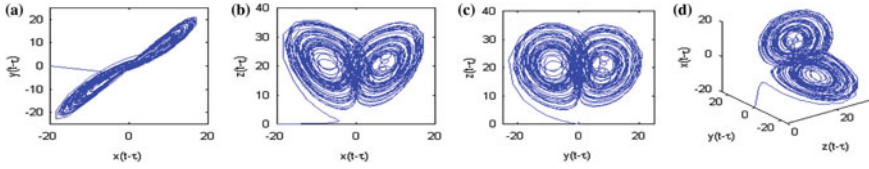


Fig. 83.1 Time-delay chaotic attractors: **a** variable $x(t - \tau)$ and $y(t - \tau)$. **b** variable $x(t - \tau)$ and $z(t - \tau)$. **c** variable $y(t - \tau)$ and $z(t - \tau)$. **d** variable $x(t - \tau)$, $z(t - \tau)$ and $w(t - \tau)$ chaotic tracks

where variables p and τ are called as control parameter and time quantum of tag respectively. If they are assigned appropriately, system (83.2) will turn into a time-delay chaotic system. We select control parameter $p = 3$, time-delay $\tau = 1.4$ ms. Figure 83.1 shows the dynamic behavior of the time-delay system. The trajectories of state variables $x(t)$ is illustrated in Fig. 83.2, and its power spectrum generated as follows:

There are three positive Lyapunov exponent which are calculated by wolf method in this constructed system, yet they prompt system to dynamics expand in more different directions simultaneously. Notably, the max Lyapunov exponent in the hyper chaotic system (83.2) is 0.6607, which indicates the possibility of chaotic motions. The Lyapunov exponent, which is plotted in Fig. 83.3.

The system (83.2) and Lorenz system could cross-transfer with each other through nonsingular continuous transformation with time-delay perturbation. On account of the infinite-dimension with high randomness and unpredictability in system (83.2), such as the topological structure. Therefore the complicated dynamical behaviors in time-delay system opens a largely potential field in information secure communication.

83.2.2 Stability Analysis of the New Time-Delay System

For the purpose of analyze the balance point in system (83.2), set:

$$\begin{cases} a(y(t) - x(t)) + px(t - \tau) = 0 \\ bx(t) + cy(t) - x(t)z(t) + w(t) = 0 \\ x(t)y(t) - dz(t) = 0 \\ -hx(t) = 0 \end{cases} \quad (83.3)$$

Nearby break-even point, the solution is not concerned with the time constant t , but rather in the time-invariant state $x(t)$, $y(t)$, $z(t)$, $w(t)$. By all appearances, system (83.3) could not depict a path line which improves over time. In this case, there is only one path curve passing through each point in the phase plane except for the balance point, which indicates the path curves could not intersect at all

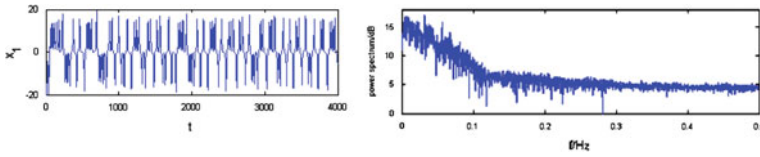
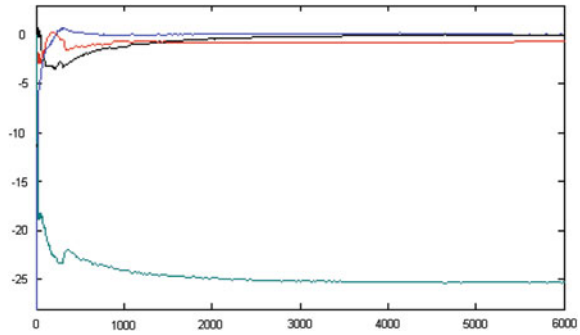


Fig. 83.2 The state variables $x(t)$ and its power spectrum

Fig. 83.3 Lyapunov exponent in the time-delay system



angles. The balance point $S_0 = (0, 0, 0, 0)$ is unique existence in discrete Eq. (83.3).

Given Jacobin matrix be denoted by J , according to time-delay system function, is written in the form of:

$$J = \begin{pmatrix} -a & a & 0 & 0 \\ b - z & c & -x & 1 \\ y & x & -d & 0 \\ -h & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} -a & a & 0 & 0 \\ b & c & 0 & 1 \\ 0 & 0 & -d & 0 \\ -h & 0 & 0 & 0 \end{pmatrix}. \tag{83.4}$$

Linearing the system (83.3) over the break-even point S_0 to structure the characteristic equation:

$$|J_s - \lambda I| = \begin{vmatrix} -a - \lambda & a & 0 & 0 \\ b & c - \lambda & 0 & 1 \\ 0 & 0 & -d - \lambda & 0 \\ -h & 0 & 0 & -\lambda \end{vmatrix} = 0. \tag{83.5}$$

Expansion of the determinant:

$$\lambda^4 + 26\lambda^3 + (-596)\lambda^2 + (-350)\lambda + 525 = 0. \tag{83.6}$$

We could acquire the necessary and sufficient condition of linear stable system which depends on Routh–Hurwitz stability criterion as Table 83.1:

Table 83.1 Routh stability analysis

System	Δ_1	Δ_2
λ^4	1	-596
λ^3	26	-350
λ^2	-582.5	525
λ^1	373.4	0
λ^0	525	0

The system (83.6) will remain steady around the balance point if and only if the values in the first column Δ_1 of the Table 83.1 are positive. According to the analysis of Table 83.1, we found that there was no way to keep stable nearby the point S_0 in system (83.6) on account of the negative value in column 1 and two positive real roots, which could affirm the existence of chaotic dynamical motions in Eq. (83.3).

83.3 Time-Delay Oscillator Circuit

Circuit design of chaotic system is an explicit way with physical realization, which can visually prove the dynamic behaviors of chaotic attractors [3]. Multisim of The National Instruments software platform, an expansion electronic simulation module, is adopted to complete the system-level switching circuit design on the basis of the Eq. (83.2).

83.3.1 Design of Chaotic Switching Circuit

Experimental switching circuit is given in Fig. 83.4, which consists of five channels. There are four integrators which correspond to system variables (V_x, V_y, V_z, V_w) and a time-delay part $X1$. In addition, we could realize the commutation circuit between different kinds of the chaotic systems by means of the on and off characteristics of circuit changer $K1$. The switch $K1$ is on means the four-dimensional Lorenz system (83.1), while the dynamical properties of time-delay system (83.2) would be shown in the state of switch $K1$ off.

Due to keep the nonlinear characteristics of self chaotic time-delay system, analog multipliers should be used to accomplish nonlinear operation, and the operational amplifiers (LF347BD) to achieve the additive operation. Otherwise, linear resistances and capacitors are assistant to accomplish operations such as addition, subtraction, multiplication and differentiation operations.

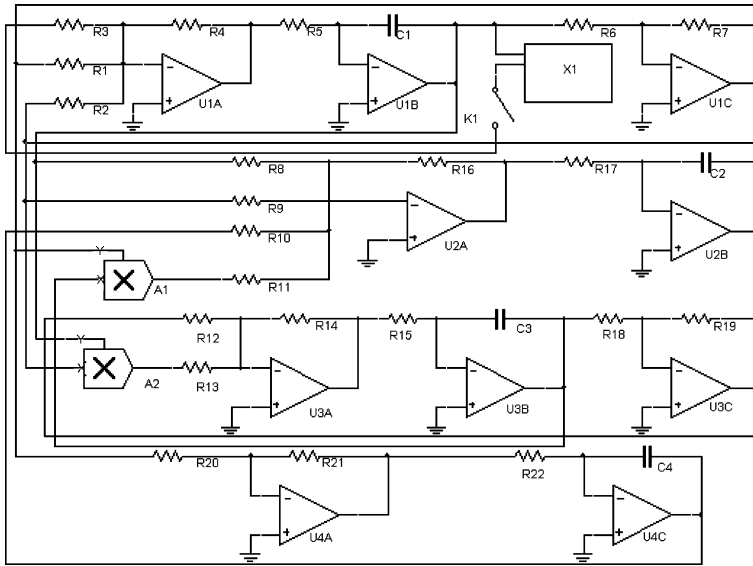


Fig. 83.4 Chaotic time-delay circuit simulation

When switch $K1$ is off, the system comes into time-delay chaotic state. The equation of circuit structure in Fig. 83.4 could be introduced with the nodal method of analysis.

$$\begin{cases} \dot{V}_x = -\frac{R_4 R_7}{R_1 R_5 R_6 C_1} V_x + \frac{R_4}{R_2 R_5 C_1} V_y + \frac{R_4}{R_3 R_5 C_1} V_x(\tau) \\ \dot{V}_y = \frac{R_{16}}{R_8 R_{17} C_2} V_x + \frac{R_{16}}{R_9 R_{17} C_2} V_y - \frac{G_1 R_7 R_{16}}{R_6 R_{11} R_{17} C_2} V_x V_z + \frac{R_{16}}{R_{10} R_{17} C_2} V_w \\ \dot{V}_z = -\frac{R_{14} R_{19}}{R_{12} R_{15} R_{18} C_3} V_z + \frac{G_2 R_{14}}{R_{13} R_{15} C_3} V_x V_y \\ \dot{V}_w = -\frac{R_7 R_{21}}{R_6 R_{20} R_{22} C_4} V_x \end{cases} \quad (83.7)$$

In respect that capacitors have no voltage before the integral, the initial value of capacitors end voltage is zero. G_1 and G_2 are output gains of analog multipliers in variable $y(t), z(t)$ channels respectively. The other resistance values are selected as $R_1, R_2, R_6, R_7, R_{14} = 10 \text{ k}\Omega$, $R_5, R_{10}, R_{15} = 100 \text{ k}\Omega$, $R_{11}, R_{13} = 1 \text{ k}\Omega$, $R_4 = 35 \text{ k}\Omega$, $R_8 = 15 \text{ k}\Omega$, $R_9 = 8.3 \text{ k}\Omega$, integrator capacitances are set $C_1, C_2, C_3, C_4 = 1 \text{ }\mu\text{F}$. As it were, the model (83.7) of circuit structure in Fig. 83.4 is in accordance with that of new time-delay system (83.2).

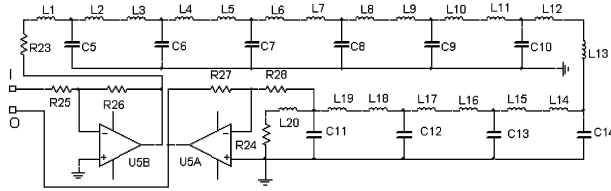


Fig. 83.5 The design of time-delay cell

83.3.2 Time-Delay Circuit Unit

Framework X1 shown in Fig. 83.5 is a time-delay cell which includes a T-LCL filter network. The filter, a typical frequency selection circuit, could transmit the signal in selected frequency band without attenuation. This filter which involves two inductors and a capacitor is suitable for filtering high-frequency signals. The capacitance would decrease with increasing frequency, as the inductance enhance with the increase of frequency. Therefore, the capacitor is connected to serial interface with the inductor connected to parallel interface. The time-lag network built by low-pass filter could behave smooth performance beneath the cut-off frequency $f_c = 1$ kHz.

The system resistant values are $R_{25}, R_{26} = 10$ k Ω , $R_{27} = 22$ k Ω . The single-section filter could not accomplish the desired effect when the frequency of noise is close to useful signal frequency. At this time we need to make use of the multi section filter. As Fig. 83.5 showing, ten pieces of filter modules which have constant values of characteristic impedance and two build-out resistors R_{23}, R_{24} are set up between the input and output ports. Delay time τ could be approximately obtained from:

$$\tau = n \left(\sqrt{2LC} \right). \tag{83.8}$$

When $L = 20$ mH, $C = 500$ nF, the delay time $\tau \approx 1.4$ ms which was equal to that in the constructed chaotic system would be approximately attained from Eq. (83.8). By means of a laboratory test in circuit layout, frequency of variable V_x is about 700 Hz, much lower than the cut-off frequency in low-pass filter, which may ensure the signals pass through the delay channel with extra low loss. To test and verify the effects on lag time, we introduce a type of 700 Hz sine wave into this lag part by the signal generator. It is found that there is 1.4 ms delay time in existence between the input and output sine waves through contrasting in Fig. 83.6.

In order to gain quite clear phase diagrams, we could adjust the capacitance values $C_1, C_2, C_3, C_4 = 1$ nF, on account of the low output frequency of state variables, generally in the range of 1–10 Hz. In this way, the signal frequency is a thousand times as great as that of previous design without the impacts on chaotic behaviors, and the signal display rate is able to be improved. Figs. 83.7 and 83.8

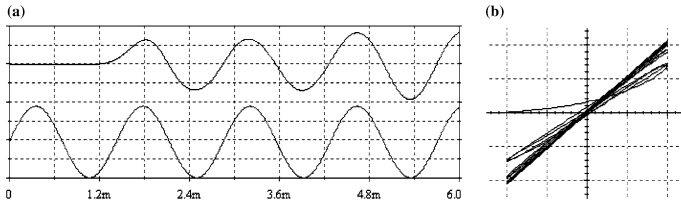


Fig. 83.6 Time-delay effect: **a** Comparison waveform (0.6 ms/div, 1 V/div). **b** Comparison diagram of signal synchronization (1 V/div,1 V/div)

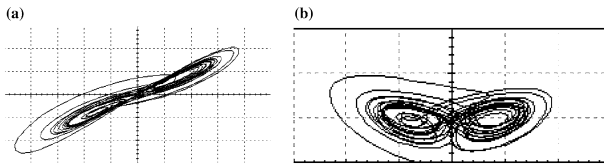


Fig. 83.7 Image of the four-dimensional Lorenz system based on voltage: **a** $V_x - V_y$ (0.5 V/div, 1 V/div). **b** $V_x - V_z$ (1 V/div, 2 V/div)

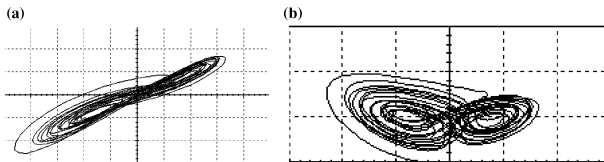


Fig. 83.8 Image of the constructed chaotic behavior based on voltage: **a** $V_x - V_y$ (0.5 V/div, 1 V/div). **b** $V_x - V_z$ (1 V/div, 2 V/div)

83.4 Conclusion

A new constructed time-delay chaotic system through introducing a linear perturbation has been designed, and an analog electronic implementation of this system by using Multisim software was proposed. Experimental results indicate that the new system is highly sensitive to initial conditions although the simple topological structure. The phase portraits in circuit which are exactly the same with the continuous one by Matlab simulation, will come to a theoretical basis for the design and realization of practical chaotic system, and the circuit design can be available for reference in the development of chaotic communication systems in the future.

Acknowledgments This work is supported by the National Natural Science Foundation of China (Grant No. 11062002).

References

1. Liu YZ, Chen LQ (2001) Nonlinearvibration. Higher Education Press, Beijing
2. Chen GR, Lü JH (2003) Dynamic analysis, control and synchronization of Lorenz family system. Science Press, Beijing
3. Torres LA, Aguirre LA (2000) Inductorless Chua's circuit. *Electron Lett* 36:1915–1916
4. Qi GY, van Wyk MA (2009) A new hyperchaotic system and its circuit implementation. *Chaos Solitons Fractals* 40:2544–2549

Chapter 84

The Design of Adaptive PI Speed Controller for Permanent Magnet Synchronous Motor Servo System

Yulin Gong and Yongyin Qu

Abstract This chapter proposes a novel adaptive PI speed controller for permanent magnet synchronous motor (PMSM) servo system. A radial basis function neural network (RBFNN) is used for the controller to identify the PMSM servo system and an improved particle swarm optimization (IPSO) algorithm is adopted to optimize the parameters of the RBFNN for fast training. The Jacobian information of PMSM obtained by RBFNN identification is used to adjust the adaptive PI parameters for speed controller online. Simulation result shows good dynamic response, strong robustness and satisfactory control performance in PMSM servo system.

Keywords PMSM · Adaptive PI controller · RBFNN · IPSO algorithm

84.1 Introduction

There are many advantages in Permanent magnet synchronous motor servo system including high torque to current ratio, large power to weight ratio and higher efficiency. PMSM has been widely used in industry, national defense and other social aspects.

Y. Gong (✉)
Changchun University of Science and Technology, Weixing Road,
Changchun 7089, China
e-mail: garrygong1983@126.com

Y. Qu
Beihua University, Huashan Road, Jilin, 3999, China
e-mail: qyy1217@yahoo.com.cn

But many uncertain factors can influence the performance of PMSM servo system such as parameter changes, external load disturbances and nonlinear system identification errors.

Numerous methods have been proposed to improve the performance of PMSM servo system including model reference adaptive control, sliding mode control, neural network control and so on. But the PI controller is still the most widely used for PMSM servo system due to its relatively simple structure.

Radial basis function neural network has been widely used in the areas of system identification and modeling because of its simple structure and approximation capabilities [1, 2].

Particle swarm optimization (PSO) algorithm as an intelligent algorithm which based on stochastic search is efficient to solve complicated engineering problem [3].

In this chapter an improved particle swarm optimization (IPSO) algorithm is used in RBFNN to improve further the learning rate and approximation capabilities.

84.2 Mathematical Model of the PMSM

The vector control technique was firstly proposed for induction motors and it is applied to PMSM later. The basic principle is to decouple the stator current to get direct axis and quadrature axis components.

Based on the vector control technique, the electrical and mechanical equations of the PMSM can be described in the rotor reference frame are summarized as follows [4]:

$$u_d = R_s i_d + p\psi_d - \omega_r \psi_q \quad (84.1)$$

$$u_q = R_s i_q + p\psi_q - \omega_r \psi_d \quad (84.2)$$

$$\psi_d = L_d i_d + \psi_f \quad (84.3)$$

$$\psi_q = L_q i_q \quad (84.4)$$

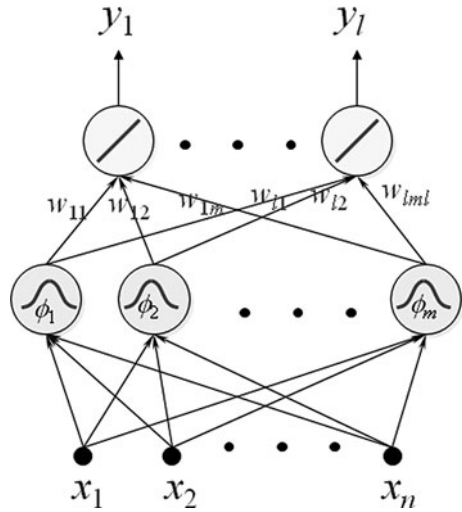
where i_d , i_q is d-axis and q-axis current in synchronous frame; u_d , u_q is d-axis and q-axis voltage in synchronous frame; L_d , L_q is d-axis and q-axis inductance; ψ_d , ψ_q is d-axis and q-axis flux linkage in synchronous frame; ψ_f is PM flux linkage in synchronous frame; R_s is motor phase resistance.

The electric torque can be expressed as follows:

$$T_{em} = p_n (\psi_d i_q - \psi_q i_d) \quad (84.5)$$

$$T_{em} = T_L + B \left(\frac{\omega_r}{p_n} \right) + J p \left(\frac{\omega_r}{p_n} \right) \quad (84.6)$$

Fig. 84.1 Structure of RBFNN



where T_{em} is electromagnetic torque; T_L is load torque; ω_r is motor electrical angular velocity; p_n is number of magnetic poles; B is viscous damping coefficient; J is moment of inertia of the motor.

84.3 Identification RBFNN Based on IPSO Algorithm

84.3.1 RBFNN

The RBFNN is a feed-forward neural network with three layers. The learning rate of RBFNN is quickened greatly and the local minimum is avoided. The Structure of RBFNN is shown in Fig. 84.1. Supposed the radial vector of RBFNN adopts Gaussian function [5, 6].

The training parameters set and kernels are described as follows:

$$\mathfrak{R} = \left\{ \left(X^{(k)}, Y^{(k)} \right) \right\}_{k=1}^p \tag{84.7}$$

$$\phi_j(X) = \exp\left(-\frac{\|X - \mu_j\|^2}{2\sigma_j^2}\right) \tag{84.8}$$

where w_{ij} are weights, μ_j is the centers of Gauss function, σ_j is the variance of Gauss function.

$$f_i(X^{(k)}) = \sum_{j=1}^l w_{ij} \phi_j(X^{(k)}) \tag{84.9}$$

The performance function of RBFNN is described as follows:

$$J = \frac{1}{2}(y(k) - y_m(k))^2 \quad (84.10)$$

According to gradient descent method, iterative algorithm of weights, node center and node radial parameters are as follows:

$$\Delta w_{ij} = \eta_1 \left(y_i^{(k)} - f_i \left(X^{(k)} \right) \right) \phi_j \left(X^{(k)} \right) \quad (84.11)$$

$$w_{ij}(k) = w_{ij}(k-1) + \Delta w_{ij} + \alpha(w_{ij}(k-1) - w_{ij}(k-2)) \quad (84.12)$$

$$\Delta \mu_j = \eta_2 \phi_j \left(X^{(k)} \right) \frac{X - \mu_j}{\sigma_j^2} \sum_{i=1}^l w_{ij} \left(y_i^{(k)} - f_i \left(X^{(k)} \right) \right) \quad (84.13)$$

$$\mu_j(k) = \mu_j(k-1) + \Delta \mu_j + \alpha(\mu_j(k-1) - \mu_j(k-2)) \quad (84.14)$$

$$\Delta \sigma_j = \eta_3 \phi_j \left(X^{(k)} \right) \frac{\|X - \mu_j\|^2}{\sigma_j^3} \sum_{i=1}^l w_{ij} \left(y_i^{(k)} - f_i \left(X^{(k)} \right) \right) \quad (84.15)$$

$$\sigma_j(k) = \sigma_j(k-1) + \Delta \sigma_j + \alpha(\sigma_j(k-1) - \sigma_j(k-2)) \quad (84.16)$$

where η_1, η_2, η_3 is learning rate and α is inertia coefficient of RBFNN.

84.3.2 IPSO Algorithm

In 1995, the PSO algorithm was first introduced by Kennedy and Eberhart. As a population-based stochastic search algorithm, it attempts to simulate the natural process of group communication, such as bird flock and fish flock.

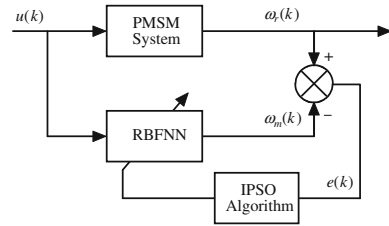
The PSO algorithm has been widely applied for reasons that it is easy to realize, has little information about the optimal problem and low demand. The PSO algorithm generates a set of solution of objective function random in order to get an initial population and each individual is called a particle. Every particles search the optimal position like birds which are searching food, following the optimal particle in the previous group along their own best track to “pass” the solution space of the problem. Every particle can modify their velocity by Eq. 84.17 and position by Eq. 84.18. The changes of velocity are represented by the former velocity, the previous personal best position and the current global best position.

$$V_i(k+1) = \omega V_i(k) + c_1 r_1 (P_i - X_i(k)) + c_2 r_2 (P_g - X_i(k)) \quad (84.17)$$

$$X_i(k+1) = X_i(k) + V_i(k+1) \quad (84.18)$$

In Eq. 84.17, $0 \leq \omega \leq 1$ is defined as the inertia weight which determines the former velocity's proportion of each particle. c_1 is cognitive parameter and c_2 is

Fig. 84.2 Structure of PMSM identification based on IPSO-RBFNN



social parameter, r_1 and r_2 random functions in the range of $[0, 1]$. Every new position must be evaluated by fitness function.

In order to accelerate the speed of be optimized and to avoid particle swarm algorithm to get into a local optimum, it is necessary to properly adjust the value of ω [7]. In the early stages, a big speed-weight ω can prevent algorithm from getting into a local optimization. In the later stage, small one is propitious to accelerate algorithm converge. In this algorithm, a variable parameters method is used in the iterative process. This way can be described as follows:

$$\omega = \omega_{initial} - (\omega_{initial} - \omega_{final})(k/N_{max})^2 \tag{84.19}$$

$$c_1 = c_{1initial} - (c_{1initial} - c_{1final})(k/N_{max})^2 \tag{84.20}$$

$$c_2 = c_{2initial} - (c_{2initial} - c_{2final})(k/N_{max})^2 \tag{84.21}$$

where ω is time-varying inertia weight factor, c_1 and c_2 are the time-varying cognitive and social parameters. $\omega_{initial}$ and ω_{final} are the initial and final values of the weight factor; $c_{1initial}$ and c_{1final} are the initial and final values of the cognitive parameter; $c_{2initial}$ and c_{2final} are the initial and final values of the social parameter; k is the current iterate time; N_{max} is the maximum iterate time.

Learning from the genetic algorithm, variation operation is introduced to reinitialize some particles in certain probability. The variation operation can make particles jump out of the continuous reduced space, keep the diversity of the particles and improved the possibility of optimizing.

84.3.3 Identification RBFNN

In this chapter, the IPSO algorithm is used to adjust the weight to train RBFNN. Let w_i be a particle, $\eta = 0.85, \alpha = 0.05$. The structure of PMSM identification based on IPSO-RBFNN is shown in Fig. 84.2.

The convergence of the IPSO algorithm toward the global optimal solution is guided by the objective function. In this chapter, the fitness function is defined as follows:

$$FIT = \omega_r - \omega_m \tag{84.22}$$

Table 84.1 Parameters of IPSO algorithm

Parameters	Values
Number of generation	200
Size of population	20
Initial inertial weight	0.9
Final inertial weight	0.4
Initial cognitive parameter	0.5
Final cognitive parameter	2.5
Initial social parameter	0.5
Final social parameter	2.5
Variation probability	0.9

where ω_r is the output of PMSM and ω_m is the output of RBFNN; $\omega_r - \omega_m$ is the identification error of PMSM.

The parameters of IPSO algorithm are showed in Table 84.1 as follows:

84.3.4 Design of Adaptive Speed Controller Based on RBFNN

PI controller has been applied to control PMSM servo systems [8–10] and the structure of adaptive PI controller with the IPSO algorithm is shown in Fig. 84.3. With changes in the PMSM servo system including parameters and external disturbances, the adaptive PI speed controller providing optimal PI parameters automatically, avoid manual tuning.

The PI Control law is:

$$u(k) = u(k-1) + \Delta u(k) \quad (84.23)$$

$$\Delta u(k) = k_p(e(k) - e(k-1)) + k_i e(k) \quad (84.24)$$

where $e(k)$ is the PMSM servo system error between desired output and actual output:

$$e(k) = \omega_r(k) - \omega_m(k) \quad (84.25)$$

$u(k)$ is the control command, k_p is the proportional gain, k_i is the integral gain.

The input of PI controller is as follows:

$$x_1(k) = e(k) - e(k-1) \quad (84.26)$$

$$x_2(k) = e(k) \quad (84.27)$$

So the output of controller is described as follows:

$$y(k) = w_1 x_1(k) + w_2 x_2(k) \quad (84.28)$$

Fig. 84.3 Structure of adaptive PI controller with the IPSO algorithm

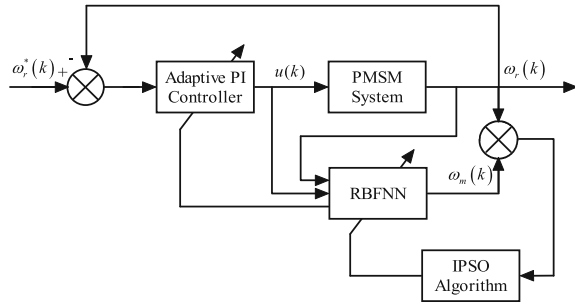


Table 84.2 The parameters of PMSM

Parameters	Values
Rated power (W)	750
Rated voltage (V)	380
Rated current (A)	2.5
Pole pairs	2
Stator resistance (Ω)	1.6
d-axis inductance (mH)	1.54
q-axis inductance (mH)	1.54
Motor inertia (kg·m ²)	0.0156

where w_i are the weights of RBFNN:

$$\Delta k_p = -\eta \frac{\partial E}{\partial k_p} = -\eta \frac{\partial E}{\partial y} \frac{\partial y}{\partial u} \frac{\partial u}{\partial k_p} = \eta e(k) \frac{\partial y}{\partial u} x_1(k) \tag{84.29}$$

$$\Delta k_i = -\eta \frac{\partial E}{\partial k_i} = -\eta \frac{\partial E}{\partial y} \frac{\partial y}{\partial u} \frac{\partial u}{\partial k_i} = \eta e(k) \frac{\partial y}{\partial u} x_2(k) \tag{84.30}$$

where $\frac{\partial y(k)}{\partial u(k)}$ is the Jacobian information which can be obtained by RBFNN identification as follows:

$$\frac{\partial y(k)}{\partial u(k)} \approx \frac{\partial y_m(k)}{\partial u(k)} = \sum_{j=1}^m w_j h_j \frac{\mu_j - u(k)}{\sigma_j^2} \tag{84.31}$$

The k_p, k_i can be update by Eq. 84.29 and 84.30.

84.4 Simulation Results

The parameters of simulated PMSM are showed in Table 84.2.

Figures 84.4 and 84.5 show the speed response curve and the tracking curve of PMSM servo system.

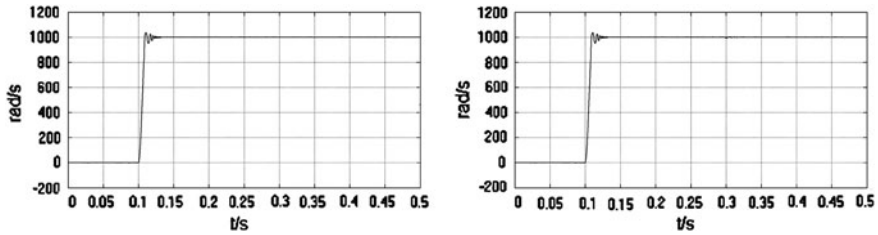


Fig. 84.4 (a) Speed response curve of PMSM servo system (b) Speed response curve of PMSM servo system when $T_L = 3N$ at 0.3s

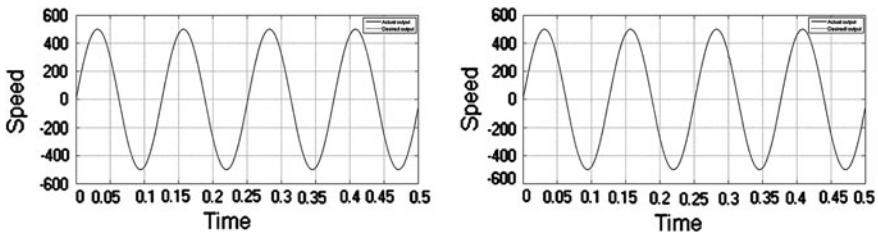


Fig. 84.5 (a) Speed tracking curve of PMSM servo system (b) Speed tracking curve of PMSM servo system when $T_L = 3N$ at 0.3s

84.5 Conclusion

An adaptive PI speed controller based on IPSO-RBFNN is proposed for PMSM servo system in this chapter. RBFNN is used to identify the PMSM servo system, and then the information of RBFNN identification is used to adjust the parameters of PI speed controller online. In order to improve the accuracy and convergence speed of RBFNN, IPSO algorithm is adopted to train the network. Simulation results show that the novel adaptive PI speed controller has adaptive ability, strong robustness and satisfactory control performance in PMSM servo system.

References

1. Elanayar SVT, Shin YC (1994) Radial basis uncton neural network for approximation and estimation of nonlinear stochastic dynamic systems. *IEEE Trans Neur Net* 5(4):584–603
2. Warwick K (1996) An introduction to radial basis functions for system identification: a comparison with other neural network methods. In: *Proceedings of the 35th conference on decision and control, Kobe*, pp 464–469
3. J Kennedy, R Eberhart (1995) Particle swarm optimization. In: *Proceedings of IEEE Conference on neural networks, IEEE service center, Piscataway* 4:1942–1948
4. Bimal KB (2002) *Modern power electronics and AC drive*[J]. Prentice Hall Inc, New Jersey

5. Zhirong G, Shunyi X, Wei G (2009) Study on a recurrent functional link-based fuzzy neural network controller with improved particle swarm optimization [J]. In: The 9th international conference on electronic measurement & instruments, pp 1095–1110
6. Rank E (2003) Application of bayesian trained RBF network to nonlinear time-series modeling. *IEEE Trans Neur Net* 83:1393–1410
7. Elwer AS, Wahsh SA (2007) Improved performance of permanent magnet synchronous motor by using particle swarm optimization techniques. In: Proceedings of the 26th Chinese control conference, Zhangjiajie, pp 326–330
8. Weider C, Shunpeng S (2010) PID controller design of nonlinear systems using an improved particle swarm optimization approach. *Commun Nonlin Sci Num Simul* (23):3632–3639
9. Mohammad RF, Mohammad SA (2010) Design an optimized PID controller for brushless DC motor by using PSO and based on NARMAX identified model with ANFIS. In: The 12th international conference on computer modelling and simulation, pp 16–21
10. Yuzeng Z, Jianxin S, Shuhan S et al (2010) Adaptive PID speed controller based on RBF for permanent magnet synchronous motor system. In: International conference on intelligent computation technology and automation, pp 425–428

Chapter 85

Application of BP Neural Network in Faults Diagnosis of Process Valve Actuator

Fukai Deng, Erqing Zhang, Qunli Shang, Shan-en Yu and Yifang He

Abstract As modern process plants become more complex, the ability to detect and identify the faulty operation of pneumatic control valves is becoming increasingly important. This chapter investigates Back Propagation (BP) neural network with improved algorithm to diagnose the pneumatic control valve faults. The particular values of six measurable quantities are shown to depend on the severity of commonly occurring faults. The relationships between these parameters from fault signatures for each operating condition that are subsequently learned by a multilayer BP neural network. Through the activity of the Development and Application of Methods for Actuator Diagnosis in Industrial Control Systems (DAMADICS) research to train the network. The simulation experiments' results prove the BP neural network has the capability to detect and identify various magnitudes of the faults of interest and can isolate multiple faults. In addition, it is observed that the network has the ability to estimate fault levels not seen by the network during training.

Keywords BP neural network · Process valve actuator · Fault diagnosis and isolation (FDI) · DAMADICS

F. Deng (✉) · E. Zhang · Q. Shang · S. Yu · Y. He
College of Automation, HangZhou Dianzi University, Hangzhou 310018, China
e-mail: fukai1015@126.com

E. Zhang
e-mail: sxhd2004@126.com

Q. Shang
e-mail: qlshang@sina.com

S. Yu
e-mail: nwcd@163.com

Y. He
e-mail: heyifang111984@163.com

85.1 Introduction

The actuator is one of foundation equipments in industrial automation process, as a terminating meter of self-regulation system, plays a very important role. The technology of actuator profoundly affects the safety and reliability of production process. Due to the frequent movements, and high temperature or pressure and other scurviness environmental, actuator is easy to get broken down. As modern process plants become more complex, the ability to detect and identify the faulty operation of pneumatic control valves is becoming increasingly important. To build highly efficient, timely and accurate fault diagnosis system, ensure production safety, has become focus research in control field. Many researchers have suggested a number of useful fault diagnosis techniques including model-based, expert system based, reliability-based, fuzzy theory-based and other technology [1–3]. In this chapter we discuss the application of BP neural network in pneumatic process valve faults diagnosis.

Modern methods to solve FDI problems in dynamic systems can be grouped into three broad categories, the first being a model-based approach. The model-based approach models the normal process using a priori mathematical information about the system [3]. However, it is well known that constructing mathematical models for complex systems or for nonlinear systems can be difficulty, so the model-based approach has some disadvantage in application. Furthermore, even make a model, there must be a great deal of experimentation to validate the model. This method is not practical in complex system. An alternate approach to the model-based FDI has been to train neural networks to estimate the model of the dynamic system [4].

The third class method is to use neural network as pattern classifiers to solve FDI problems. Here, the networks are trained to recognise data representing the different failure modes of the dynamic system. This approach has been used in a variety of applications [4]. Other methods have also demonstrated the usefulness of this approach in the applications of sensor fault diagnosis and fault diagnosis of chemical processes, respectively.

This chapter investigates a BP neural network with a improved algorithm to diagnose faults in the process valve. The multilayer feed forward structure was selected for use as it has been shown to perform well in similar FDI problems (Sorsa, Koivo, and Koivisto 1991; Koivo 1994; Crowther et al. 1998) [4].

The organisation of the remainder of this chapter is as follows: [Sect. 85.2](#), The typical process pneumatic control valve is described. The faults can be easily quantified and simulated experimentally via six measured features. [Section 85.3](#) describes the BP neural network structure and the application in pneumatic actuator fault diagnosis. The structures and training of neural network is proposed. It is shown that the six measured features of the pneumatic actuator are sufficient to distinguish characteristic the faults and are suitable inputs to the neural network. Finally, the performance of the BP neural network is discussed in [Sect. 85.4](#). It is concluded that

Fig. 85.1 Actuator with spring-and-diaphragm pneumatic servomotor positioner and control valve



the BP neural network with a improved algorithm successfully diagnose the pneumatic actuator faults and estimate the degree of faults. It has a ability to isolate the multiple faults. Through Matlab/Simulink simulation verify the network has perfect performance. In Sect. 85.5 some concluding remarks are brought to be a close.

85.2 Experimental Test Rig

85.2.1 Pneumatic Control Valve

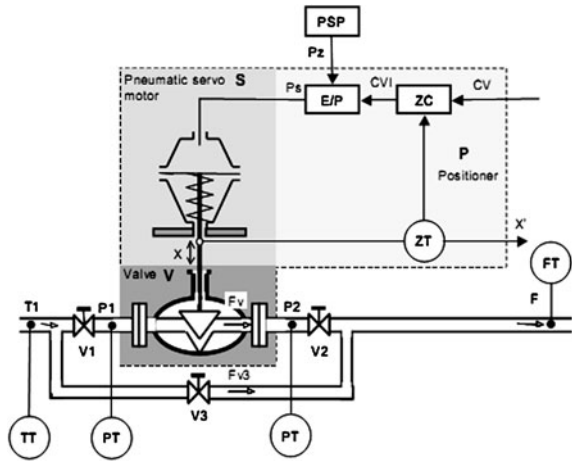
The pneumatic control valve include three parts: pneumatic servo motor, positioner, and valve, as shown in Fig. 85.2. The valve was modified so that the fault of interest could be introduced experimentally. The flow control device, or valve body, is stroked by a pneumatically driven actuator [5, 6]. With locator pneumatic control valve is shown in Fig. 85.1.

Positioner module consists of a electricity to pressure (E/P) converter, and a pneumatic relay receiving DCS system (or regulator) control signal CV (ordinary 4–20 mA DC), and valve position feedback signal to Produce residual. After sufficient dynamic PI or PID calculation to deliver to E/P . So that a relatively large volume of air can be delivered to, or vented from the upper chamber of the pneumatic servo motor. Positioner module closed the loop so that position control of the valve stem is realised. Referring to Fig. 85.2, it is observed that these parameters describe the main features of the valve states and are indicative of the dynamic and static characteristics of the response.

85.2.2 The Set of Available Signals

The limited set of 6 available measurements and 1 control value signal have been considered : process control external signal CV , values of liquid pressure on the

Fig. 85.2 Schematic diagram of the flow control valve



valve inlet P_1 and outlet P_2 , stem displacement X , main pipeline flow rate F , supply air pressure p_z and liquid temperature T_1 . A distinction has been made between real values of physical variables and its measurements. The six measured features of the pneumatic actuator are sufficient to distinguish characteristic the faults and are suitable inputs to the neural network. The set of main variables and notion used in training network as input vector is as follows: P_z supply air pressure; CV process control external signal; T_1 liquid temperature; P_1 pressures on valve inlet; P_2 pressures on valve outlet; F main pipeline flow rate; X valve plug displacement. These parameters can be calculated through communicates with onboard sensors of the valve. The causal graph of the benchmark actuator has been given. Physical relations between variables have been formulated in the actuator benchmark and have been given in the succeeding sections [6].

85.2.3 Control Valve Faults

The pneumatic control valve include three parts: pneumatic servo motor, positioner and valve. If there are faults emerge, it is commonly caused by one or more part broken down. The 19 actuator faults have been considered in the DAMADICS actuator FDI benchmark study [6]. The faults are classified into four following groups: Control valve faults; Pneumatic servo-motor faults; Positioner faults; General faults/external faults.

For the most part, single actuator faults are observed in industrial practice whilst multiple faults rarely occur. Control valve faults classified seven categories as Table 85.1 follows:

Table 85.1 Control valve faults

Number	Faults
1	Valve clogging
2	Valve or valve seat sedimentation
3	Valve or valve seat erosion
4	Increase of valve friction
5	External leakage (leaky bushing, covers, terminals)
6	Internal leakage (valve tightness)
7	Medium evaporation or critical flow

85.3 BP Neural Network

85.3.1 Network Structure

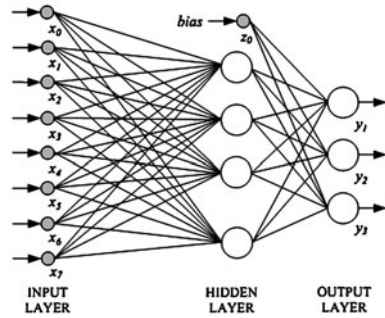
Single layer networks were not considered due to their poor generalisation ability as compared with multilayer networks [4]. With a single hidden layer, the network can precisely approach any nonlinear system. As contains a hidden layer the multilayer network can dramatically improve network classification ability. In this chapter, the networks are trained to recognise data representing the different failure modes of the dynamic system. It has three layers: one input layer, one hidden layer and one output layer, as Fig. 85.3 shown.

The input layer contains six nodes, one node corresponds to each available performance parameters. The number of neurons in the hidden layer is adjustable. Since the network does not contain any feedback loops, its output can be calculated as an explicit function of the network inputs and the network weights [8]. The seven linear output neurons estimate the magnitude of the faults according to the following relation, i.e.

$$y_k = \sum_{j=1}^m w_{jk} \tanh \left(\sum_{i=0}^7 w_{ji} x_i \right) + w_{0k} z_0, k = 1, 2, 3 \tag{85.1}$$

where x_i is the input and y_k is the network output. w_{jk} is the adjustable weight between input i and hidden neuron j and w_{jk} is the adjustable weight between hidden neuron j and output neuron k . Index m denotes the number of neurons in the hidden layer. Note that the bias terms are absorbed into Eq. 85.1 by setting $x_0 = 1$ and $z_0 = 1$. The hidden layer neurons use hyperbolic tangent activation functions. The number of output neurons is set equal to the number of class vectors. The network is trained such that the neuron with the largest output indicates the condition of the valve. In this chapter, we used seven faults of the pneumatic control valve. So we defined the neural network structure is 6-13-7.

Fig. 85.3 Network architecture



85.3.2 Network Training

The Levenberg–Marquardt optimization technique has been used to train multi-layer neural networks and was shown to converge faster than the gradient descent method. Levenberg–Marquardt optimization technique is one of the most popular and effective second-order algorithms for training feedforward neural networks. One of the characteristics of this method is that it combines the stability of gradient descent with the speed of Newton’s algorithm [7]. The Levenberg–Marquardt learning expression is formulated as follows, i.e.

$$\Delta W(k) = G(k)^{-1}J(k)e(k). \tag{85.2}$$

where $G(k) = J^T(k)J(k) + \lambda(k)I$. The resulting weight update rule is given by

$$\Delta W(k) = -\frac{\lambda_1}{2\lambda_2}G(k)^{-1}g(k) + \frac{1}{2\lambda_2}\Delta W(k - 1). \tag{85.3}$$

The regularization parameter $\lambda(k)$ is decreased by a factor of ten when the following wolfe condition holds.

$$f(W(k) + a(k)d(k)) \leq f(W(k) + C_1a(k)g(k)d(k)). \tag{85.4}$$

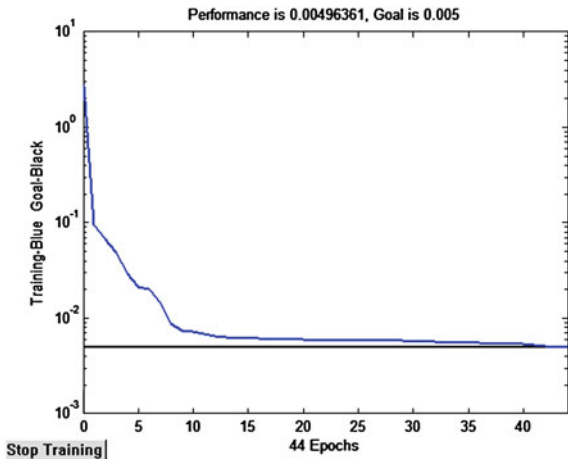
where C_1 is a constant in the interval (0,1), $f(\cdot)$ is the objective function [9]. The previous inequality can be simplified as follows, i.e.

$$E(k + 1) < E(k + 1) + 0.1g(k)^T \Delta W(k). \tag{85.5}$$

In this case the constant C_1 is set to 0.1. To facilitate the assessment of the generalisation ability of the trained network, the available experimental data were subdivided into training and test sets. Each vector of input data is assigned a class vector that represents a specific mode of operation (either normal or faulty). The number of output neurons is set equal to the number of class vectors.

We through DAMADICS [6] make the simulation experiment, get the available experimental data in normal or fault cases. In a particular situation of determined

Fig. 85.4 Residual responses



strength of fault, making control signal CV is a sine signal and $\Delta p = p_1 - p_2$ is a constant. Setting several different fault type and intensity, respectively is normal $f = 0$, valve clogging $f = 0.75$, valve or valve seat sedimentation $f = 0.5$, valve or valve seat erosion $f = 0.75$, increase of valve friction $f = 0.75$, external leakage (leaky bushing, covers, terminals) $f = 0.75$, internal leakage (valve tightness) $f = 0.25$, and valve clogging or increase of valve friction happening at the same time $f_1 = 0.75, f_2 = 0.75$. Get eight different fault conditions of system parameters, and the value of the parameter, all recording every measure volume CV, p_1, p_2, T_1, f, F and X . Another simulation experimental at the other condition: $\Delta p = p_1 - p_2$ is a sine signal and CV is a constant, with the same intensity of the same fault. So the data can be used to train and test the network performance.

85.4 Results

This chapter use the improved algorithm of Levenberg–Marquardt algorithm to train network. Hidden layer transfer function is logsig function, and the purelin function for the transfer function of output layer. Maximum training for 2000 steps. Interval for 100 steps. Residual is seted as $5e-3$. Via Levenberg–Marquardt algorithm training the neural network, performance of the network as follow, Fig. 85.4

Here, the networks are trained to recognise data representing the different failure modes of the Control valve. The BP neural network is implemented to detect and identify actuator faults in a typical pneumatic control valve. The output of trained neural network as the follow Table 85.2. It is concluded that BP neural network has perfect ability to diagnosis control valve faults.

Table 85.2 BP neural network diagnosis fault result

Ideal output	The output of network	Diagnosis style
[0.00 0.00 0.00 0.00 0.00 0.00 0.00]	(0.0000 -0.0000 -0.0000 0.0000 - 0.0000 0.000 -0.0000)	Normal
[0.00 0.75 0.00 0.00 0.00 0.00 0.00]	(0.0024 0.7489 0.0053 0.0010 0.0012 -0.0037 -0.0071)	Valve clogging
[0.00 0.00 0.50 0.00 0.00 0.00 0.00]	(0.0003 -0.0006 0.5005 -0.0001 0.0022 0.0024 0.0016)	Valve or valve seat sedimentation
[0.00 0.00 0.00 0.75 0.00 0.00 0.00]	(-0.0097 0.0047 0.0011 0.7547 0.0039 -0.0068 -0.0078)	Valve or valve seat erosion
[0.00 0.00 0.00 0.00 0.75 0.00 0.00]	(0.0033 0.0051 0.0007 0.0007 0.7515 -0.0030 -0.0041)	Increase of valve friction
[0.00 0.00 0.00 0.00 0.00 0.75 0.00]	(-0.0067 0.0058 0.0046 0.0014 0.0035 0.7606 0.0038)	External leakage (leaky bushing, covers, terminals)
[0.00 0.00 0.00 0.00 0.00 0.00 0.25]	(0.0066 0.0008 0.0001 -0.0002 0.0039 0.0022 0.2321)	Internal leakage (valve tightness)
[0.00 0.75 0.00 0.00 0.75 0.00 0.00]	(-0.0000 0.7531 0.0028 0.0010 0.0022 0.7488 0.0017)	Valve clogging or increase of valve friction

It is observed that these parameters describe the main features of the condition of valve and are indicative of the dynamic and static characteristics of the response. With one fault or multi-fault occur in process, the network can detect, diagnosis and isolate them. The simulation experimental results prove the BP neural network has the capability to detect and identify various magnitudes of the faults of interest and can isolate multiple faults. This proposed approach is practical to implement.

85.5 Conclusions

This chapter has investigated a BP neural-network-based scheme for detection and identification of pneumatic control valve faults. The faults of interest were various. The specific values of six estimable parameter were observed to depend upon both the magnitude and type of fault. For each operating condition, the performance parameters formed a discriminatory fault signature that was subsequently learned by a multilayer feedforward neural network with the goal of successfully detecting and identifying the faults. The experimental results proved that the trained network has the capability to detect and identify the various magnitudes of the faults of interest, as they occur singly. This information is vital to a process valve condition monitoring strategy since it can be used to detect problems and assist in the timely scheduling of repairs to the valve, before a severe failure occurs. Future work in this area includes training the network to discriminate against other control valve failure modes. The scheme presented here can be used directly to perform the detection and isolation of control valve faults.

References

1. Megahed AI, Malik OP (1998) Synchronous generator internal fault computation and experimental verification. *IEEE Proc Gener Trans Distrib* 145(5):604–610
2. Kulig TS, Buchley GW et al (1986) A new approach to determine transient winding and damper currents in case of internal and external faults and abnormal operation. *Fundamentals, IEEE/PES winter meeting, New Orleans, 1986*
3. Gonzalez AJ (1986) An expert system for on-line diagnosis of turbine generators. *International conference on large high voltage electric systems, 1986, Session*
4. Karpenko M, Sepehri N, Scuse D (2003) Diagnosis of process valve actuator faults using a multilayer neural network. *Control Eng Pract* 11:1289–1299
5. Paten K, Parisini T (2003) Dynamic neural networks for actuator fault diagnosis: application to the damadics benchmark problem. In: *Proceedings of IFAC symposium SAFEPROCESS*, pp 1077–1082
6. Michal B, Ron P et al (2006) Introduction to the DAMADICS actuator FDI benchmark study. *Control Eng Pract* 14:577–596
7. Pawel Syfert M, Jegorov S (2003) On-line actuator diagnosis based on neural models and fuzzy reasoning: the damadics benchmark study. In: *Proceeding of IFAC symposium safe process*, pp 1083–1088
8. Marcu T, Mirea L, Frank PM (1999) Development of dynamic neural networks with application to observer-based fault detection and isolation. *Intern J Appl Math Comput Sci* 9(3):547–570
9. Choi YS, Noah ST (1992) Response and stability analysis of piecewise linear oscillators nuer multi-forcing frequencies. *Nonl Dyn* 3:105–121

Chapter 86

Optimal Detection of Distributed Target with Fluctuating Scatterers

Tao Jian

Abstract In the non-Gaussian clutter modeled as a spherically invariant random vector, the optimal detection of a distributed target is addressed with fluctuating scatterers in high resolution radar scenarios, by exploiting the generalized likelihood ratio test design procedure and the binary integrator. The formula relating the false alarm probability to the detection threshold is given, which implies the constant false alarm rate property with respect to both the clutter covariance matrix structure and the clutter power level. Moreover, the optimal detection parameter is also obtained for a distributed target with fluctuating scatterers. Finally, the performance assessment conducted by Monte Carlo simulation confirms the effectiveness of the proposed detectors.

Keywords Optimal detection · Distributed target · Fluctuating scatterer · Binary integrator

86.1 Introduction

The point-like target detection against Gaussian clutter for the traditional low-resolution radar has been addressed partly in [1]. However, a high-resolution radar can resolve a target into a number of scatterers, which is referred to as a distributed target [2, 3]. Increasing the radar range resolution can reduce the amount of energy per cell backscattered by distributed clutter and can enhance the radar detection

T. Jian (✉)

Research Institute of Information Fusion, Naval Aeronautical and Astronautical University, 264001 Yantai, China
e-mail: iamjiantao@yahoo.com.cn

performance largely by appropriate detection strategies [4, 5]. Whereas, at the higher range resolution, the radar system receives target-like spikes that result in non-Gaussian observations, which can be suitably modeled by a spherically invariant random vector (SIRV) [6–8].

In this work, the optimal detection of a distributed target is addressed with fluctuating scatterers, by exploiting the generalized likelihood ratio test (GLRT) design procedure [9] and the binary integrator (BI).

86.2 Problem Formulation

It is assumed that data are collected from N sensors and the problem of detecting the presence of a target across K range cells $z_t s$, $t = 1, \dots, K$, is dealt with. It is supposed that the possible target is completely contained within those data [10]. Herein, the clutter-dominant environment is considered, and the internal noise is ignored. Hence the detection problem can be formulated as the following binary hypotheses test

$$\begin{aligned} H_0 : z_t &= c_t, & t &= 1, \dots, K \\ H_1 : z_t &= \alpha_t \mathbf{p} + c_t, & t &= 1, \dots, K \end{aligned} \quad (86.1)$$

where \mathbf{p} denotes the normalized steering vector, such that $\mathbf{p}^H \mathbf{p} = 1$ ($(\cdot)^H$ implies conjugate transpose), and the $\alpha_t s$, $t = 1, \dots, K$ are unknown parameters accounting for both the target and the channel effects. Note that, for the uniform linear array, $\mathbf{p} = (1, e^{j\phi}, e^{j2\phi}, \dots, e^{j(N-1)\phi})^T / \sqrt{N}$, where ϕ denotes a constant phase shifting and $(\cdot)^T$ represents transpose.

The clutter returns are modeled as a SIRV for representing non-Gaussian clutter [6]. Thus the N -dimension clutter vector \mathbf{c}_t at range t can be given by

$$\mathbf{c}_t = \sqrt{\tau_t} \cdot \boldsymbol{\eta}_t, \quad t = 1, \dots, K + R \quad (86.2)$$

where $\boldsymbol{\eta}_t = (\eta_t(1), \eta_t(2), \dots, \eta_t(N))^T$, $\eta_t(n) s$, $n = 1, \dots, N$ are zero-mean complex circular Gaussian random variables (RVs) with variance equal to one, and the texture component τ_t is a semipositive real RV with probability distribution f_{τ_t} , which is called mixing distribution. Moreover, $\boldsymbol{\eta}_t$ and τ_t are assumed to be independent from range cell to range cell. Here an $N \times N$ clutter covariance matrix structure $\boldsymbol{\Sigma}$ associated with η_t, s , $t = 1, \dots, K + R$ is defined as

$$\boldsymbol{\Sigma} = E\{\boldsymbol{\eta}_t \boldsymbol{\eta}_t^H\} \quad (86.3)$$

where $\boldsymbol{\Sigma}$ is the positive definite and Hermitian matrix.

It is assumed that the underlying mixing distribution f_{τ_t} is unknown. Thereby each component of clutter vector \mathbf{c}_t is modeled as conditionally Gaussian with the unknown variance τ_t . It is also assumed that α_t is unknown but \mathbf{p} is known.

According to the previous assumptions, the probability density function (PDF) of \mathbf{z}_t s, $t = 1, \dots, K$ under each hypothesis is given by [11]

$$f(\mathbf{z}_t | \Sigma, \tau_t, \mathbf{H}_0) = \frac{1}{\pi^N \tau_t^N \det(\Sigma)} \times \exp \left[-\frac{1}{\tau_t} \mathbf{z}_t^H \Sigma^{-1} \mathbf{z}_t \right] \quad (86.4)$$

$$f(\mathbf{z}_t | \Sigma, \alpha_t, \tau_t, \mathbf{H}_1) = \frac{1}{\pi^N \tau_t^N \det(\Sigma)} \times \exp \left[-\frac{1}{\tau_t} (\mathbf{z}_t - \alpha_t \mathbf{p})^H \Sigma^{-1} (\mathbf{z}_t - \alpha_t \mathbf{p}) \right] \quad (86.5)$$

where $\det(\cdot)$ denotes determinant.

According to the Neyman-Pearson criterion, the optimal solution to the hypotheses testing problem (86.1) is the likelihood ratio test, but for the case at hand, it cannot be implemented due to total ignorance of the parameters $\{\alpha_t | t = 1, \dots, K\}$ and $\{\tau_t | t = 1, \dots, K + R\}$. We resort, instead, to GLRT-based decision schemes [9].

86.3 Binary Integrator

In this section, with the known matrix Σ , the BI is introduced.

To simplify the analysis, only one scatterer is supposed to occupy one resolution cell. In most cases of target scattering, the scatterers may occupy only a fraction of the K range cells. Furthermore, the echo amplitudes of range cells occupied by target scatterers are significantly greater than that of range cells with clutter only.

The traditional detection strategies of a point-like target only utilizes target energy in a single range cell, and may fail for distributed targets. In order to make the best of target energy in all K resolution cells of the range extent of target, we can accumulate target scatterers by BI, after single target scatterer detection in each range cell.

With known Σ , the derivation of target scatterer detection in single range cell is begun by writing the GLRT as follows [3]

$$\frac{\max_{\tau_t} \max_{\alpha_t} f(\mathbf{z}_t | \Sigma, \alpha_t, \tau_t, \mathbf{H}_1)}{\max_{\tau_t} f(\mathbf{z}_t | \Sigma, \tau_t, \mathbf{H}_0)} \quad (86.6)$$

By replacing the unknown parameters with their maximum likelihood estimates under each hypothesis, the GLRT statistic for target scatterer detection in single range cell can be denoted as

$$\lambda_1(\mathbf{z}_t) = -N \ln \left[1 - \frac{|\mathbf{p}^H \Sigma^{-1} \mathbf{z}_t|^2}{(\mathbf{z}_t^H \Sigma^{-1} \mathbf{z}_t) (\mathbf{p}^H \Sigma^{-1} \mathbf{p})} \right] \quad (86.7)$$

Therefore, the first detection threshold T_1 of BI for the given first false alarm probability P_{fa1} can be given by [5]

$$P_{fa1} = \exp[-(N - 1)T_1/N] \tag{86.8}$$

Set

$$d_t = \begin{cases} 1, & \text{if } \lambda_1(\mathbf{z}_t) > T_1 \\ 0, & \text{otherwise} \end{cases}, t = 1, \dots, K \tag{86.9}$$

Herein, the BI (or M/K detector) is designed to detect all scatterers for a distributed target. The detection decision is based on at least M threshold crossings, out of K observations [12], where K is the integrated cell number and M ($1 \leq M \leq K$) is the threshold of BI. The first threshold level of a single range cell must be determined, for the given M and K , to produce the desired integrated false alarm probability P_{fa2} . Since the choice of M affects this result, each M requires a different first threshold. It is necessary to determine an optimal or nearly optimal value for the parameter M .

The d_t s, $t = 1, \dots, K$ are inputted into the M/K detector. Moreover, the hypothesis that a distributed target is present is tested as follows

$$\lambda_2 = \sum_{t=1}^K d_t \underset{H_0}{\overset{H_1}{\geq}} T_2 \tag{86.10}$$

where the second detection threshold of BI is given by

$$T_2 = M, 1 \leq M \leq K \tag{86.11}$$

Accordingly, the second false alarm probability P_{fa2} is simply expressed as

$$P_{fa2} = \sum_{k=M}^K P_{fa1}^k (1 - P_{fa1})^{K-k} K! / (k!(K - k)!) \tag{86.12}$$

where P_{fa1} is the first false alarm probability in (86.8). For the given overall false alarm probability $P_{fa} = P_{fa2}$, M and K , the first false alarm probability P_{fa1} can be determined from (86.12) either iteratively or approximately [12]. Finally, the first threshold T_1 can be computed from (86.8) for the given P_{fa1} . It is shown that both of two detection thresholds T_1 and T_2 are independent of Σ and τ_t , $s, t = 1, \dots, K$. It implies that, with known Σ , the BI is constant false alarm rate (CFAR) with respect to both the clutter covariance matrix structure and the clutter power level.

86.4 Optimal Detection Threshold

In this section, the optimal threshold M (M_{opt}) of BI is calculated for distributed target detection.

The matrix Σ is assumed to be Toeplitz. The clutter samples were generated assuming an exponential correlation structure, i.e., the matrix Σ has elements [8]

Table 86.1 Values of $P_{\text{fa}1}$ with different M for $P_{\text{fa}}=10^{-4}$ and $K=15$

M	1	2	3	4	5
$P_{\text{fa}1}$	6.6670×10^{-6}	9.8006×10^{-4}	6.1472×10^{-3}	1.70846×10^{-2}	3.38243×10^{-2}
M	6	7	8	9	10
$P_{\text{fa}1}$	5.60672×10^{-2}	8.3561×10^{-2}	1.16188×10^{-1}	1.53997×10^{-1}	1.97231×10^{-1}
M	11	12	13	14	15
$P_{\text{fa}1}$	2.46399×10^{-1}	3.02418×10^{-1}	3.66965×10^{-1}	4.4346×10^{-1}	5.4117×10^{-1}

$$[\Sigma]_{i,j} = \gamma^{|i-j|}, 1 \leq i, j \leq N \quad (86.13)$$

where γ is the one-lag correlation coefficient.

The distribution f_τ is modeled as a Gamma distribution with the following PDF

$$f_\tau(x) = (L/b)^L x^{L-1} e^{-(L/b)x} / \Gamma(L), x \geq 0 \quad (86.14)$$

where $\Gamma(\cdot)$ is the gamma function, b indicates the mean of the distribution, and L controls the deviation from Gaussian statistics.

It is assumed that each of the K range cells has a clutter component and each of the h_0 target range cells has a signal component. The quantity σ_s^2/σ_c^2 indicates the average signal-to-clutter ratio (SCR) per range cell taken over K range cells, where σ_s^2 and σ_c^2 indicate the average signal and clutter power per range cell respectively. The returns from target scatterers are modeled as independent and identically distributed (IID) zero-mean complex circular Gaussian RVs with the variance σ_s^2 . It means that the target amplitude fluctuates with Rayleigh law over range cells. Moreover, the input SCR of distributed target detectors is defined as [10]

$$SCR = \sigma_s^2 \mathbf{p}^H \Sigma^{-1} \mathbf{p} / \sigma_c^2 \quad (86.15)$$

For the given P_{fa} , M and K , the first detection threshold of BI can be computed from (86.8). Moreover, $P_{\text{d}s}$ for all detectors are estimated based on Monte Carlo simulation. For analytical convenience, the $P_{\text{fa}1}$ with different M is given for $P_{\text{fa}} = 10^{-4}$ and $K = 15$ in Table 86.1.

For only one scatterer is supposed to occupy one resolution cell, we just consider the values of M for $M \leq h_0$. For the space consideration, herein, we only present some representative M_{opt} for $h_0 = 2, 5, 8$. In addition, M_{opt} for other values of $h_0 \leq K$ has the similar rules and can be calculated from the resultant equation of M_{opt} with respect to h_0 .

In Fig. 86.1, the plots of P_{d} versus SCR_{in} of BI ($M = 1, 2$) are given for $N = 2, L = 1, \gamma = 0, K = 15$ and $h_0 = 2$. It is observed that, the BI with $M = 2$ outperforms the BI with $M = 1$. We determine that $M_{\text{opt}} = 2$ for $h_0 = 2$. With the other preferences same as Figs. 86.1 and 86.2 refers to the detection performance of BI ($M = 1, 2, 3, 4, 5$) for $h_0 = 5$. It highlights that the BI with $M = 1$ performs worst, and the performance gets better as M increases. Moreover,

Fig. 86.1 P_d versus SCR of BI for $N=2, L=1, \gamma=0, K=15, M=1,2, h_0=2, P_{fa}=10^{-4}$

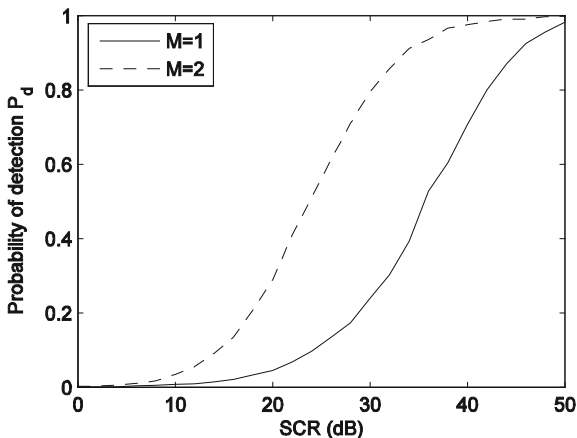
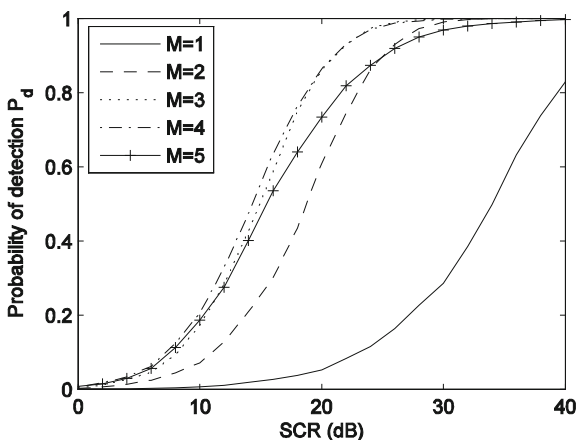


Fig. 86.2 P_d versus SCR of BI for $N=2, L=1, \gamma=0, K=15, M=1,2,3,4,5, h_0=5, P_{fa}=10^{-4}$



the BI with $M = 4$ performs best, but the performance gets worse as M increases for $M \geq 4$. Hence, we determine that $M_{opt} = 4$ for $h_0 = 5$.

Furthermore, with the other preferences same as Figs. 86.1, 86.3 refers to the detection performance of BI ($M = 1, 2, 3, 4, 5, 6, 7, 8$) for $h_0 = 8$. It is indicated that, for $M \leq 5$, the performance gets better as M increases, however, for $M \geq 5$, the performance gets worse as M increases. Thereby we determine that $M_{opt} = 5$ for $h_0 = 8$. In like manner, we also calculate the optimal M for other values of $h_0 \leq K$, and the values of M_{opt} with different h_0 are given for $K = 15$ in Table 86.2. It is observed that M_{opt} is a monotonically increasing function of h_0 .

According to Table 86.2, it is concluded that, for $1 < h_0 \leq K, M_{opt}$ satisfies

$$M_{opt} = \text{round}(h_0/2 + 1) \tag{86.16}$$

where $\text{round}(\cdot)$ denotes rounding the parameter to the nearest integer.

Fig. 86.3 P_d versus SCR of BI for $N = 2, L = 1, \gamma = 0, K = 15, M = 1, \dots, 8, h_0 = 8, P_{fa} = 10^{-4}$

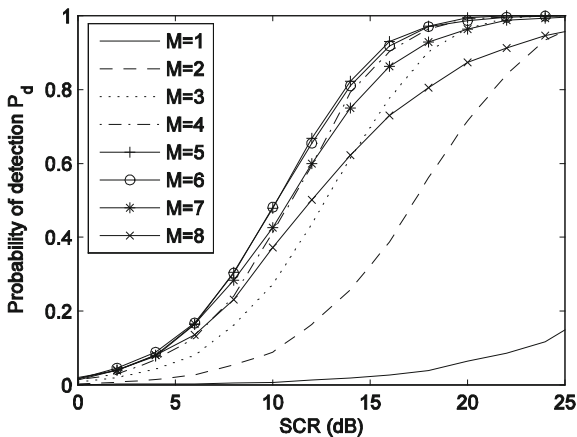


Table 86.2 Values of M_{opt} with different h_0 for $K = 15$

h_0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
M_{opt}	1	2	3	3	4	4	5	5	6	6	7	7	8	8	9

86.5 Conclusions

The optimal detection of a distributed target is addressed with fluctuating scatterers, by exploiting the GLRT design procedure and the BI. The formula relating the false alarm probability to the detection threshold of BI implies the CFAR property with respect to both the clutter covariance matrix structure and the clutter power level. Moreover, the optimal parameter of BI is also obtained by performance assessment.

Acknowledgments This work was supported by National Natural Science Foundation of China (61032001) and Scientific Research Foundation of Naval Aeronautical and Astronautical University for Young Scholars (HYQN201013).

References

1. He Y, Guan J, Peng YN et al (1999) Radar automatic detection and constant false alarm processing. Tsinghua University Press, Beijing (in Chinese)
2. Wehner DR (1995) High-resolution radar. Artech House, Boston
3. Jian T, Su F, He Y et al (2010) Target detection of high-resolution radar in non-Gaussian clutter. In: Proceedings of the 10th international conference on signal processing, vol 3, pp 2047–2050

4. Bandiera F, De Maio A, Greco AS et al (2007) Adaptive radar detection of distributed targets in homogeneous and partially homogeneous noise plus subspace interference. *IEEE Trans Signal Process* 55:1223–1237
5. He Y, Jian T, Su F et al (2010) Novel range-spread target detectors in non-Gaussian clutter. *IEEE Trans Aerosp Electron Syst* 46:1312–1328
6. Yao K (1973) A representation theorem and its applications to spherically invariant random processes. *IEEE Trans Inf Theory* 19:600–608
7. Rangaswamy M, Weiner M, Ozturk A (1995) Computer generation of correlated non-Gaussian radar clutter. *IEEE Trans Aerosp Electron Syst* 31:106–116
8. He Y, Jian T, Su F et al (2010) Adaptive detection application of covariance matrix estimation for correlated non-Gaussian clutter. *IEEE Trans Aerosp Electron Syst* 46:2108–2117
9. Van Trees HL (2001) *Detection, estimation, and modulation Theory, Part I: detection, estimation, and linear modulation theory*. Wiley, New York
10. Gerlach K (1999) Spatially distributed target detection in non-Gaussian clutter. *IEEE Trans Aerosp Electron Syst* 35:926–934
11. Papoulis A, Pillai SU (2002) *Probability, random variables and stochastic processes*. McGraw-Hill, New York
12. Shnidman DA (1998) Binary integration for swerling target fluctuations. *IEEE Trans Aerosp Electron Syst* 34:1043–1053

Chapter 87

Output Feedback Control for an Active Heave Compensation System

Jia-Wang Li, Tong Ge and Xu-Yang Wang

Abstract To reduce the adverse effect of the unexpected vessel heave motion on the response of underwater payloads, a control strategy is presented for an active heave compensation system using an electro-hydraulic system driven by a double-rod actuator. An adaptive observer is designed to estimate the unmeasured system states and the unmodeled forces. An observer is also proposed to asymptotically reconstruct the vessel motion. By using these observers, the Lyapunov's direct method and backstepping technique, an output feedback controller is proposed to force the heave compensation error to converge to a small bounded area around the origin. Simulations illustrate the effectiveness of the proposed control scheme.

Keywords Heave compensation · Output feedback control · Adaptive observer · Backstepping

87.1 Introduction

In offshore installations and deep sea marine operations, one of the most important issues is how to provide safety and high operability of payloads. This means that

J.-W. Li (✉) · T. Ge · X.-Y. Wang
Institute of Underwater Engineering, Shanghai Jiaotong University,
Shanghai 200240, China
e-mail: jwli@sjtu.edu.cn

T. Ge
e-mail: tongge@sjtu.edu.cn

X.-Y. Wang
e-mail: wangxuyang@sjtu.edu.cn

the payload motion should be kept unaffected by the supporting vessel motion, since waves, wind and ocean currents can easily cause an unexpected motion of the vessel, which in turn has adverse effects on the cable connecting between the payload and the vessel. The unexpected horizontal motion of the vessel is often controlled by a dynamic positioning system. To reduce the adverse effects in the vertical direction, active heave compensation systems are usually been used.

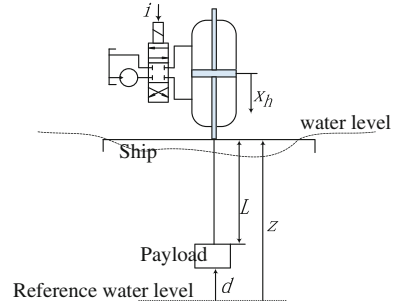
The control problems for active heave compensation systems have been addressed by numerous researchers in the past. In [1], a linear control scheme was presented for an active heave compensation system. However, the authors assumed that the vessel motion due to waves was known, which is generally hard to be accomplished in practice. To remove this restriction, the jointed problems of wave synchronization and heave compensation were studied in [2, 3]. In both works, the authors assumed that only the heave acceleration was measured and its integrals were obtained via high-pass filters. In [4], a nonlinear controller for an electro-hydraulic system driven by a double-rod cylinder was proposed, whereas the vessel motion and acting force were estimated by disturbance observers. The authors in [5] designed an autopilot for the autonomous landing of a vertical take off and landing vehicle on a ship oscillating in the vertical direction, which was based on the approach introduced in [6]. An improved work was presented in [7], where the reconstruction of the wave disturbances was accomplished by using the adaptive external models proposed in [8]. For the system with time delays in sensors and actuators, a prediction algorithm was designed in [9] to predict the vessel motion.

This chapter focuses on active heave compensation control of an electro-hydraulic system driven by a double-rod actuator. An adaptive observer is proposed to estimate the system states and the force acting on the cable. By assuming that the vessel motion can be represented by a set of harmonics with known frequencies as [10], an observer is also developed to asymptotically reconstruct these harmonic signals. These observers are then implemented in the control design procedure. The control development and stability analysis are based on the Lyapunov's direct method and backstepping technique.

87.2 Problem Formulation

The active heave compensation system under consideration is depicted in Fig. 87.1. This system consists of an electro-hydraulic system driven by a double rod actuator, which is fixed to the vessel. The payload connects to the piston of the hydraulic system via a cable and a ball joint, where the cable is assumed rigidly. In Fig. 87.1, the reference water level is a horizontal line fixed to the earth. The heave motion of the vessel with respect to this level is denoted by z . The position of the piston with respect to the vessel is represented by x_h . L and d denote respectively the cable length and the desired position of the payload with respect to the reference water level.

Fig. 87.1 Sketch of an active heave compensation system



Following [4], the scaled model of the active heave compensation system can be written as

$$\begin{aligned}\ddot{x}_h &= \theta_1 \bar{P} - \theta_2 \dot{x}_h + \theta_3 f(t) \\ \dot{\bar{P}} &= -\theta_4 \dot{x}_h - \theta_5 \bar{P} + \theta_6 \bar{x}_v \\ \dot{\bar{x}}_v &= -\theta_7 \bar{x}_v + \theta_8 i\end{aligned}\quad (87.1)$$

where $\theta_i (i = 1, \dots, 8)$ are model parameters, \bar{P} and \bar{x}_v denote the scaled load pressure and spool displacement of the servo valve respectively, i is the control current input of the servo valve, and $f(t)$ represents the resulting forces in the vertical direction acting on the piston. Generally, it is hard to derive the explicit expression of $f(t)$ since it depends on too many factors. Hence, following [4], we treat it as a disturbance. On the other hand, it should be noted that only part of the states of the system (87.1) can be practically measured. To facilitate the control development, we assume that the velocity of the piston, \dot{x}_h , is the measured output.

In this chapter, the control objective is to make the heave velocity of the payload with respect to the reference water level, $\dot{d}(t)$ (see Fig. 87.1), to track a desired velocity reference. On the other hand, one can see from Fig. 87.1 that

$$d(t) = z(t) + x_h(t) - L \quad (87.2)$$

where the vessel motion $z(t)$ can be generally decomposed into a set of harmonic oscillations and an additional slow time-varying term, i.e.,

$$z(t) = \underbrace{\sum_{i=1}^n A_i \sin(\omega_i t + \phi_i)}_{w(t)} + v(t) \quad (87.3)$$

where $n \geq 1$, A_i , ω_i and ϕ_i are the amplitude, frequency and phase of the i th mode respectively. In this chapter, we assume that the term $v(t)$ is a constant.

Then, from (87.2) and (87.3), one can see that the control objective is equivalent to the problem of stabilizing the following heave tracking error

$$e(t) = \dot{w}(t) + \dot{x}_h(t) - c(t) \quad (87.4)$$

where $c(t)$ is the desired speed, and without loss of generality, we set it as a constant.

Note that (87.4) is generally not available for feedback since $\dot{w}(t)$ is usually unknown, whereas its derivative, $\dot{w}(t)$, is assumed to be measured via high precision accelerometers. To overcome this difficulty, we will also design an observer for $w(t)$ in the next section. On the other hand, to facilitate the observer design procedure, the following assumptions for $w(t)$ and $f(t)$ are made.

Assumption 1 The frequencies of the harmonics, $\omega_i (i = 1, \dots, n)$, are known, whereas the amplitudes and phases of the harmonics are not.

Assumption 2 The variables $f(t)$ is globally bounded and there exists a nonnegative constant σ_f such that the first order time derivative of $f(t)$ satisfies

$$|\dot{f}(t)| \leq \sigma_f. \quad (87.5)$$

Furthermore, since the heave position is left uncontrolled, the state x_h in (87.1) is negligible, and thus, the entire system can be reduced to a third-order system given by the following state space form

$$\begin{aligned} \dot{x} &= Ax + Bu + Df(t) \\ y &= Cx \end{aligned} \quad (87.6)$$

with $x = [\dot{x}_h, \bar{P}, \bar{x}_v]^\top$, $B = [0, 0, \theta_8]^\top$, $C = [1, 0, 0]$, $D = [\theta_3, 0, 0]^\top$ and

$$A = \begin{bmatrix} -\theta_2 & \theta_1 & 0 \\ -\theta_4 & -\theta_5 & \theta_6 \\ 0 & 0 & -\theta_7 \end{bmatrix}. \quad (87.7)$$

It can be seen that the pair (A, B) is controllable, and (A, C) is observable.

87.3 Adaptive Observer Design

In this section, an adaptive observer will be developed for the unmeasured states x and $f(t)$. And then, another observer for $w(t)$ will be presented to asymptotically recover $w(t)$ and its any order time derivatives. The first observer scheme is mainly improved on [4].

To estimate the states x , from (87.6), we interpret the following observer

$$\dot{\hat{x}} = A\hat{x} + Bu + D\hat{f}(t) + K_1(y - C\hat{x}) \quad (87.8)$$

where \hat{x} is the estimate of x , $K_1 = [k_{11}, k_{12}, k_{13}]^\top$ is such that the matrix $(A - K_1C)$ is Hurwitz, $\hat{f}(t)$ denotes the estimate of $f(t)$ and is given by

$$\hat{f} = C(x - \xi), \quad \dot{\xi} = A\hat{x} + Bu + D\hat{f} + K_2(y - C\hat{x}) \quad (87.9)$$

where $K_2 = [k_{21}, k_{22}, k_{23}]^\top$ is a vector of control gains to be determined later.

Let $\tilde{x} = x - \hat{x}$ and $\tilde{f} = f - \hat{f}$ be the estimation errors. Then, from (87.6), (87.8) and (87.9), we yield the observation error dynamics as follows

$$\begin{aligned} \dot{\tilde{x}} &= (A - K_1C)\tilde{x} + D\tilde{f} \\ \dot{\tilde{f}} &= -C(A + K_2C)\tilde{x} - CD\tilde{f} + \dot{f}. \end{aligned} \quad (87.10)$$

Proposition 1 Consider the observation error dynamics (87.10). Under Assumption 2, if there exists a symmetric, positive-definite matrix M satisfying

$$\begin{bmatrix} A - K_1C & D \\ -C(A + K_2C) & -CD \end{bmatrix}^\top M + M \begin{bmatrix} A - K_1C & D \\ -C(A + K_2C) & -CD \end{bmatrix} \leq \lambda I, \quad M \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

with λ a positive constant and I an identity matrix, then the estimation errors \tilde{x} and \tilde{f} are globally convergent to a bounded area around the origin.

Proof Consider the following Lyapunov function $V_0 = [\tilde{x}^\top, \tilde{f}^\top]M[\tilde{x}^\top, \tilde{f}^\top]^\top$. Differentiating it along the solutions of (87.10), we yield

$$\dot{V}_0 \leq -\lambda(\tilde{x}^\top, \tilde{x} + \tilde{f}^2) + 2\tilde{f}\dot{f}.$$

Due to the Young's inequality and $|\dot{f}| \leq \sigma_f$, we have

$$\dot{V}_0 \leq -\lambda\tilde{x}^\top\tilde{x} - \lambda_f\tilde{f}^2 + \varepsilon_1^{-1}\sigma_f^2 \leq -\eta V_0 + \varepsilon_1^{-1}\sigma_f^2$$

where $\lambda_f = \lambda - \varepsilon_1$ and $\eta = \lambda_f/\bar{l}$, ε_1 is a positive constant such that $\lambda_f > 0$ and \bar{l} is the maximum eigenvalue of the matrix M . Then, one can see that $V_0(t)$ globally converges to a ball around zero with the radius $\sigma_f^2/\varepsilon_1\eta$. As a consequence, the estimation errors (\tilde{x}, \tilde{f}) converge to a ball centered at the origin with the radius $\sigma_f/\sqrt{\varepsilon_1\eta\bar{l}}$, with, where \underline{l} is the minimum eigenvalue of M .

In the rest of this section, we will design an observer for $w(t)$. At first, it worth noting that the dynamics of $w(t)$ can be expressed as

$$\dot{W} = S_w W, \quad \dot{w} = C_w W \quad (87.11)$$

with $W \in \mathbb{R}^{2n}$, $S_w = \text{block diag}[S_1, \dots, S_n]$ and $C_w = [C_1, \dots, C_n]$, where $S_i = \begin{bmatrix} 0 & \omega_i \\ -\omega_i & 0 \end{bmatrix}$, $C_i = [\omega_i^2, 0]$, $\forall i \in [1, n]$. Note that S_w and C_w are known under Assumption 1, and the pair (S_w, C_w) is observable. Since \dot{w} is assumed to be a known variable, then the following observer is designed

$$\dot{\hat{W}} = S_w \hat{W} + K_3(\ddot{w} - C_w \hat{W}), \quad \hat{w} = C_w \hat{W} \quad (87.12)$$

where \hat{W} and \hat{w} are the estimates, $K_3 \in \mathbb{R}^{2n}$ is such that the matrix $(S_w - K_3 C_w)$ is Hurwitz. This observer can be used to reconstruct any order time derivative of $w(t)$.

Proposition 2 *The output of system (87.12) defined as*

$$\hat{w}^{(i)} = C_w S_w^{i-2} \hat{W}, \quad \forall i \geq 0 \quad (87.13)$$

yields converging estimate of the i th order derivative of $w(t)$.

Proof From (87.11), one can see that the i th order derivative of $w(t)$ can be given as $w^{(i)} = C_w S_w^{i-2} W$. Then, we have

$$w^{(i)} - \hat{w}^{(i)} = C_w S_w^{i-2} (W - \hat{W})$$

Let $\tilde{W} = W - \hat{W}$ be the estimation error and choose the Lyapunov function candidate $V_1 = \tilde{W}^\top \tilde{W}$. Differentiating V_1 along the solutions of (87.11) and (87.12), we have

$$\dot{V}_1 = \tilde{W}^\top (S_w - K_3 C_w) \tilde{W}$$

which implies that $\lim_{t \rightarrow \infty} \|\tilde{W}\| = 0$ due to the fact that $(S_w - K_3 C_w)$ is Hurwitz. Then, one can obtain $\lim_{t \rightarrow \infty} |w^{(i)} - \hat{w}^{(i)}| = \lim_{t \rightarrow \infty} |C_w S_w^{i-2} (W - \hat{W})| = 0$.

87.4 Controller Design

At first, we want to note that the structure of the system (87.8) allows us to use the Lyapunov's direct method and backstepping technique for the controller design procedure, which can be divided into three steps.

Step 1.

Consider the Lyapunov function candidate $V_2 = 0.5\gamma_1 e^2$, where γ_1 is a positive constant to be chosen later. The time derivative of V_2 is given by

$$\dot{V}_2 = \gamma_1^e (-\theta_2 \hat{x}_1 + \theta_1 \hat{x}_2 + \theta_3 \hat{f} + K_{11} C \tilde{x} + \hat{w}). \quad (87.14)$$

Let $\hat{x}_{2e} = \hat{x}_2 - \alpha_2$ be the error state with α_2 a virtual control of \hat{x}_2 . Choosing α_2 as

$$\alpha_2 = \theta_1^{-1} (-k_e \bar{e} + \theta_2 \hat{x}_1 - \theta_3 \hat{f} - \hat{w}) \quad (87.15)$$

with k_e a positive constant, and substituting (87.15) into (87.14), we have

$$\dot{V}_2 = -k_e\gamma_1\bar{e}^2 + \gamma_1\bar{e}^2(\theta_1\hat{x}_{2e} + k_{11}C\bar{x}). \quad (87.16)$$

Step 2.

To regulate the new error \hat{x}_{2e} , we choose the Lyapunov function candidate $V_3 = V_2 + 0.5\gamma_2\hat{x}_{2e}^2$ with $\gamma_2 > 0$. The dynamics of V_3 satisfies

$$\begin{aligned} \dot{V}_3 = & -k_e\gamma_1\bar{e}^2 + \gamma_1\bar{e}(\theta_1\hat{x}_{2e} + k_{11}C\bar{x}) \\ & + \gamma_2\hat{x}_{2e}(-\theta_4\hat{x}_1 - \theta_5\hat{x}_2 + \theta_6\hat{x}_3 + k_{12}C\bar{x} - \dot{\alpha}_2). \end{aligned} \quad (87.17)$$

Let $\hat{x}_{3e} = \hat{x}_3 - \alpha_3$ be the virtual control error, and α_3 is given by

$$\alpha_3 = \theta_6^{-1}(-k_x\hat{x}_{2e} + \theta_4\hat{x}_1 + \theta_5\hat{x}_2 + \zeta_1) \quad (87.18)$$

with k_x a control gain to be determined later, and

$$\zeta_1 = -\theta_1^{-1}[k_e(-\theta_2\hat{x}_1 + \theta_1\hat{x}_2 + \theta_3\hat{f} + \hat{w}) - \theta_2\hat{x}_1 + \hat{w}]. \quad (87.19)$$

Substituting (87.18) and (87.19) into (87.17) and according to the expression of $\dot{\alpha}_2$, we yield

$$\begin{aligned} \dot{V}_3 = & -k_e\gamma_1\bar{e}^2 + \gamma_1\bar{e}(\theta_1\hat{x}_{2e} + k_{11}C\bar{x}) \\ & - k_x\gamma_2\hat{x}_{2e}^2 + \gamma_2\hat{x}_{2e}(\theta_6\hat{x}_{3e} + \bar{k}_{12}C\bar{x} + \theta_1^{-1}\theta_3\dot{f}). \end{aligned} \quad (87.20)$$

with $\bar{k}_{12} = k_e k_{11} \theta_1^{-1} + k_{12}$.

Step 3.

This is the final step. To regulate the error state \hat{x}_{3e} , we consider the Lyapunov function $V_4 = V_3 + 0.5\gamma_3\hat{x}_{3e}^2$ with $\gamma_3 > 0$. The time derivative of V_4 is

$$\dot{V}_4 = \dot{V}_3 + \gamma_3\hat{x}_{3e}(-\theta_7\hat{x}_3 + \theta_8i + k_{13}C\bar{x} - \dot{\alpha}_3). \quad (87.21)$$

Then, we choose the input i as

$$i = \theta_8^{-1}(-k_i\hat{x}_{3e} + \theta_7\hat{x}_3 - \bar{k}_{13}C\bar{x} + \zeta_2) \quad (87.22)$$

with

$$\zeta_2 = -k_x\hat{x}_{3e} + \theta_6^{-1}[\zeta_1\dot{\hat{x}}_1 + \zeta_2\dot{\hat{x}}_2 + k_x^2\hat{x}_{2e} - \theta_1^{-1}(k_e\hat{w}^{(3)} + \hat{w}^{(4)})] \quad (87.23)$$

where k_i is a control gain, $\bar{k}_{13} = k_{13} + \theta_6^{-1}k_x\bar{k}_{12}$, $\zeta_1 = \theta_4 + \theta_1^{-1}\theta_2(k_e - \theta_2)$ and $\zeta_2 = \theta_5 - k_e + \theta_2$. Substituting (87.22) and (87.23) into (87.21) and following (87.20) and the definition of α_3 , we have

$$\begin{aligned} \dot{V}_4 = & -k_e\gamma_1\bar{e}^2(\theta_1\hat{x}_{2e} + k_{11}C\bar{x}) - k_x\gamma_2\hat{x}_{2e}^2 + \gamma_2\hat{x}_{2e}^2(\theta_6\hat{x}_{3e} \\ & + \bar{k}_{12}C\bar{x} + \theta_1^{-1}\theta_3\hat{f}) - k_i\gamma_3\hat{x}_{3e}^2 + \gamma_3\hat{x}_{3e}^2(\zeta_3\hat{x} + \zeta_4\hat{f}) \end{aligned} \quad (87.24)$$

with $\zeta_3 = -\theta_2k_{11}C/\theta_2\theta_6$ and $\zeta_4 = -\theta_3(k_x + k_e + \theta_2)/\theta_2\theta_6$. Then, we can state our main result of this chapter in the following theorem.

Theorem 1 Consider the system (87.6) with the output-feedback controller (87.22) and the observers given in (87.8) and (87.12). Under Assumption 1 and 2, the heave compensation error $e(t)$ asymptotically tends to a small bounded area around the origin.

Proof From (87.8) and (87.9), we can rewrite (87.24) as

$$\begin{aligned} \dot{V}_4 = & -k_e\gamma_1\bar{e}^2 + \gamma_1\theta_1\bar{e}\hat{x}_{2e} + (k_{11}C\bar{x}) - k_x\gamma_2\hat{x}_{2e}^2 + \gamma_2\theta_6\hat{x}_{2e}\hat{x}_{3e} - k_i\gamma_3\hat{x}_{3e}^2 \\ & + (\Omega_1\bar{e} + \Omega_1\hat{x}_{2e} + \Omega_3\hat{x}_{3e})\bar{x} + (\Omega_4\bar{e} + \Omega_5\hat{x}_{2e} + \Omega_6\hat{x}_{3e})\hat{f} \end{aligned}$$

with $\Omega_i^\top \in \mathbb{R}^3 (i = 1, 2, 3)$ and $\Omega_i \in \mathbb{R} (i = 4, 5, 6)$ the appropriate vectors and scalars, respectively. By using the Young's inequalities and from Proposition 1, after a lengthy but simple calculation, we have

$$\dot{V}_4 \leq -\bar{k}_e\bar{e}^2 - \bar{k}_x\hat{x}_{2e}^2 - \bar{k}_i\hat{x}_{3e}^2 + A$$

with $\bar{k}_e = k_e\gamma_1\varepsilon_2(\gamma_1\theta_1 + 2)/2$, $\bar{k}_x = k_x\gamma_2 - \gamma_1\theta_1/2\varepsilon_2 - \varepsilon_3(\gamma_2\theta_6 + 2)/2$, $\bar{k}_i = k_i\gamma_3 - \gamma_2\theta_6/2\varepsilon_3 - \varepsilon_4$ and $A = \sigma_f^2(\varepsilon_1\eta L)^{-1}[\sum_{i=1}^3(\|\Omega_i\| + |\Omega_{i+3}|^2)/2\varepsilon_{i+1}]$, where $\varepsilon_i (i = 2, 3, 4)$ are chosen such that the control gains \bar{k}_e , \bar{k}_x and \bar{k}_i are positive. Then, following the proof of Proposition 1, one can conclude that the error states $(\bar{e}, \hat{x}_{2e}, \hat{x}_{3e})$ globally asymptotically converge to a bounded ball centered at zero with the radius $\sqrt{A/\bar{\eta}}$, where $\bar{\eta} = 2 \min(\bar{k}_e, \bar{k}_x, \bar{k}_i)/\max(\gamma_1, \gamma_2, \gamma_3)$. Furthermore, it is not hard to check that this radius can be made arbitrarily small by choosing the gains, k_e , k_x and k_i , sufficiently large.

For $e(t)$ from Proposition 1, 2 and above result, we have

$$\lim_{t \rightarrow \infty} |e(t)| \leq \lim_{t \rightarrow \infty} |\bar{e}(t)| + \lim_{t \rightarrow \infty} |x_1 - \hat{x}_1| + \lim_{t \rightarrow \infty} |\dot{w} - \hat{w}| \leq \sqrt{A/\bar{\eta}} + \sigma_f / \sqrt{\varepsilon_1\eta L}.$$

This means that the trajectory of $e(t)$ reduces to a bounded value as time goes to infinity, which completes this proof.

Fig. 87.2 The time evaluation of $e(t)$

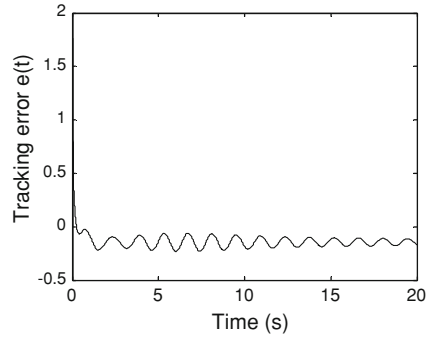
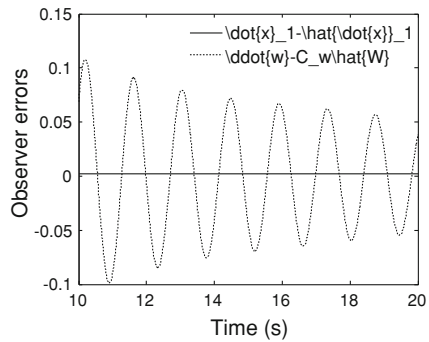


Fig. 87.3 Short presentations of the estimation errors



87.5 Simulation Results

To illustrate the effectiveness of the proposed controller and observer, we carry out some simulations in this section. The system parameters are given by: $\theta_1 = 390$, $\theta_2 = 0.04$, $\theta_3 = 0.001$, $\theta_4 = 490.75$, $\theta_5 = \theta_6 = 1.0$, $\theta_7 = 157.233$, $\theta_8 = 1.02e7$ and $f(t) = 1000 + \sin(15t)$. The desired velocity is $c = 0$. The harmonics $w(t)$ is set as

$$w(t) = \sin(4.3t - 0.4\pi) + 1.5 \sin(4.8t + 0.3\pi) + 0.7 \sin(5.3t).$$

The controller and observer gains are chosen as: $k_e = 100$, $k_x = k_i = 150$, $K_1 = [500, -490.75, 0]^T$, $K_2 = [0.04, 0,]^T$ and $K_3 = [2, 0, 4, 0, 0, 5]^T$. The initial conditions of the system and observer are selected at the origin. The simulation results are depicted in Figs. 87.2 and 87.3.

From Fig. 87.2, one can see that the tracking error $e(t)$ converges a small area around the origin as expected. The plots in Fig. 87.3 show a short time presentation of the convergence of the observation errors. It can be seen that the estimation error $\ddot{w} - \hat{C}_w \hat{W}$ is asymptotically stable and the error $\dot{x}_h - \hat{\dot{x}}_h$ does not converge to zero due to the fact that the term $f(t)$ is time-varying.

87.6 Conclusions

In this chapter, a control strategy for an active heave compensation system is presented. For the system with unknown disturbances and unmeasurable states, two observers are designed respectively to estimate the system states and asymptotically reconstruct the vessel motion represented by a set of harmonic signals. By using the Lyapunov's direct method and backstepping technique, a output-feedback tracking controller is presented. The effectiveness of the proposed control strategy is tested by means of simulations.

Acknowledgments This work was supported by the National Hi-Tech Research and Development (863) Program of China (Grant No. 2007AA09Z215).

References

1. Korde UA (1998) Active heave compensation on drill-ships in irregular waves. *Ocean Eng* 25:541–561
2. Johansen TA, Fossen TI, Sagatun SI et al (2003) Wave synchronizing crane control during water entry in offshore moonpool operations-experimental results. *IEEE J Ocean Eng* 28:720–728
3. Skaare B, Egeland O (2006) Parallel force/position crane control in marine operations. *IEEE J Ocean Eng* 31:599–613
4. Do KD, Pan J (2008) Nonlinear control of an active heave compensation system. *Ocean Eng* 35:558–571
5. Marconi L, Isidori A, Serrani A (2002) Autonomous vertical landing on an oscillating platform: an internal-model based approach. *Automatica* 38:21–32
6. Serrani A, Isidorim A, Marconi L (2001) Semiglobal nonlinear output regulation with adaptive internal model. *IEEE Trans Autom Control* 46:1178–1194
7. Messineo S, Serrani A (2009) Offshore crane control based on adaptive external models. *Automatica* 45:2546–2556
8. Serrani A (2006) Rejection of harmonic disturbances at the controller input via hybrid adaptive external models. *Automatica* 42:1977–1985
9. Kuchler S, Mahl T, Neupert J et al (2011) Active control for an offshore crane using prediction of the vessel's motion. *IEEE/ASME Trans Mechatronics* 16:297–309
10. Messineo S, Celani F, Egeland O (2008) Crane feedback control in offshore moonpool operations. *Control Eng Pract* 16:356–364

Chapter 88

Models for Multiple Attribute Decision Making with Intuitionistic Trapezoidal Information

Liang Xiao-yi

Abstract The aim of this chapter is to investigate the multiple attribute decision making problems to deal with supplier selection with intuitionistic trapezoidal fuzzy information, in which the information about attribute weights is completely known, and the attribute values take the form of intuitionistic trapezoidal fuzzy number. We developed a multiple attribute decision making method by closeness to ideal alternative in intuitionistic trapezoidal fuzzy setting. Then, we calculate the distances between the ideal alternative and all alternatives to determine the ranking of all alternatives. Finally, an example about supplier selection is shown to highlight the procedure of the proposed algorithm.

Keywords Multiple attribute decision making • Supplier selection • Intuitionistic trapezoidal fuzzy number • Weight information • Distances measure

88.1 Introduction

Atanassov [1, 2] introduced the concept of intuitionistic fuzzy set (IFS) characterized by a membership function and a non-membership function, which is a generalization of the concept of fuzzy set [3] whose basic component is only a membership function. The intuitionistic fuzzy set has received more and more attention since its appearance [4]. Later, Atanassov and Gargov [5] further introduced the interval-valued intuitionistic fuzzy set (IVIFS), which is a

L. Xiao-yi (✉)

Shaanxi University of Science and Technology, Northern Suburb of University Park
Weiyang, Xi'an 710021, Shaanxi, China
e-mail: liangxiaoyi99@126.com

generalization of the IFS. The fundamental characteristic of the IVIFS is that the values of its membership function and non-membership function are intervals rather than exact numbers. Xu [6] developed some aggregation operators with interval-valued intuitionistic fuzzy information. Xu [7] investigated the interval-valued intuitionistic fuzzy MADM with the information about attribute weights is incompletely known or completely unknown, a method based on the ideal solution was proposed. Wang [8] investigated the interval-valued intuitionistic fuzzy MADM with incompletely known weight information. A nonlinear programming model is developed. Then using particle swarm optimization algorithms to solve the nonlinear programming models, the optimal weights are gained. And ranking is performed through the comparison of the distances between the alternatives and idea/anti-idea alternative. Shu et al. [9–12] gave the definition and operational laws of intuitionistic triangular fuzzy number and proposed an algorithm of the intuitionistic fuzzy fault-tree analysis.

The aim of this chapter is to investigate the multiple attribute decision making problems to deal with supplier selection with intuitionistic trapezoidal fuzzy information, in which the information about attribute weights is completely known, and the attribute values take the form of intuitionistic trapezoidal fuzzy number. We developed a multiple attribute decision making method by closeness to ideal alternative in intuitionistic trapezoidal fuzzy setting. Then, we calculate the distances between the ideal alternative and all alternatives to determine the ranking of all alternatives. The remainder of this chapter is set out as follows. In the next section, we introduce some basic concepts related to intuitionistic trapezoidal fuzzy numbers. In Sect. 88.3 we introduce intuitionistic trapezoidal fuzzy multiple attribute decision making problems with completely known weight information. Then, we calculate the distances between the ideal alternative and all alternatives to determine the ranking of all alternatives. In Sect. 88.4 we conclude the chapter and give some remarks.

88.2 Preliminaries

In the following, we shall introduce some basic concepts related to intuitionistic trapezoidal fuzzy numbers.

Definition 1 Let \tilde{a} is an intuitionistic trapezoidal fuzzy number, its membership function is [10–12]:

$$\mu_{\tilde{a}}(x) = \begin{cases} \frac{x-a}{b-a} \mu_{\tilde{a}}, & a \leq x < b; \\ \mu_{\tilde{a}}, & b \leq x \leq c; \\ \frac{d-x}{d-c} \mu_{\tilde{a}}, & c < x \leq d; \\ 0, & \text{others.} \end{cases} \quad (88.1)$$

Its non-membership function is:

$$v_{\tilde{a}}(x) = \begin{cases} \frac{b-x+v_{\tilde{a}}(x-a)}{b-a}, & a_1 \leq x < b; \\ v_{\tilde{a}}, & b \leq x \leq c; \\ \frac{x-c+v_{\tilde{a}}(d_1-x)}{d_1-c}, & c < x \leq d_1; \\ 0, & \text{others.} \end{cases} \tag{88.2}$$

where $0 \leq \mu_{\tilde{a}} \leq 1; 0 \leq v_{\tilde{a}} \leq 1$ and $\mu_{\tilde{a}} + v_{\tilde{a}} \leq 1; a, b, c, d \in R$. Then $\tilde{a} = \langle ([a, b, c, d]; \mu_{\tilde{a}}), ([a_1, b, c, d_1]; v_{\tilde{a}}) \rangle$ is called an intuitionistic trapezoidal fuzzy number.

For convenience, let $\tilde{a} = ([a, b, c, d]; \mu_{\tilde{a}}, v_{\tilde{a}})$.

Definition 2 Let $\tilde{a}_1 = ([a_1, b_1, c_1, d_1]; \mu_{\tilde{a}_1}, v_{\tilde{a}_1})$ and $\tilde{a}_2 = ([a_2, b_2, c_2, d_2]; \mu_{\tilde{a}_2}, v_{\tilde{a}_2})$ be two intuitionistic trapezoidal fuzzy number, and $\lambda \geq 0$, then [24–26]

- (1) $\tilde{a}_1 + \tilde{a}_2 = ([a_1 + a_2, b_1 + b_2, c_1 + c_2, d_1 + d_2]; \mu_{\tilde{a}_1} + \mu_{\tilde{a}_2} - \mu_{\tilde{a}_1} \cdot \mu_{\tilde{a}_2}, v_{\tilde{a}_1} \cdot v_{\tilde{a}_2})$;
- (2) $\tilde{a}_1 \cdot \tilde{a}_2 = ([a_1 \cdot a_2, b_1 \cdot b_2, c_1 \cdot c_2, d_1 \cdot d_2]; \mu_{\tilde{a}_1} \cdot \mu_{\tilde{a}_2}, v_{\tilde{a}_1} + v_{\tilde{a}_2} - v_{\tilde{a}_1} \cdot v_{\tilde{a}_2})$;
- (3) $\lambda \tilde{a}_1 = ([\lambda a_1, \lambda b_1, \lambda c_1, \lambda d_1]; 1 - (1 - \mu_{\tilde{a}_1})^\lambda, v_{\tilde{a}_1}^\lambda)$;
- (4) $\tilde{a}_1^\lambda = ([a_1^\lambda, b_1^\lambda, c_1^\lambda, d_1^\lambda]; \mu_{\tilde{a}_1}^\lambda, 1 - (1 - v_{\tilde{a}_1})^\lambda)$

Definition 3 Intuitionistic trapezoidal fuzzy positive ideal solution and intuitionistic trapezoidal fuzzy negative ideal solution are defined as follows:

$$\tilde{r}^+ = ([a^+, b^+, c^+, d^+]; \mu^+, v^+) = ([1, 1, 1, 1]; 1, 0) \tag{88.3}$$

Definition 4 Let $\tilde{a}_1 = ([a_1, b_1, c_1, d_1]; \mu_{\tilde{a}_1}, v_{\tilde{a}_1})$ and $\tilde{a}_2 = ([a_2, b_2, c_2, d_2]; \mu_{\tilde{a}_2}, v_{\tilde{a}_2})$ be two intuitionistic trapezoidal fuzzy number, then the normalized Hamming distance between \tilde{a}_1 and \tilde{a}_2 is defined as follows [26]:

$$d(\tilde{a}_1, \tilde{a}_2) = \frac{1}{8} (|(1 + \mu_{\tilde{a}_1} - v_{\tilde{a}_1})a_1 - (1 + \mu_{\tilde{a}_2} - v_{\tilde{a}_2})a_2| + |(1 + \mu_{\tilde{a}_1} - v_{\tilde{a}_1})b_1 - (1 + \mu_{\tilde{a}_2} - v_{\tilde{a}_2})b_2| + |(1 + \mu_{\tilde{a}_1} - v_{\tilde{a}_1})c_1 - (1 + \mu_{\tilde{a}_2} - v_{\tilde{a}_2})c_2| + |(1 + \mu_{\tilde{a}_1} - v_{\tilde{a}_1})d_1 - (1 + \mu_{\tilde{a}_2} - v_{\tilde{a}_2})d_2|). \tag{88.4}$$

88.3 Models for Multiple Attribute Decision Making with Intuitionistic Trapezoidal Fuzzy Information

The following assumptions or notations are used to represent the MADM problems with completely known weight information in intuitionistic trapezoidal fuzzy setting. Let $A = \{A_1, A_2, \dots, A_m\}$ be a discrete set of alternatives, and $G = \{G_1, G_2, \dots, G_n\}$ be the set of attributes. Suppose that $\tilde{R} = (\tilde{r}_{ij})_{m \times n} =$

$\left([a_{ij}, b_{ij}, c_{ij}, d_{ij}]; \mu_{ij}, \nu_{ij} \right)_{m \times n}$ is the intuitionistic trapezoidal fuzzy decision matrix, $\mu_{ij}^{(k)} \in [0, 1], \nu_{ij}^{(k)} \in [0, 1], \mu_{ij}^{(k)} + \nu_{ij}^{(k)} \leq 1, i = 1, 2, \dots, m, j = 1, 2, \dots, n, k = 1, 2, \dots, t$. The information about attribute weights is incompletely known. Let $w = (w_1, w_2, \dots, w_n)$ be the weight vector of attributes, where $w_j \geq 0, j = 1, 2, \dots, n, \sum_{j=1}^n w_j = 1$.

In the following, we develop a practical method for solving the MADM problems, in which the information about attribute weights is completely known, and the attribute values take the form of intuitionistic trapezoidal fuzzy variables. The method involves the following steps:

Step 1. Suppose that $\tilde{R} = \left(\tilde{r}_{ij} \right)_{m \times n} = \left([a_{ij}, b_{ij}, c_{ij}, d_{ij}]; \mu_{ij}, \nu_{ij} \right)_{m \times n}$ is the intuitionistic trapezoidal fuzzy decision matrix, $\mu_{ij}^{(k)} \in [0, 1], \nu_{ij}^{(k)} \in [0, 1], \mu_{ij}^{(k)} + \nu_{ij}^{(k)} \leq 1, i = 1, 2, \dots, m, j = 1, 2, \dots, n, k = 1, 2, \dots, t$. The information about attribute weights is incompletely known. Let $w = (w_1, w_2, \dots, w_n)$ be the weight vector of attributes, where $w_j \geq 0, j = 1, 2, \dots, n, \sum_{j=1}^n w_j = 1$.

Step 2. Let $\tilde{R} = \left(\tilde{r}_{ij} \right)_{m \times n} = \left([a_{ij}, b_{ij}, c_{ij}, d_{ij}]; \mu_{ij}, \nu_{ij} \right)_{m \times n}$ be the intuitionistic trapezoidal fuzzy decision matrix, the ideal alternative can be defined as follows:

$$A^+ = (\tilde{r}_1^+, \tilde{r}_2^+, \dots, \tilde{r}_n^+),$$

where $\tilde{r}_j^+ = ([a^+, b^+, c^+, d^+]; \mu^+, \nu^+) = ([1, 1, 1, 1]; 1, 0)$.

Step 3. Utilize the weight vector $w = (w_1, w_2, \dots, w_n)$ and utilize the (88.4) to derive the distances $d(A_i, A^+)(i = 1, 2, \dots, m)$, by which we can get the ranking of all alternatives $A_i(i = 1, 2, \dots, m)$.

$$d(A_i, A^+) = \sum_{j=1}^n w_j d(\tilde{r}_j^+, \tilde{r}_{ij}) \tag{88.5}$$

Step 4. Rank all the alternatives $A_i(i = 1, 2, \dots, m)$ and select the best one(s) in accordance with $d(A_i, A^+)(i = 1, 2, \dots, m)$. The smaller $d(A_i, A^+)$, the better the alternatives A_i .

Step 5. End.

88.4 Conclusion

In this chapter, we have investigated the multiple attribute decision making problems to deal with supplier selection with intuitionistic trapezoidal fuzzy information, in which the information about attribute weights is completely known, and the attribute values take the form of intuitionistic trapezoidal fuzzy number. We developed a multiple attribute decision making method by closeness to ideal alternative in intuitionistic trapezoidal fuzzy setting. Then, we calculate

the distances between the ideal alternative and all alternatives to determine the ranking of all alternatives. At last, a practical example about supplier selection is provided to illustrate the proposed method.

References

1. Atanassov K (1986) Intuitionistic fuzzy sets. *Fuzzy Sets Syst* 20:87–96
2. Atanassov K (1989) More on intuitionistic fuzzy sets. *Fuzzy Sets Syst* 33:37–46
3. Zadeh LA (1965) Fuzzy sets. *Inf Control* 8:338–356
4. Gau WL, Buehrer DJ (1993) Vague sets. *IEEE Trans Syst Man Cybern* 23(2):610–614
5. Bustine H, Burillo P (1996) Vague sets are intuitionistic fuzzy sets. *Fuzzy Sets Syst* 79:403–405
6. Xu ZS, Yager RR (2006) Some geometric aggregation operators based on intuitionistic fuzzy sets. *Int J Gen Syst* 35:417–433
7. Xu ZS (2007) Intuitionistic preference relations and their application in group decision making. *Inf Sci* 177(11):2363–2379
8. Xu ZS (2007) Intuitionistic fuzzy aggregation operators. *IEEE Trans Fuzzy Syst* 15(6):1179–1187
9. Wei GW (2008) Maximizing deviation method for multiple attribute decision making in intuitionistic fuzzy setting. *Knowl Based Syst* 21(8):833–836
10. Wei GW (2009) Some geometric aggregation functions and their application to dynamic multiple attribute decision making in intuitionistic fuzzy setting. *Int J Uncertainty Fuzz Knowl Based Syst* 17(2):179–196
11. Wei GW (2010) GRA method for multiple attribute decision making with incomplete weight information in intuitionistic fuzzy setting. *Knowl Based Syst* 23(3):243–247
12. Wei GW (2010) Some induced geometric aggregation operators with intuitionistic fuzzy information and their application to group decision making. *Appl Soft Comput* 10(2):423–431

Chapter 89

Modeling the Risk Factors in Ergonomic Processes Using Fuzzy Logic

Prafulla Kumar Manoharan, Sanjay Jha and Bijay Kumar Singh

Abstract This paper presents an approach to minimize the risk for transforming measured body data between various postures. In this research the measured human body is substituted by a proper set of critical points using fuzzy logic. They are used as a basis of transforming the data, and they are required to describe specific body postures. Artificial neural networks have been applied to the actual conversion of data. The input is a set of demographic data and the coordinates of the critical points characterizing a given posture.

Keywords Ergonomics · Fuzzy logics (FL) · Body postures

89.1 Introduction

Engineers and ergonomists are keen to exploit the potential of the three dimensional (3D) anthropometric technologies. The 3D measurements are more deployable and provide a more elaborated micro level description of the human

P. K. Manoharan (✉)

Department of Production Engineering, Birla Institute of Technology, Mesra,
Deoghar Campus, Jharkhand, India
e-mail: prafkumar@hotmail.com

S. Jha · B. K. Singh

Department of Production Engineering, Birla Institute of Technology, Mesra,
Ranchi, Jharkhand, India
e-mail: sanjujha@hotmail.com

B. K. Singh

e-mail: Bksingh@bitmesra.ac.in

bodies in comparison with the traditional manual 1 or 2D data processing. In many industrial design cases, there is a need to take into consideration various postures of the human body when the product is designed. This paper presents an approach to minimize the risk for transforming measured body data between various postures. In this research the measured human body is substituted by a proper set of critical points using fuzzy logic. They are used as a basis of transforming the data, and they are required to describe specific body postures. Artificial neural networks have been applied to the actual conversion of data. The input is a set of demographic data and the coordinates of the critical points characterizing a given posture.

The rest of the paper has been organized as follows:

Section 89.2 presents overview of the similar work.

Section 89.3 describes mathematical formation of ergonomics process, followed by sect. 89.4 which processes an algorithm to evaluate the demographic data feed into the model.

Section 89.5 finally presents scope of work and conclusion.

89.2 Literature Survey

Recently there are substantial numbers of research breakthroughs in ergonomic modeling. The problem of sharing and reusing information in the lifecycle (LC): from design, manufacturing to operation, maintenance and recycling have been directed through perfect ergonomics modelling [3]. The results reported in those publications have been obtained in the European research project ManuVAR (211548)—Manual Work Support throughout System Lifecycle by Exploiting Virtual and augmented Reality (VR/AR) [1].

In many fields of science, including biology, psychology, and so on, human observers have provided linguistic descriptions and explanations of various systems. However, to study these phenomena in a systematic manner, there is a need to construct a suitable mathematical model, a process that usually requires subtle mathematical understanding. Fuzzy modeling is a simple, direct, and natural approach for transforming the linguistic description to a mathematical model [4]. Considering this line of research, ergonomics interventions often focus on reducing exposure in those parts of the job having the highest exposure levels, while leaving other parts unattended. A successful Intervention will thus change the form of the job exposure distribution. This disqualifies standard methods for assessing the ability of various exposure measurement strategies to correctly detect an intervention's effect on the overall job exposure of an individual worker, in particular for the safety or ergonomics practitioner who with limited resources can only collect a few measurements [5].

89.3 Mathematical Proposition

Fuzzy Logics (FL) provides an appropriate logical mathematical framework to handle problems with such characteristics, since [6]:

deals with uncertainty and imprecision of reasoning processes;
 allows the modelling of the heuristic knowledge (that cannot be described by traditional mathematical equations); and

Allows the computation of linguistic information.

The basic fuzzy linguistic to formulate fuzzy model can be shown in following scheme: Fig. 89.1

Hence, the particular ergonomic modelling for assessing the risk and uncertainty has been accomplished through fuzzy logic [7]. In this chapter, we consider

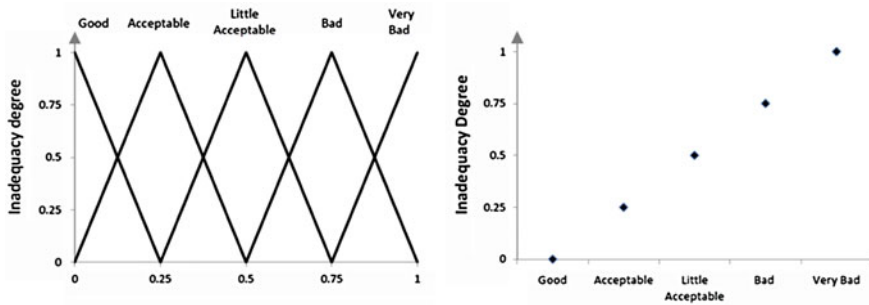
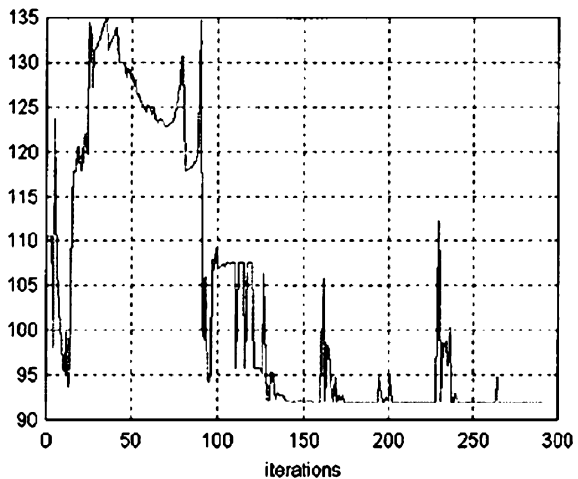


Fig. 89.1 Fuzzy linguistic components

Fig. 89.2 Different Iterations during working movements

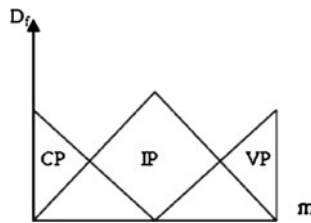


the basic formulation of fuzzy to keep track the different risk evaluation ergonomic movement.

$$G = \frac{\text{the sum of } (\mu_i * \beta_i)}{\text{the sum of } (\mu_i)}$$

where : $1 \leq i \leq m$, m : number of rule, B : centroid of the backend membership function correspond for each rule. μ : factor of membership correspond for each rule. This intelligent task uses the fuzzy linguistic terms and calculates for each degree of membership functions under expertise of an expert system (ES). An ES is a computer program that functions, is in a narrow domain, dealing with specialized knowledge, generally possessed by human experts. Fig. 89.2

89.4 Proposed Algorithm

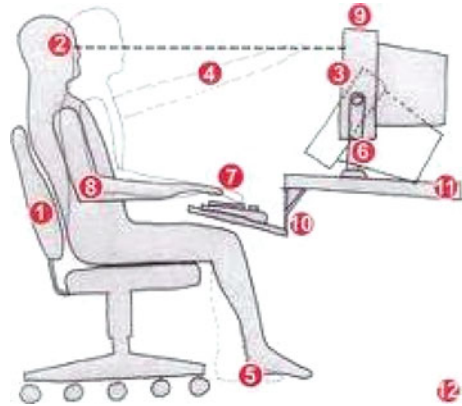


```

Begin
  Start set up environment for postures/* Point 1
  Fig. 89.3*/
  Initialization
  Move
    IF {the movement is Recorded} DO End task
      ELSE
        Begin
          L1:If {the 1st Posture is detected?} DO
            Begin
              Change the direction.
              Move
                Free ()
                is attended.
            End task.
          Else Goto L1.
        End.
      End
    End
  End

```


Fig. 89.3 Courtesy: <http://www.armannd.com/>



89.4.1 Data Set and Analysis

The developed model focused on in simple office environment, and identified a total of certain primary risk associated with task. Using the proposed algorithm, the different critical ergonomic points in a posture has been evaluated (shown in red marks)

Risk factor category	Relative weight	Work iterations	Instances o fuzzy	Membership value
Task	$w1 = 0.2657$	x1:	Total typing hours (per day)	(0.231,0)
		x2:	Typing speed	(0.144,0)
		x3:	Continuous typing time	(-0.078,0)
		x4:	Hand temperature	(0.039,0.298)
		x5:	Comfortability of workstation	(-0.133,0.044)
		x6:	Work surface hardness	(0.059,0)
		x7:	Hand rest time	(-0.007,0)
		x8:	Rating of perceived exertion	(-0.001,0)
		x9:	Working under pressure	(0.051,0)

89.5 Conclusion

The paper elaborates the risk in ergonomics movement during office environment using fuzzy logic as a primary tool. The data set adopts different linguistic attribute to evaluate the de-fuzzified results.

References

1. Krassi B, D'Cruz M, Vink P (2010a) ManuVAR: a framework for improving manual work through virtual and augmented reality. Conference on applied human factors and ergonomics—AHFE 2010, USA
2. FAST ERGO_X—A tool for ergonomic auditing and work-related musculoskeletal disorders prevention work: a journal of prevention, assessment and rehabilitation, IOS Press 1051–9815 (Print) 1875–9270 (Online) 34(2/2009):133–148
3. Nunes IL (2009) National patent no 103446: Método de Análise Ergonómica de Postos de Trabalho. [Workstation Ergonomic Analysis Method]
4. Tron E, Margalot M (2003) Modeling observed natural behavior using fuzzy logic. Proc IEEE Int Conf Fuzzy Syst 1:74–78
5. Mathiassen SE, Paquet V. The ability of limited exposure sampling to detect effects of interventions that reduce the occurrence of pronounced trunk inclination. Applied Ergonomics—rok 2010, ročník 41 Číslo 2 (March 2010) SPECIAL pp. 295–304
6. Zadeh LA (1996) Fuzzy logic = computing with words. IEEE Trans Fuzzy Syst 4(2):103–111
7. Nunes IL (2006) ERGO_X—The model of a fuzzy expert system for workstation ergonomic analysis. In: Karwowski W (ed) International encyclopaedia of ergonomics and human factors. CRC Press, New York, pp 3114–3121

Chapter 90

Modified Projective Synchronization of Uncertain Chaotic Systems

R. Z. Luo, Y. L. Wang and S. C. Deng

Abstract In this chapter, we investigate the issue on the modified projective synchronization (MPS) of chaotic systems in the presence of uncertainty by combining using the impulsive control scheme and nonlinear controller. We construct a uncertain response chaotic system to implement MPS with the uncertain driven chaotic system and propose some sufficient conditions to make them achieve MPS.

Keywords Modified projective synchronization · Impulsive control · Chaotic system

90.1 Introduction

The first idea of synchronizing two identical chaotic systems with different initial conditions was introduced by Pecora and Carroll in 1990 [1], since then, a wide variety of different techniques and methods have been proposed to achieve chaos synchronization, which include adaptive control, sliding mode control, fuzzy control, optimal control, digital redesign control, backstepping control, impulsive

R. Z. Luo (✉) · Y. L. Wang · S. C. Deng
Department of Mathematics, Nanchang University,
Nanchang, 330031 People's Republic of China
e-mail: luo_rz@163.com

Y. L. Wang
e-mail: 1040006247@qq.com

S. C. Deng
e-mail: deng548629@163.com

control [2], and so on. Recently, impulsive control has been widely used to stabilize and synchronize chaotic systems [2–4]. Its necessity and importance lie in that impulsive control may give a more efficient method to deal with systems that cannot endure continuous disturbance. Additionally, impulsive method can also greatly reduce the control cost. In the past 20 years, various types of synchronization have been proposed and investigated, such as complete synchronization (CS) [1], phase synchronization [5], generalized projective synchronization [6], modified projective synchronization (MPS) [7].

However, most of the existing synchronization methods by using impulsive control are mainly concerned the CS [2–4]. On the other side, the uncertainty of chaotic system is unavoidable. For example, it is impossible to implement two identical systems and/or maintain them to be identical all the time during their operation. As far as the authors know, very few chapters have considered the MPS of uncertain chaotic system via impulsive control.

Motivated by the aforementioned comments, the main aim of this chapter is to study the MPS of uncertain autonomous chaotic system by combining using the impulsive control scheme and nonlinear controller. For a uncertain chaotic driven system we construct a response system and propose some new sufficient conditions for achieving MPS between the driven system and response system.

90.2 Synchronization Criteria

Consider the uncertain chaotic system in the form of

$$\dot{x} = M(x)x + \Delta M_1(x, t)x, \quad (90.1)$$

where $x \in R^n$ is the state vector, $M(x), \Delta M_1(x, t) \in R^{n \times n}$ are functional matrix, and $\Delta M_1(x, t)x$ is the uncertainty of system (90.1).

Suppose system (90.1) is the driven system, in order to synchronize system (90.1) we construct the response system:

$$\dot{y} = \alpha^{-1}M(x)(\alpha y) + \alpha^{-1}\Delta M_2(x, y, t)(\alpha y) + \alpha^{-1}u(t), \quad (90.2)$$

where y is the state vector of the response system, $\alpha = \text{diag}(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$, α_i are the scaling factors, $i = 1, 2, 3, 4$, and α^{-1} is the inverse matrix of matrix α . $\Delta M_2(x, y, t)y$ is the uncertainty of system (90.2), where $\Delta M_2(x, y, t) \in R^{n \times n}$ is a functional matrix. $u(t)$ is a control signal used to regulate the states of response system towards those of the driven system.

In order to derive our main results, we make the following assumption.

Assumption. Suppose the uncertainties are bounded, i.e. there exists a positive number ρ , such that $\|\Delta M_1(x, t)\| \leq \rho$, $\|\Delta M_2(x, y, t)\| \leq \rho$.

Note that at the discrete time instants, t_i , $i = 1, 2, \dots$, the state variable of system (90.1) is transmitted to the response system (90.2) and leads to a “jump” of

the state of the response system. In this case, the model of response system can be rewritten as

$$\begin{cases} \dot{y} = \alpha^{-1}M(x)(\alpha y) + \alpha^{-1}\Delta M_2(x, y, t)(\alpha y) + \alpha^{-1}u(t), & t \neq t_i \\ \Delta y(t_i) = y(t_i^+) - y(t_i^-) = -B_i(x(t_i^-) - \alpha y(t_i^-)) = -B_i(x(t_i) - \alpha y(t_i)), & t = t_i, i = 1, 2, \dots, \\ y(t_0^+) = y_0, \end{cases} \tag{90.3}$$

where $B_i \in R^{n \times n}$ is impulse strength. $t_i, i = 1, 2, \dots$ are constants satisfying:

$$t_0 < t_1 < t_2 < \dots < t_i < t_{i+1} < \dots, \lim_{i \rightarrow \infty} t_i \rightarrow \infty.$$

Let $e = x - \alpha y$ and subtract (90.3) from (90.1) we have $\dot{e} = (M(x) + \Delta M_1(x, t))e + \Delta M(x, y, t) - u(t)$, where $\Delta M(x, y, t) = \Delta M_1(x, t)(\alpha y) - \Delta M_2(x, y, t)(\alpha y)$. Thus, the following error dynamics system is obtained:

$$\begin{cases} \dot{e} = (M(x) + \Delta M_1(x, t))e + \Delta M(x, y, t) - u(t), & t \neq t_i \\ \Delta e(t_i) = e(t_i^+) - e(t_i^-) = \alpha B_i e(t_i^-) = \alpha B_i e(t_i), & t = t_i, i = 1, 2, \dots, \\ e(t_0^+) = e_0. \end{cases} \tag{90.4}$$

Suppose

$$\Delta M_1(x, t) = \begin{pmatrix} m_{11}(x, t) & m_{12}(x, t) & \dots & m_{1n}(x, t) \\ m_{21}(x, t) & m_{22}(x, t) & \dots & m_{2n}(x, t) \\ \dots & \dots & \dots & \dots \\ m_{n1}(x, t) & m_{n2}(x, t) & \dots & m_{nn}(x, t) \end{pmatrix},$$

$$\Delta M_2(x, y, t) = \begin{pmatrix} \tilde{m}_{11}(x, y, t) & \tilde{m}_{12}(x, y, t) & \dots & \tilde{m}_{1n}(x, y, t) \\ \tilde{m}_{21}(x, y, t) & \tilde{m}_{22}(x, y, t) & \dots & \tilde{m}_{2n}(x, y, t) \\ \dots & \dots & \dots & \dots \\ \tilde{m}_{n1}(x, y, t) & \tilde{m}_{n2}(x, y, t) & \dots & \tilde{m}_{nn}(x, y, t) \end{pmatrix},$$

then we have

$$\begin{aligned} \Delta M(x, y, t) &= \Delta M_1(x, t)(\alpha y) - \Delta M_2(x, y, t)(\alpha y) \\ &= \begin{pmatrix} (m_{11} - \tilde{m}_{11})(\alpha_1 y_1) \\ (m_{21} - \tilde{m}_{21})(\alpha_1 y_1) \\ \dots \\ (m_{n1} - \tilde{m}_{n1})(\alpha_1 y_1) \end{pmatrix} + \dots + \begin{pmatrix} (m_{1n} - \tilde{m}_{1n})(\alpha_n y_n) \\ (m_{2n} - \tilde{m}_{2n})(\alpha_n y_n) \\ \dots \\ (m_{nn} - \tilde{m}_{nn})(\alpha_n y_n) \end{pmatrix}. \end{aligned} \tag{90.5}$$

The goal of this chapter is to find some conditions on the controller $u(t)$, the control gains B_i , and the impulse distances $\tau_{i+1} = t_{i+1} - t_i < \infty (i = 1, 2, \dots)$ such that the response system (90.3) is modified projective synchronized with system (90.1), i.e.

$$\|x(t) - \alpha y(t)\| \rightarrow 0 \text{ for } t \rightarrow \infty.$$

For convenience, define the following notations:

$$\begin{aligned} \lambda_M &= \lambda_{\max}((M(x) + \Delta M_1(x, t))^T + (M(x) + \Delta M_1(x, t))), \lambda_i \\ &= \lambda_{\max}(I + \alpha B_i)^T(I + \alpha B_i), \end{aligned}$$

Where $\lambda_{\max}(M)$ is the maximal eigenvalue of matrix M . Let

$$\begin{aligned} u(t) &= \begin{pmatrix} |m_{11}(x, t) - \tilde{m}_{11}(x, y, t)|\alpha_1 y_1 |sgn(e_1)| \\ |m_{21}(x, t) - \tilde{m}_{21}(x, y, t)|\alpha_1 y_1 |sgn(e_2)| \\ \dots \\ |m_{n1}(x, t) - \tilde{m}_{n1}(x, y, t)|\alpha_1 y_1 |sgn(e_n)| \end{pmatrix} + \dots \\ &+ \begin{pmatrix} |m_{1n}(x, t) - \tilde{m}_{1n}(x, y, t)|\alpha_n y_n |sgn(e_1)| \\ |m_{2n}(x, t) - \tilde{m}_{2n}(x, y, t)|\alpha_n y_n |sgn(e_2)| \\ \dots \\ |m_{nn}(x, t) - \tilde{m}_{nn}(x, y, t)|\alpha_n y_n |sgn(e_n)| \end{pmatrix} \end{aligned} \tag{90.6}$$

then we obtain the following results.

Theorem 1 If $u(t)$ is designed based on (90.6) and (90.1) If $\lambda_M < 0$ and there exist constants $\beta_i(0 \leq \beta = \max\{\beta_1, \beta_2, \dots, \beta_i, \dots\}, < -\lambda_M)$ such that

$$\ln \lambda_i - \beta_i(t_i - t_{i-1}) \leq 0, \quad i = 1, 2, \dots, \tag{90.7}$$

then the origin of system (90.4) is globally exponentially stable, which implies that system (90.1) and (90.3) can achieve modified projective synchronization.

(2) If $\lambda_M \geq 0$ and there exist constants $\beta_i(\beta_i > 1)$ such that

$$\ln(\beta_i \lambda_i) + \lambda_M(t_{i+1} - t_i) \leq 0, \quad i = 1, 2, \dots, \tag{90.8}$$

then, when $\lim_{i \rightarrow \infty} \beta_1 \beta_2 \dots \beta_i = k(0 < k < +\infty)$, the origin of system (90.4) is stable; when $\lim_{i \rightarrow \infty} \beta_1 \beta_2 \dots \beta_i = +\infty$, the origin of system (90.4) is globally exponentially stable, which implies that system (90.1) and (90.3) can achieve modified projective synchronization.

Proof Construct a Lyapunov function in the form of $V(t) = e^T e$. When $t \in (t_{i-1}, t_i]$ the total derivative of $V(t)$ with respect to (90.4) is

$$\begin{aligned} \dot{V}(t) &= ((M(x) + \Delta M_1(x, t))e + \Delta M(x, y, t) - u(t))^T e + e^T ((M(x) + \Delta M_1(x, t))e + \Delta M(x, y, t) - u(t)) \\ &= e^T ((M(x) + \Delta M_1(x, t))^T + (M(x) + \Delta M_1(x, t)))e + 2e^T (\Delta M(x, y, t) - u(t)) \lambda_M e^T e = \lambda_M V(t), \end{aligned} \tag{90.9}$$

which implies that

$$V(t) \leq V(t_{i-1}^+) \exp(\lambda_M(t - t_{i-1})), \quad t \in (t_{i-1}, t_i], \quad i = 1, 2, \dots, \tag{90.10}$$

On the other hand, it follows from system (90.4) that

$$V(t_i^+) = ((I + \alpha B_i)e(t_i))^T (I + \alpha B_i)e(t_i) \leq \lambda_i e^T(t_i)e(t_i) = \lambda_i V(t_i). \tag{90.11}$$

From (90.10) and (90.11), we know for $t \in (t_i, t_{i+1}]$ we have

$$V(t) \leq V(t_0^+) \lambda_1 \lambda_2 \cdots \lambda_i \exp(\lambda_M(t - t_0)), i = 1, 2, \dots, \tag{90.12}$$

(i) When $\lambda_M < 0$, it follows from (90.7) and (90.14) that for $t \in (t_i, t_{i+1}]$,

$$V(t) \leq$$

Thus, the origin of system (90.4) is globally exponentially stable, which implies that system (90.1) and (90.3) can achieve modified projective synchronization.

(ii) When $\lambda_M \geq 0$, it follows from (90.8) and (90.12) that for, $t \in (t_i, t_{i+1}]$

$$V(t) \leq$$

Thus, if $\lim_{i \rightarrow \infty} \beta_1 \beta_2 \cdots \beta_i = k (0 < k < +\infty)$, then the origin of system (90.4) is stable, and if $\lim_{i \rightarrow \infty} \beta_1 \beta_2 \cdots \beta_i = +\infty$, then the origin of system (90.4) is globally exponentially stable, which implies that system (90.1) and (90.3) can achieve modified projective synchronization. This completes the proof.

In the following, some remarks are listed in order:

Remark 1 In Theorem 1 we don't need β_i identically equal to a constant and β_i can be less than 1. Thus our results may be more useful from a practical point of view.

Remark 2 Since $\|\Delta M_1(x, t)\| \leq \rho$, $\|\Delta M_1(x, t)\| \leq \rho$, then there must exist some constants a_{ij} , $i, j = 1, 2, \dots, n$, such that $|m_{ij}(x, t) - \tilde{m}_{ij}(x, y, t)| \leq a_{ij}$, where $a_{ij} \geq 0$. Thus from the proof of Theorem 1 it is easy to know that the controller $u(t)$ in Theorem 1 can be chosen as:

$$u(t) = \begin{pmatrix} a_{11}|\alpha_1 y_1|sgn(e_1) \\ a_{21}|\alpha_1 y_1|sgn(e_2) \\ \dots \\ a_{n1}|\alpha_1 y_1|sgn(e_n) \end{pmatrix} + \dots + \begin{pmatrix} a_{1n}|\alpha_n y_n|sgn(e_1) \\ a_{2n}|\alpha_n y_n|sgn(e_2) \\ \dots \\ a_{nn}|\alpha_n y_n|sgn(e_n) \end{pmatrix}. \tag{90.13}$$

Furthermore, by noting that the state variables of chaotic system are bounded, the controller $u(t)$ in Theorem 1 can be simply chosen as:

$$u(t) = \begin{pmatrix} \gamma sgn(e_1) \\ \gamma sgn(e_2) \\ \dots \\ \gamma sgn(e_n) \end{pmatrix}, \tag{90.14}$$

where γ is a constant. We can evaluate γ by simulation, or choose γ large enough. The controller (90.14) is more useful from the point of practice.

90.3 Conclusions

In this chapter, we investigate the MPS of autonomous chaotic system in the presence of uncertainty. We construct a uncertain response chaotic system to implement MPS with the uncertain driven chaotic system and propose some sufficient conditions to make them achieve modified projective synchronization.

Acknowledgment The authors acknowledge the support provided by the Science and Technology Foundation of the Education Department of Jiangxi Province (GJJ11296).

References

1. Pecora LM, Carroll TL (1990) Synchronization in chaotic systems. *Phys Rev Lett* 64:821–824
2. Luo RZ (2008) Coordination of supply chains by option contracts: A cooperative game theory approach. *Dyn Continuous Discret Impulsive Sys Ser B Appl Algorithms* 15:831–841
3. Liu XZ (2009) Existence and efficiency of oligopoly equilibrium under toll and capacity competition. *Nonlinear anal* 71:e1320–e1327
4. Jiang HB, Bi QS (2010) First-principles study of the formation mechanisms of nitrogen molecule in annealed ZnO. *Phys Lett A* 374:2723–2729
5. Rosenblum MG, Pikovsky AS, Kurth J (1996) Phase Synchronization of Chaotic Oscillators. *Phys Rev Lett* 76:1804–1807
6. Kocarev L, Parlitz U (1996) Synchronization of uncertain chaotic systems using active sliding mode control. *Phys Rev Lett* 76:1816–1819
7. Li GH (2007) Synchronization of uncertain chaotic systems using active sliding mode control [J]. *Chaos Solitons Fractals* 32:1786–1790

Chapter 91

Multi-Agent Based Architecture Supporting Collaborative Product Lightweight Design

Zheng Liu, Fuyuan Xu, Kai Pan, Xinjian Gu and Yongwei Zhang

Abstract Increasingly notable environmental problems facilitate more and more considerable emphasis all over the world on them that will somewhat effectively be reduced by less utilization of raw materials. Aiming at diminishing using materials and hence resulting in delicate environment, product lightweight design, indirectly promoting economic and social interest, is developed and refined with multi-agent support in this paper. Not only involving in performance parameter designing, such lightweight design also consider environmental influence, manufacturability and maintainability in the product life cycle, therefore being a collaborative design process with multi-objective optimization and iterative repetition. Firstly, multi-agent based architecture supporting collaborative product lightweight design is put forward, including detailed description of structural

Z. Liu (✉) · K. Pan · X. Gu · Y. Zhang

Institute of Modern Manufacturing Engineering, Zhejiang University,
Hangzhou 310027, China
e-mail: athena@zju.edu.cn

K. Pan
e-mail: pankachn@gmail.com

X. Gu
e-mail: xjgu@zju.edu.cn

Y. Zhang
e-mail: exinanzhizi@sina.com

F. Xu
Business School, University of Shanghai for Science and Technology (USST),
Shanghai 200093, China
e-mail: xufy@usst.edu.cn

composition. Secondly, the core agent-CAX Agent-is introduced thoroughly. Finally, self-adjusting arm shell, as a case study, is redesigned and implemented with such collaborative architecture to validate the operation process of lightweight design.

Keywords Agent · Multi-agent system · Lightweight design

91.1 Introduction

In the era of rapid technological development and economic globalization leading to severe competition among enterprises, customers claim much more steadfastly for such requirements as due date, quality, cost and individuation. Accordingly, to maintain subsistence and keep competent advantages, enterprises are shifting traditional product developing paradigm, moving away from single consideration on technique requirements such as function and cost to on environmental nature of certain problems [1].

Nowadays, numerous countries place sustainable development on strategic level to decrease consumption of energy to save them, to reduce greenhouse gases that deplete our ozone layer and heat the Earth's atmosphere, and to develop green commerce to protect environment. The strength of lightweight design, diminishing utilization of raw material, perfectly amounts to the target of sustainable development as economic and social benefits are also produced. For example, 10% reduction of the weight of one automobile would correspondingly save 6–8% oil [2].

As a gradual process developing from sketch to detail, product lightweight design not only consider the fundamental performance parameter of a product such as size, intensity, longevity, cost and reliability, but also take into account environmental influence, manufacturability and maintainability, resulting in such characters as high technology, high integration of knowledge, etc. Thus, this design plays significant effect on following process planning, manufacturing, sale and maintenance. Generally, two methods for lightweight design are widely accepted: (1) using the materials with high intensity and low density; (2) optimizing the structure of the parts.

In this research, we propose multi-agent based architecture collaborative product lightweight design to hone cooperative lightweight design. And [Sect. 91.2](#) introduces related work. The generic system architecture supporting collaborative product lightweight and its details are described in [Sect. 91.4](#). [Section 91.5](#) concludes the research and briefly suggests future work.

91.2 Related Work

91.2.1 Agent and Multi-Agent System

Multi-agent system (MAS) represents one of the most promising technological paradigms for the development of open, distributed, cooperative, and intelligent software systems [3, 4]. While no uniform definition of Agent is given, briefly speaking, Agent, with relatively high autonomy, is an entity that operates in dynamic environment and is driven by a target. It has some noticeable characters such as autonomy, social ability, reactivity, pro-activeness, mobility, rationality [5–10]. Since the way agents effectively collaborate with each other is similar to that designers accomplish cooperative lightweight design, multi-agent based lightweight design with excellent superiority transcend traditional design driven by passively knowledge support.

91.2.2 Collaborative Product Lightweight Design Systems

Different geographically location of designers and product cycle reduction in the market gives rise to necessary cooperation of lightweight design with the help of computer technology to enhance the efficiency, performance and economy [11]. Based on finite element analysis (FEA), literature [12–14] research lightweight design methods of coal gondola car, heavy truck compartment, car's engine cover plate respectively. Literature [15] applies FEA to optimize the ribs of car body, resulting in higher rigidity and lower cost.

Furthermore, some dominant CAX software such as Pro/Engineer, SolidWorks provides us fine tools for collaborative design, although they lack integral functions and synchronization. A collaborative CAD system needs two kinds of capabilities and facilities: distribution and collaboration [16]. Hence, multi-agent system serves to realize distributive, parallel and effective lightweight design, also implementing more efficient cooperation than Web and eventually bringing more eminent functions.

91.3 Generic System Architecture Supporting Collaborative Product Lightweight

91.3.1 Generic System Architecture

In the premise of having satisfied essential capabilities of product, multi-agent based collaborative lightweight design includes a group of software agents making joint efforts to achieve as lighter quality and more environmental friendly as

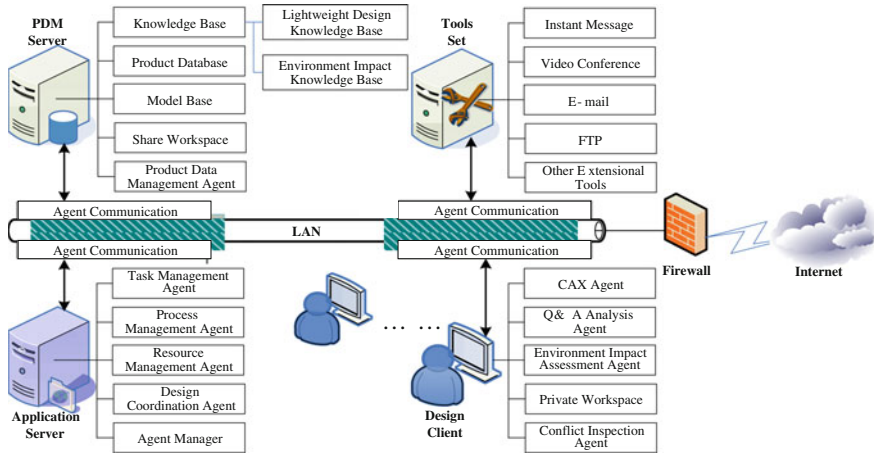


Fig. 91.1 Multi-agent based generic system architecture supporting collaborative product lightweight

possible. Such design, as an intricate process of knowledge discovery, not only demands book learning, but also requires affluent experience and artistry of designers who ought to share and communicate their ideas and knowledge with others to attain the common target.

The multi-agent based generic system architecture supporting collaborative product lightweight is organized as follows, integrating software agents with design tools and designers in an open environment (Fig. 91.1). These agents are Product Data Management Agent, Task Management Agent, Process Management Agent, Resource Management Agent, Design Coordination Agent, CAX Agent, Q&A Analysis Agent, EIA Agent, Conflict Inspection Agent, and Agent Manager while detailed description of them are as follows.

Product Data Management Agent: Managing product information, document and users who should have specific authorization, guaranteeing integrity and logical coherence of data.

Task Management Agent: Dividing complicated tasks into sub-tasks and then rationally assigning them to corresponding designers.

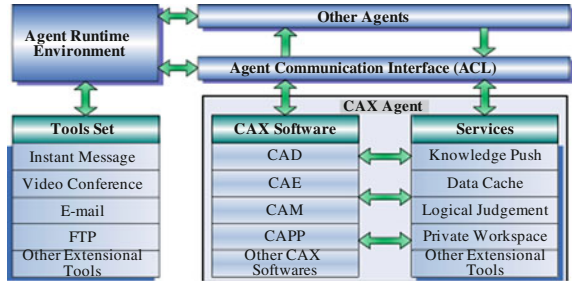
Process Management Agent: Establishing, maintaining and tracking certain task of lightweight design, managing versions of documents in the process of design.

Resource Management Agent: Detecting and managing software and hardware resources in the network nodes.

Design Coordination Agent: Coordinating the design process according to quality & asset assessment, environmental influence assessment and conflict detection results.

CAX Agent: Constituting designing analysis panel to communicate with designers through uniform user interface, providing them the functions of CAD/CAE/CAM/CAPP.

Fig. 91.2 The composition of CAX Agent



Q&A Analysis Agent: Carrying out quality & asset assessment for the results of design.

EIA Agent: Developing environmental influence assessment for the results of design.

Conflict Inspection Agent: Accomplishing conflict detection to help designers solve it with the help of decision knowledge base.

Agent Manager: Managing all the activities of the agents above and utilizing agent negotiation mechanism to uphold their rational, effective, collaborative and timely work.

Connecting with Internet through firewall, LAN serves as a communication channel to connect the entire agents, who could contact with others irrespective of their disparate location, bringing distributive lightweight design into effect. To conveniently allocate and maintain the system, Product Data Management Agent operates in PDM server that is also responsible for managing knowledge base, product database, and model base. Moreover, the knowledge base includes lightweight design knowledge base and environmental influence assessment knowledge base. And Task Management Agent, Process Management Agent, Resource Management Agent, Design Coordination Agent and Agent Manager operate in the Application Server, as CAX Agent, Q&A Analysis Agent, EIA Agent, and Conflict Inspection Agent run in the computer of design client.

91.3.2 CAX Agent

As a key and prominent agent, CAX Agent paramountly assist human designer to develop lightweight design and its composition is shown in Fig. 91.2.

CAX software and some services are integrated and encapsulated in the CAX Agent. On one hand, CAX software carries out lightweight in the design, concerning the effects on mechanics capability, manufacturing and processes. One the other hand, with respect to these services, knowledge push positively provides specific knowledge relating to lightweight to the designers, including the speculative information from knowledge base; data cache help rapidly preread and

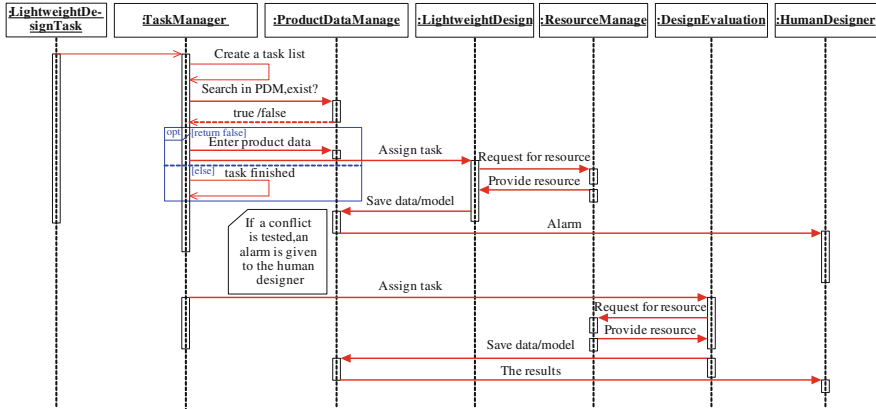


Fig. 91.3 Lightweight design process in the multi-agent system

transitorily store data, therefore promoting the calculation efficiency; logical judgment functions when it is required to decide the status as information is inputted from outside.

Agent Communication interface serves as a channel to help CAX Agent visits the outside, which is realized through Agent Communication language (ACL) defined by Foundation for Intelligent Physical Agents (FIPA). Thereby, different platforms transmit encapsulated information through such communication protocol as Socket, CORBA, RMI, etc.

Tools Set provides auxiliary assistance for lightweight, acting as communication and coordination among human designers.

91.4 Operation Process of the System

Sequence chart is applied to vividly and explicitly show the detailed lightweight design process in the multi-agent system in Fig. 91.3, which is validated perfectly using self-adjusting arm as a case. This arm is an important part of car's brake system. Subsequently, the shell (Fig. 91.4a) of such arm is analyzed and optimized with collaborative lightweight design to describe the design process.

The department of design has been finished the preliminary design of self-adjusting arm shell and save it in the PDM system. Firstly, the lightweight design task of this shell will be sent to Process Management Agent that establishes design task and actuates Task Management Agent, which divides this task to several sub-tasks in a task list. This list is labeled generally with the parameters such as No, task name, starting time, ending time, status, etc. and stored by Process Management Agent for tracking and version control.

Task Management Agent, through Agent Manager, appoints Product Data Management Agent to check, according to Drawing No, whether there is a

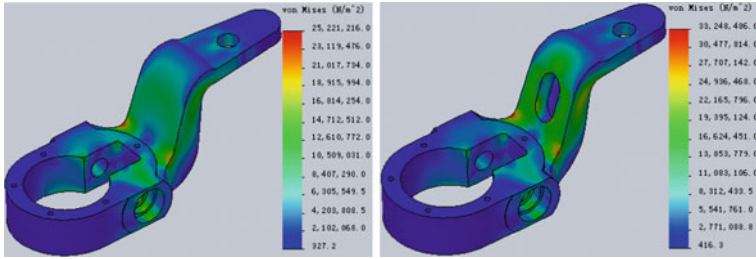


Fig. 91.4 Static analysis results before and after lightweight design for self-adjusting arm shell

lightweight solution of the shell. If there is an extant solution, it is applied directly; otherwise, lightweight design task will be sent to CAX Agent.

After receiving the task, CAX Agent asks for resources from Resource Management Agent. The resources of CAD/CAE serve to accomplish product lightweight design, as CAM/CAPP help to evaluate the effects on processes and manufacturing, such as manufacturability, the facilities to fabricate this product, etc. Then devised model and data will be sent to be saved in Product Data Management Agent, sending messages of task done to Process Management Agent and Task Management Agent.

Task Management Agent continues to dispense different tasks, sending the foregoing solution to Q&A Analysis Agent and EIA Agent to develop assessment on quality, asset and environment. These two agents execute the task also after acquiring resources from Resource Management Agent. The assessment results will be sent to be stored in Product Data Management Agent, sending messages of task done to Process Management Agent and Task Management Agent.

Finally, Product Data Management Agent presents the assessment results to human designers. If the results do not reach anticipatory target, CAX Agent will start another lightweight design solution, involving in successive circulation until the assessment results achieve target.

In the design process above, Product Data Management Agent also activates Conflict Inspection Agent to inspect conflict. If conflict inspected, alert will be sent to human designers and help them solve it. Moreover, quite vital is communication among different agents.

Final lightweight design solution is shown in Fig. 91.4. As mechanics capabilities are satisfied, 15% of previous weight is reduced and no worse effects on quality, asset and environment is discovered.

91.5 Conclusion

With the promising agent technology, this paper proposes multi-agent based architecture supporting collaborative product lightweight design to help designers effectively conduct collaborative lightweight design. CAX Agent, as the key agent,

is introduced at length, and operation process of lightweight design is also demonstrated in detail with self-adjusting arm shell as a case.

Since lightweight design would not become totally automatic, human designers should involve in design. Thus, in future, more concrete work listed as follows should be done.

An endless stream of knowledge and accumulative experience required in collaborative lightweight design reveal that better knowledge base should be designed to carry out significant knowledge push, lending compelling assistance to lightweight design.

Current assessment methods for lightweight design solutions are so sketchy that comprehensive and substantial assessment for such solutions is necessarily demanded.

As a general problem in lightweight design, conflict is very difficult. So, one of provident and considerable emphasis on future research lies on how to excellently eliminate conflicts for human designers.

Acknowledgments This work is supported by Shanghai Leading Academic Discipline Project, Project Number: S30504 and Chinese Science and Technology Support Plan Project (2011BAB02B01).

References

1. George H, Mark H (2006) A sustainable product design model. *Mater Des* 27:1128–1133
2. Benedyk JC (2000) Light metals in automotive applications. *Light Metal Age* 10(1):34–35
3. Hao Q, Shen W, Zhang Z, Park S-W, Lee J-K (2006) Agent-based collaborative product design engineering: an industrial case study. *Comput Ind* 57(1):26–38
4. Marvin M (1987) *The society of mind*. Simon and Schuster, New York
5. Wooldridge MJ, Jennings NR (1995) Intelligent agent: theory and practice. *Knowl Eng Rev* 10(2):115–152
6. Castelfranchi C (1995) Guarantees for autonomy in cognitive agent architecture. In: Wooldridge M, Jennings NR (eds) *Intelligent agents: theories, architectures, and languages*, LNAI, vol 890. Springer, Heidelberg, pp 56–70
7. Genesereth MR, Ketchpel S (1994) Software agents. *Commun ACM* 37(7):48–53
8. Galliers JR (1989) A theoretical framework for computer models of cooperative dialogue, acknowledging multi-agent conflict, Ph.D. thesis, Open University, UK
9. Rosenschein SJ (1985) Formal theories of knowledge in AI and robotics. *New Gener Comput* 3:345–357
10. White JE (1994) Telescript technology: the foundation for the electronic marketplace, White paper, General Magic, CA 94040
11. Wang Y, Nnaji BO (2006) Document-driven design for distributed CAD services in service-oriented architecture. *J Comput Inf Sci Eng* 6(2):127–138
12. Yang A, Zhang Z, Yang J (2007) Lightweight design of open top wagon using updated finite element models. *China Railw Sci* 28(3):79–83
13. Xiaonan W, Hongshuang D, Bing-jie L et al (2010) FEA numerical simulation for weight lightening design via carriage's stress of heavy-duty truck. *J Northeast Univ (Nat Sci)* 31(1):60–63

14. Zhang Y, Zhu P, Chen G et al (2006) The lightweight design of bonnet in auto-body based on finite element method. *J Shanghai Jiaotong Univ* 40(1):163–166
15. Wang L, Prodyot KB, Juan PL (2004) Automobile body reinforcement by finite element optimization. *Finite Elem Anal Des* 40(8):879–893
16. Li WD, Lu WF, Fuh JYH, Wong YS (2005) Collaborative computer-aided design, research and development status. *Comput Aided Des* 37(9):931–940

Chapter 92

Fuzzy Controller Design with Fault Diagnosis System Condition On-line Monitor Using Neural Network

Xiaochun Lou

Abstract This chapter presents a practical method to design and implement a fuzzy controller with system condition on-line monitor for temperatures of continuous soaking process in sugar plant. A new fuzzy control strategy is proposed to improve the control performances. The proposed strategy utilizes an innovative idea based on sectionalizing the error signal of the step response into four different functional zones. The supporting philosophy behind these four functional zones is to decompose the desired control objectives in terms of rising time, settling time and steady-state error measures maintained by an appropriate PID-type controller in each zone. Then, fuzzy membership factors are defined to configure the control signal on the basis of the fuzzy weighted PID outputs of all four zones. A method of system condition on-line monitor using neural network is presented, base on dead time, peak time, percent overshoot, steady state error, rise times, and gain of system step response. The obtained results illustrate the effectiveness of the proposed fuzzy control scheme in improving the performance and intelligent maintenance of the implemented control systems for temperatures of continuous soaking process in sugar plant.

Keywords Fuzzy controller · On-line monitor · Neural network

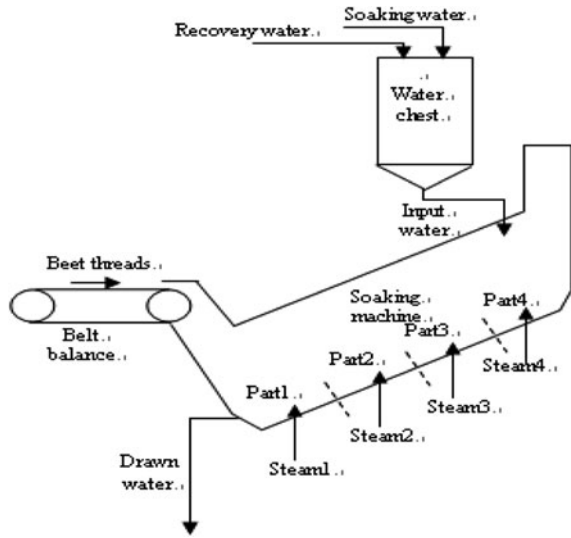
92.1 Introduction

The continuous soaking process (CSP) is an important step in sugar plant. The CSP is carried out in Fig. 92.1. The soaking machine have actuating medium input of the beet threads from belt balance and the input waters from water chest, and

X. Lou (✉)

Hangzhou Vocational and Technical College, 310018 Hangzhou, Zhejiang, China
e-mail: hznwb@163.com

Fig. 92.1 Continuous soaking process



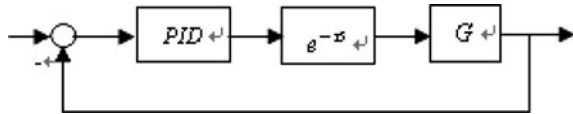
produce output of drawn water, its beet threads levees of four ports are constrained, its temperatures of four ports are contolee by heating steam, the input water is composed of soaking water and recovery water. The controlee variables are flow of drawn water, temperatures of four ports, flow of input water, and water level of water chest, mass of beet threads. The operating variables are four ports flow of heating steam, flow of input water, flow of soaking water, motor speed in belt propeller. The major disturbances are beet threads quantity to get into soaking machine and recovery water quantity [1].

Due to time delay, the temperatures controls of four ports are large influence to product quality. The chapter is an attempt to study this problem.

To analyse this observation, consider the general time delay system in Fig. 92.2.

To improve the control performance under the control delay degradation effect, appropriate control methods should be developed. Although considerable research interest has been paid to the implementation of advanced controllers, PID controllers are still being used in the majority of industrial processes. This is mainly due to the fact that the PID control schemes have a simple structure which can be easily understood and maintained by field engineer. However the tuning of conventional PID remains a difficult task due to insufficient knowledge of the analytical process dynamics. Therefore many classical PID control loops suffer from poor tuning due to the nonlinear and time-varying nature of industrial process. As a consequence, recourse to the automatic PID tuning approach is unavailable in most practical situations for maintaining a consistent performance in the presence of real process uncertainty. This is attracting recent attention from researchers and practicing engineers [2].

Fig. 92.2 General time delay system diagram



Currently, the fuzzy logic control technique has demonstrated great promise to provide a reasonable and effective alternative to the classical controllers in the face of proven model complexity and uncertainty. In this direction, fuzzy PID control type has been the subject of intense interest during the last two decades because of its ability to induce the familiar conventional PID control law on the basis of approximate fuzzy reasoning. Hence, it has received considerable attention in the field of process control to improve its performance in dealing with process model uncertainties compared to its alternative conventional PID counterparts [3].

In the fuzzy PID control approach, the fuzzy rules can be implemented either as an error driven direct control action type or a gain scheduling type [4]. The error driven type controllers constitute the majority in which the rules are expressed to produce the controller output. The gain scheduling type controllers are based on fuzzy tuning rules to adjust the PID gains. Different research attempts have been made to develop fuzzy PID controllers with automatic tuning concepts [5].

This chapter proposes a new fuzzy PID control tuning methodologies in which the fuzzy PID control function is partitioned into four fuzzy regional-based PD, P, PID, PI whose contributions in derivation of the overall PID control output are adjusted on the basis of their fuzzy membership values. Then, the chapter is an attempt to undertake the demonstrations of how control strategies can be practically implemented on real industrial processes.

92.2 Fuzzy Controller Design

92.2.1 Fuzzy Control Strategy

PID control is by far the most commonly used control scheme in process control applications. This scheme, however, suffers from poor parameter tuning, resulting in degraded performance. Real processes are essentially nonlinear and their dynamics change with the operating point, calling for an adaptive PID tuning mechanism. Furthermore, the PID control designer needs to consider the time delay characteristics. For invariable PID tuning, control performance would be degraded due to the stochastic nature of the time delay as experimentally illustrated in the previous section. It has been observed that fuzzy PID controllers can represent better control performances in the time delay processes compared to their conventional counterparts [6]. In the fuzzy PID scheduler, the error signal

plays a decisive role in defining membership values and the control action mainly depends on gain scheduling [7–9]. The proposed method is based on

Fig. 92.3 Proposed membership function based on error percent signal

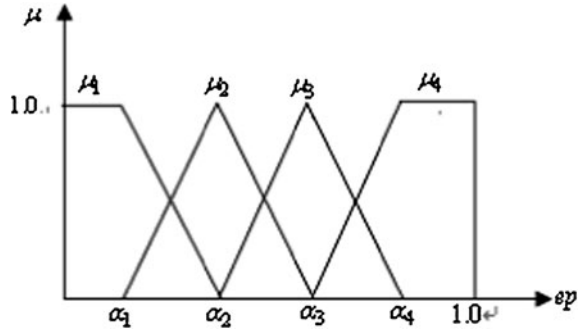
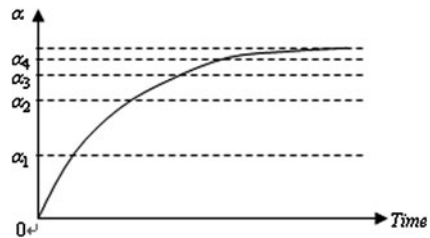


Fig. 92.4 Desired PID zones functionality

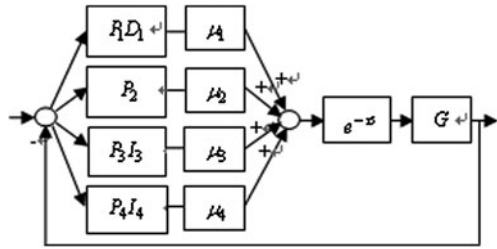


sectionalizing the error percent signal of the step response into three functional zones and fuzzy membership factors are used to configure the overall control signal contributed from the PID outputs of all zones. These functional zones have been derived from our desired functional expectation of PID controller in different zones. To discriminate among the operating zones of the four different PID type controllers, four membership variables (i.e. μ_1 , μ_2 , μ_3 , and μ_4) are defined as shown in Fig. 92.3.

The desired PID controller for the first zone is a PD type due to a large error signal to speed up the response, leading to a minimum rising time. In the second zone, P type is suitable for leading to a minimum rising time when a largish error signal to speed up the response. Next, the controller should minimize the settling time without oscillation or instability; therefore the PI type should be applied in the third zone. Finally, in the case of a small error signal in the fourth zone, a PI type controller should be applied to eliminate or minimize the steady state error. The supporting philosophy of defining these four functional zones is to decompose the desired control objectives in terms of three important control performance measures, which are the rising time, settling time and steady state error maintained by an appropriate PID controller in each zone. The proposed desired PID functionality is depicted in Fig. 92.4 in which α_1 , α_2 , α_3 and α_4 denote the parameters to be determined by designer in order to specify the zones boundaries.

This fuzzy PID control structure offers more robustness due to its ability to set proper actuator high and low limits in the predefined functional zones

Fig. 92.5 Fuzzy PID control algorithm



(i.e., actuator low limit in zone 1 and actuator high limit in zone 4). The robustness advantage could be become more obvious in the case of large time delays. Fuzzy membership factors are used to configure the overall control signal by aggregating the weighted PID outputs of all zones to facilitate a gradual bump-less control transfer between two adjacent control zones, as depicted in Fig. 92.5. In order to increase stability assurance, the proportional gain of the PID controller should be reduced when the feedback signal approaches the set-point. P_1 and D_1 indicate the proportional gain and derivative factor of the PD type controller which is enabled in the first and second zones. P_2 is the proportional gain of the P type controller which is enabled in the second and third zones. P_3, I_3 and D_3 are the proportional gain, derivative factor and integral coefficient of the PID type controller which is enabled in the third and fourth zones while P_4 and I_4 are the proportional gain and integral factor of PI type controller which is enabled only in the fourth zone. Therefore, the output of fuzzy PID controller can be determined as follows in terms of the fuzzy aggregated contribution of each regional PID type controller:

$$\begin{aligned}
 u_c &= \mu_1(P_1(e + D_1 \dot{e})) + \mu_2(P_2e) + \mu_3(P_3(e + I_3 \int edt + D_3e)) + \mu_4(P_4(e + I_4 \int edt)) \\
 &= (\mu_1P_1 + \mu_2P_2 + \mu_3P_3 + \mu_4P_4)e + (\mu_1P_1D_1 + \mu_3P_3D_3) \dot{e} + (\mu_3P_3I_3 + \mu_4P_4I_4) \int edt
 \end{aligned}
 \tag{92.1}$$

where u_c is the overall output of fuzzy PID control signal which is applicable for all zones. Therefore, the output of fuzzy PID control signal for each zone ($u_{c1}, u_{c2}, u_{c3}, u_{c4}$) is derived as follows:

Zone 1:

$$u_{c1} = \mu_1(P_1(e + D_1 \dot{e})) \tag{92.2}$$

Zone 2:

$$u_{c2} = (\mu_1P_1 + \mu_2P_2 + \mu_3P_3)e + (\mu_1P_1D_1) \dot{e} + \mu_3P_3I_3 \int edt \tag{92.3}$$

Zone 3:

$$u_{c3} = (\mu_2P_2 + \mu_3P_3 + \mu_4P_4)e + (\mu_3P_3I_3 + \mu_4P_4I_4) \int edt \tag{92.4}$$

Zone 4:

$$u_{c4} = (\mu_3 P_3 + \mu_4 P_4)e + (\mu_3 P_3 I_3 + \mu_4 P_4 I_4) \int edt \quad (92.5)$$

Examining Eqs. 92.2–92.5 reveals that the proposed fuzzy controller behaves like a gain scheduling PID controller with the following general characteristics:

In the early stage of the response in zone 1 when $\mu_1 = 1$ and $\mu_2 = \mu_3 = \mu_4 = 0$ the controller acts as a PD controller with maximum proportional and derivative gains leading to a fast rising response without overshooting or instability concerns.

As the response moves into zone 2, μ_1 decreases and μ_2 increases leading to a reduction derivative gains and also an increase in the proportional gain to a faster rising response without overshooting or instability concerns.

As the response moves into zone 3, μ_2 decreases and μ_3 increases leading to a reduction of proportional gains and also an increase in the integral gain to make the controller achieve the final steady state error elimination.

As the response enters zone 4, the fuzzy controller behaves as a PI controller. Thus, the steady state-state error is eliminated by the gradual increase in integral gain, the gradual decrease of proportional gain and the absence of derivative gain when the system is about to settle down.

92.2.2 Error Percent Determination

The error percent is defined as follows:

$$ep = \frac{S.P - P.V}{S.P - P.V_0} \quad (92.6)$$

where $P.V_0$ denotes the process value when the set-point changes. Therefore, the complete error percent signal ep is partitioned into three fuzzy sets, as shown in Fig. 92.3, by the following four membership variables:

For

$$ep \geq \alpha_1 : \mu_1 = 1, \mu_2 = \mu_3 = \mu_4 = 0; \quad (92.7)$$

For

$$\alpha_2 \leq ep \leq \alpha_1 : \mu_2 = \frac{\alpha_1 - ep}{\alpha_1 - \alpha_2}, \mu_1 = 1 - \mu_2, \mu_3 = \mu_4 = 0; \quad (92.8)$$

For

$$\alpha_3 \leq ep \leq \alpha_2 : \mu_3 = \frac{\alpha_2 - ep}{\alpha_2 - \alpha_3}, \mu_2 = 1 - \mu_3, \mu_1 = \mu_4 = 0; \quad (92.9)$$

For

$$\alpha_4 \leq ep \leq \alpha_3 : \mu_4 = \frac{\alpha_3 - ep}{\alpha_3 - \alpha_4}, \mu_3 = 1 - \mu_4, \mu_1 = \mu_2 = 0; \tag{92.10}$$

For

$$ep \leq \alpha_4 : \mu_4 = 1, \mu_1 = \mu_2 = \mu_3 = 0. \tag{92.11}$$

In this chapter, the membership variables are utilized to set the proper actuator limits to enhance the desired control system operation. This could be implemented as a low limit for the actuator set-point in zone 1 and a high limit for the actuator set point in zone 4. This limitation scheme has an obvious operational advantage. In zone 1, the process output value is far from the desired set-point. Thus, proper setting of the actuator lower limit can cause a fast rising time response. However, there is a small error between the desired set-point and the actual process value in zone 4. Therefore, reducing the actuator high limit can prevent instability. The operational benefit of this scheme becomes more obvious with large time delays.

92.3 System Condition On-line Monitor Using Neural Network

The system parameters of dead time (τ), peak time (t_p), percent overshoot (M_p), steady state error (e_{ss}), rise times (t_r), and gain (K) are shown to depend on the system condition. They are fixed when systems do normally. To bring about a improvement of intelligent maintenance, we use the system parameters to identified normal or faulted system by means of neural network.

92.3.1 Network Structure

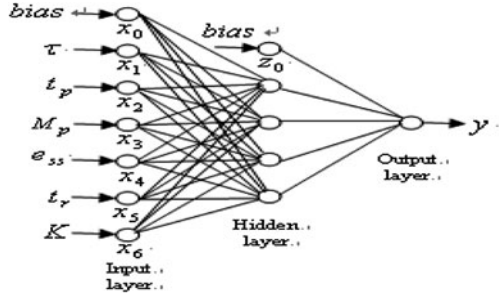
Figure 92.6 shows the neural network structure by means of research determine.

The six inputs to the network included all of the aforesaid step response features. Neurons in the hidden layer used hyperbolic tangent activation functions. The linear output neurons estimate the system condition according to the following relation:

$$y = \sum_{j=1}^4 w_{jk} \tan h \left(\sum_{i=0}^6 w_{ij} x_i \right) + w_0 z_0 \tag{92.12}$$

where x_i is the input and y is the network output. w_{ij} is the adjustable weight between input i and hidden neuron j and w_{jk} is the adjustable weight between hidden neuron j and output neuron k . Note that the bias terms are absorbed into Eq. 92.12 by setting $x_0=1$ and $z_0 = 1$.

Fig. 92.6 Neural network structure



92.3.2 Training Data

The faulty operating conditions were chosen to reflect a wide range of variation in the severity of each failure mode. While possible in practice, situations where two or more faults occur simultaneously were not considered here due to the exceedingly large number of possible fault combinations. This represents a potential limitation of the approach. Step response tests were carried out to ascertain the six performance parameters associated with each operating condition, such as normal operating condition and faulty operating conditions of controller, sensor, actuator and controlled object. The test signal consisted of a series of steps. The set point was increased 10% from current values and back to the 10% after the testing. This sequence was repeated 15 times so that the slight variations in the responses would be captured.

92.3.3 Network Training

The algorithm was implemented by first assigning the weights random numbers on the interval $[-0.5, 0.5]$. The weights were then scaled as follows:

$$w_{ij,new} = \frac{0.6\sqrt[6]{4}}{\sqrt{\sum_{i=1}^6 w_{ij}^2}} w_{ij,old}, \quad i = 1, \dots, 6, \quad j = 1, \dots, 4 \quad (92.13)$$

the bias weights, w_{0j} , were given random initial values on the interval $[-0.6\sqrt[6]{4}, 0.6\sqrt[6]{4}]$. The remaining adjustable parameters were given random initial values on the interval $[1,1]$.

The goal of training is to minimize the mean squared error between the desired network outputs and the actual network outputs which is written for batch mode training, where the error is evaluated over the entire training data set, as

$$E(T) = \frac{1}{P} \sum_{p=1}^P (d_p - y_p(T))^2 \quad (92.14)$$

where d_p is the desired output corresponding to input vector X_p , $y_p(T)$ is the network output corresponding to input vector X_p after training epoch T , and P is the total number of input/output vectors in the training data set. To minimize (92.14), the gradient descent method with momentum was employed. The weights were updated according to

$$\Delta W(T) = -\eta \nabla E|_{W(T)} + \mu \Delta W(T-1) \quad (92.15)$$

In Eq. 92.15, W is the matrix of adjustable weights, η is the learning rate, μ is the momentum term, ∇ is the gradient operator and $\Delta W(T) = W(T) - W(T-1)$. To accelerate convergence, the learning rate was continuously adjusted during training as follows:

$$\eta_{new} = \begin{cases} \rho \eta_{old}, & \Delta E < 0 \\ \sigma \eta_{old}, & \Delta E > 0 \end{cases} \quad (92.16)$$

where ΔE is the change in the network error, $E_{new} - E_{old}$, resulting from the last weight update. Parameters ρ and σ were set to standard values of 1.1 and 0.5, respectively. Setting the momentum term, μ , in (92.15) to 0.9 was also found to increase the speed of learning.

The experimental results proved that the trained network has the capability to detect and identify the various magnitudes of the faults of interest, as they occur singly. Furthermore, the results indicate that the network can accurately estimate fault levels unseen during the training process. This information is vital to a process system condition monitoring strategy since it can be used to detect problems and assist in the timely scheduling of repairs to the system, before a severe failure occurs.

92.4 Practical Implementation

Figure 92.1 shows a temperatures control system in sugar plant processes with large time delay. The controlled variables are temperatures of four ports. The operating variables are four ports flow of heating steam.

Four different PID controllers are employed in different time response zones. The PID parameters in this experiment have been set as follows: P_1 (Proportional gain) = 8, $D_1 = T_{d1}$ (derivative time) = 480 ms, P_2 (Proportional gain) = 7, P_3 (Proportional gain) = 4, $(I_3)^{-1} = T_{i3}$ (Integral time) = 1 s, P_4 (Proportional gain) = 3, $(I_4)^{-1} = T_{i4}$ (Proportional gain) = 0.5 s, Cycle time = 500 ms.

Furthermore, α parameters, determining the boundary of the functional zones in Fig. 92.4, are specified as follow: $\alpha_1 = 0.5$, $\alpha_2 = 0.2$, $\alpha_3 = 0.15$, $\alpha_3 = 0.05$.

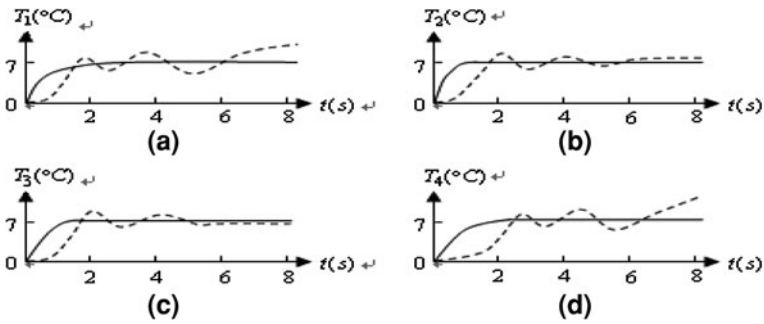


Fig. 92.7 Effect compare

Table 92.1 Condition on-line monitor experimental results

Number	y	Fault by off-line check
1	0.00	Normal condition
2	0.10	Controller open circuit
3	0.11	Control software chaos
4	0.15	Controller short circuit
5	0.30	Sensor open circuit
6	0.36	Sensor wander
7	0.38	Sensor short circuit
8	0.50	Supply pressure of actuator
9	0.63	Percent vent blockage of actuator
10	0.66	Percent diaphragm leakage of actuator
11	0.80	Controlled object leak
12	0.89	Controlled object performance wander
13	0.91	Controlled object blockage

In Fig. 92.7, the solid line shows a effect of the fuzzy PID controller, the dotted line shows a effect of rule PID. The practical observations indicate the promising capability of the proposed fuzzy PID control scheme. System condition on-line monitor experimental results are listed in Table 92.1 along with the normal condition and 13 faulty conditions.

92.5 Conclusions

In this chapter, a new fuzzy PID control approach is proposed to improve the control performance due to the time delay degradation effect. The proposed method is based on sectionalizing the error signal of the step response into four functional zones and fuzzy membership factors are used to configure the control

signal by aggregating the weighted PID outputs of all zones to facilitate a gradual bump-less control transfer between two adjacent control zones, and that A method of system condition on-line monitor using neural network is presented. The obtained results illustrate the effectiveness of the proposed fuzzy control scheme in improving the performance and intelligent maintenance of the implemented control systems for temperatures of continuous soaking process in sugar plant.

References

1. Harbeck C, Faurie R, Scheper T (2004) Application of near-infrared spectroscopy in the sugar industry for the detection of betaine. *Anal Chim Acta* 2:249–253
2. Hang CC, Astrom KJ, Wang QG (2002) Relay feedback auto-tuning of process controllers—A tutorial review. *J Process Control* 12:143–162
3. Mohan BM, Sinha A (2008) Analytical structure and stability analysis of a fuzzy PID controller. *Appl Soft Comput* 8:749–758
4. Mann GKI, Gosine RG (2005) Three-dimensional min–max-gravity based fuzzy PID inference analysis and tuning. *Fuzzy Sets Syst* 156:300–323
5. Woo ZW, Chung HY, Lin JJ (2000) A PID type fuzzy controller with self-tuning scaling factors. *Fuzzy Sets Syst* 115:321–326
6. Li G, Tsang KM, Ho SL (1998) Fuzzy based variable step approaching digital control for plants with time delay. *ISA Trans* 37:167–176
7. Blanchett TP, Kember GC, Dubay R (2000) PID gain scheduling using fuzzy logic. *ISA Trans* 39:317–325
8. Qiao WZ, Mizumoto M (1996) PID type fuzzy controller and parameters adaptive method. *Fuzzy Sets Syst* 78:23–35
9. Kazemian HB (2001) Comparative study of a learning fuzzy PID controller and a self-tuning controller. *ISA Trans* 40:245–253

Part IX
Forensics, Recognition Technologies and
Applications

Chapter 93

Fingerprint Orientation Template Matching Based on Mutual Information

Xuying Zhao, Xiaokun Zhang, Geng Zhao, Rong Qian, Xiaodong Li and Kejun Zhang

Abstract A novel fingerprint matching method based on mutual information is proposed. Fingerprint Orientation is estimated in blocksize and quantized in an appropriate interval. MI of two fingerprint images is calculated in a joint probability distribution of their orientation fields with high noise immunity. The fingerprint matching result involves a combination of MI and minutia matching score using the product fusion strategy. Experimental results show the better performance compared with the alternative approach.

Keywords Fingerprint recognition · Mutual information · Orientation field · Template matching

X. Zhao (✉) · X. Zhang · G. Zhao · R. Qian · X. Li · K. Zhang
Beijing Electronic Science and Technology Institute, No.7 Fufeng Road, Fengtai
District, Beijing 100070, People's Republic of China
e-mail: xyzhao@besti.edu.cn

X. Zhang
e-mail: sam@besti.edu.cn

G. Zhao
e-mail: zg@besti.edu.cn

R. Qian
e-mail: rqian@besti.edu.cn

X. Li
e-mail: lxd@besti.edu.cn

K. Zhang
e-mail: zkj@besti.edu.cn

93.1 Introduction

Shannon entropy is interpreted not only as the amount of information an event gives when it takes place and the uncertainty about the outcome of an event, but also as the dispersion of the probabilities with which the events take place [1]. Therefore, the entropy can be computed for an image as a measure to weight the distribution of the grey values of the image when we define the grey value of a point as an event. A joint histogram of two images is used to estimate a joint probability distribution of their grey values by dividing each entry in the histogram by the total number of entries. Consequently, the Shannon entropy for a joint histogram is introduced as a registration measure for multimodality images by finding the transformation that minimizes their joint entropy [2, 3]. Generally, relative entropy or mutual information is introduced as a more powerful measure for the registration of multimodality medical images that have showed great promise and have been used in a wide variety of applications [4–10]. Liu proposed an algorithm for fingerprint registration by maximization of mutual information [11], and Su presented an approach to face recognition using MI feature [12].

In this Chapter, we do not focus on the grey values of pixels in fingerprint images but on orientation fields that represent fingerprint texture patterns describing the global appearance with high noise immunity. Several algorithms for estimating fingerprint orientation fields have been proposed in the literature [13–16]. The MI of orientation fields of corresponding fingerprint areas is computed as a measure to evaluate similarity of two fingerprints. The orientation fields from two different fingers will have a low MI value while a high MI value will be yielded by orientation fields of two fingerprints that are geometrically aligned. Most of the fingerprint verification systems are prominently based on minutiae matching required to provide a high processing speed and high degree of security and several efficient minutia-based matching approaches have been presented in the literature [17–20]. The principle categories of minutiae are ridge endings and bifurcations which are essential features of fingerprint describing the local fine characteristics, so combining minutiae information and orientation information describing the global patterns can enhance the performance of the fingerprint recognition system.

This Chapter is organized as following. Firstly, mutual information theory is presented in Sect. 93.2. The approach to estimation and quantization of fingerprint orientation is described in Sect. 93.3, while MI of fingerprint orientation is described in Sect. 93.4. In Sect. 93.5, we discuss fingerprint matching algorithm integrating minutia information with orientation information. Finally, experimental results and conclusions are given in Sect. 93.6.

93.2 Mutual Information

Mutual information is a basic concept in information theory, measuring the statistical dependence between two random variables or the amount of information that one variable contains about the other. Two random variables, X and Y , with marginal probability distribution, $p_X(x)$ and $p_Y(y)$, and joint probability distribution, $p_{XY}(x, y)$, are statistically independent if $p_{XY}(x, y) = p_X(x) \cdot p_Y(y)$. The Shannon entropy is defined as

$$H(X) = - \sum_x p_X(x) \log p_X(x), \quad (93.1)$$

$$H(Y) = - \sum_y p_Y(y) \log p_Y(y), \quad (93.2)$$

while the Shannon entropy for joint probability distribution is

$$H(X, Y) = - \sum_{x,y} p_{XY}(x, y) \log p_{XY}(x, y), \quad (93.3)$$

where marginal probability distribution can be given by

$$p_X(x) = \sum_y p_{XY}(x, y), \quad (93.4)$$

$$p_Y(y) = \sum_x p_{XY}(x, y). \quad (93.5)$$

The mutual information of X and Y is calculated by the equation

$$MI(X, Y) = \frac{H(X) + H(Y)}{H(X, Y)}. \quad (93.6)$$

MI measures the degree of dependence of X and Y by computing the distance between the joint distribution $p_{XY}(x, y)$ and distribution associated to the case of complete independence $p_X(x) \cdot p_Y(y)$, analogous to the Kullback–Leibler measure. Considering the fingerprint orientation fields in two images that are to be registered to be random variables X and Y respectively, estimations for the joint and marginal distributions $p_{XY}(x, y)$, $p_X(x)$ and $p_Y(y)$ can be obtained by normalization of the joint and marginal histogram of the overlapped parts of both images. MI is a measure of dependence between two images based on assumption that there is maximal dependence between the orientation fields if they are from the same finger and correctly aligned, while unrelated fingerprints or misregistration will result in a decrease in the measure.

93.3 Estimation and Quantization of Fingerprint Orientation

The method recommended in this paper to estimate fingerprint orientation fields that describe fingerprint texture information mainly involves four steps as below. In the first place, the fingerprint image is divided into blocks of a fixed size of $w \times w$ pixels. Secondly, the gradient of each block is calculated as follows:

$$V_x(i, j) = \sum_{u=i-\frac{w}{2}}^{i+\frac{w}{2}} \sum_{v=j-\frac{w}{2}}^{j+\frac{w}{2}} 2G_x(u, v)G_y(u, v) \quad (93.7)$$

$$V_y(i, j) = \sum_{u=i-\frac{w}{2}}^{i+\frac{w}{2}} \sum_{v=j-\frac{w}{2}}^{j+\frac{w}{2}} (G_x^2(u, v) - G_y^2(u, v)) \quad (93.8)$$

where $G_x(u, v)$, $G_y(u, v)$ are gradients at each pixel (u, v) , w is the width of each block. Thirdly, the direction of each block is estimated by

$$\theta(i, j) = \frac{1}{2} \arctan\left(\frac{V_x(i, j)}{V_y(i, j)}\right) \quad (93.9)$$

in which $\theta(i, j)$ is the orientation angle of the block centered in (i, j) . At last, some smoothing work is carried out on the orientation fields. Generally the texture line angle is limited to $[-\frac{\pi}{2}, \frac{\pi}{2})$ by

$$\phi(i, j) = \begin{cases} \theta(i, j) + \pi/2 & \theta(i, j) \leq 0 \\ \theta(i, j) - \pi/2 & \theta(i, j) > 0 \end{cases} \quad (93.10)$$

Suppose the quantization parameter is N , then we define a quantization interval by equally dividing $[-\frac{\pi}{2}, \frac{\pi}{2})$ and the direction angle is assigned to an integer value by

$$I\phi(i, j) = \left\lceil \frac{\phi(i, j)}{\pi} \times 2^N \right\rceil + 2^{N-1} \quad (93.11)$$

where $\lceil \cdot \rceil$ denotes the nearest integer function. Accordingly, Fingerprint orientation template consists of block-size orientation fields in which the quantized angle value ranges from 0 to $2^N - 1$. Taking $N = 4$ for example, the quantization orientation field is shown in Fig. 93.1b corresponding to the fingerprint image in Fig. 93.1a from which we can find that the quantized data includes the main texture information of the source image.

93.4 MI of Fingerprint Orientation

Two fingerprint images are aligned by determining a rigid transformation using minutiae coordinates and minutiae angles of corresponding pairs. The query fingerprint is lapped over to the reference one as is shown in Fig. 93.2, and the

Fig. 93.1 Contrast between the fingerprint image and the quantized orientation field

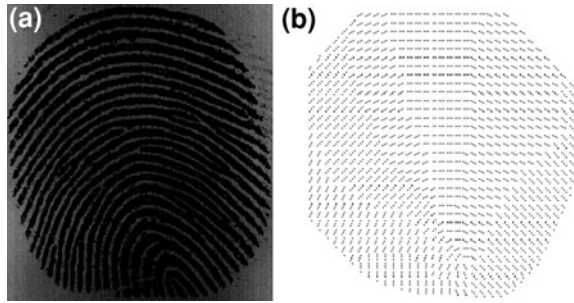
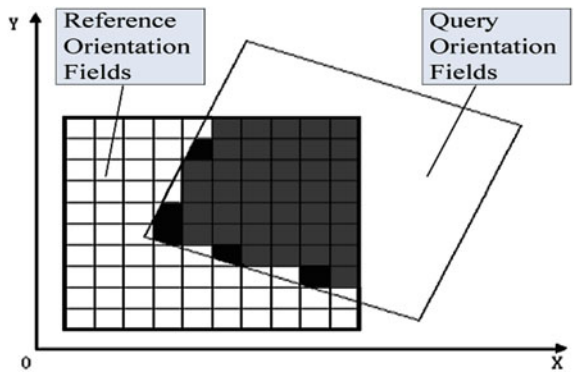


Fig. 93.2 Sketch map of overlapping orientation fields of query fingerprint and reference one



fingerprint orientation fields are recalculated on the transformed image with the same blocksize.

The joint probability distribution $P_{XY}(x, y)$ of the orientation fields can be estimated by counting the number of times $p_n(x, y)$ each quantized block direction value occurs in the overlapping area and dividing those numbers by the total number of occurrences. It is

$$P_{XY}(x, y) = \frac{p_n(x, y)}{\sum_{x=0}^{2^N-1} \sum_{y=0}^{2^N-1} p_n(x, y)} \tag{93.12}$$

The marginal probability distribution, $p_X(x)$ and $p_Y(y)$, of query and reference fingerprint can be obtained by (93.4), (93.5), which is rewritten as

$$P_X(x) = \sum_{y=0}^{2^N-1} P_{XY}(x, y) \tag{93.13}$$

$$P_Y(y) = \sum_{x=0}^{2^N-1} P_{XY}(x, y) \tag{93.14}$$

The MI of two fingerprint images is evaluated by (93.1, 93.3, and 93.6) which is essentially a normalized measure of mutual information that is less sensitive to changes in overlap.

93.5 Fingerprint Matching

In order to evaluate the similarity between two fingerprint impressions, the matching score M is calculated as a measure of the similarity level. Suppose there are maximum P pairs of minutiae matched in two fingerprint images, the matching score M_s in minutia-based matching approaches can be given by:

$$M_s = \frac{P^2}{M \times N} \quad (93.15)$$

where M , N are the number of minutiae in query and reference fingerprints respectively.

In this Chapter, we compute the final matching score M using the product fusion strategy to combine the two classifiers based on minutia and orientation features.

$$M = \begin{cases} M_s \times \frac{MI}{T_m}, & \text{if } C_{b1}, C_{b2} \geq C_T, \\ M_s \times \frac{T_0}{T_m}, & \text{others.} \end{cases} \quad (93.16)$$

C_{b1} and C_{b2} are ratios of overlapped area to query and reference fingerprint images estimated by dividing the overlapped blocks by the total blocks available in each fingerprint. The threshold C_T of the ratio is set to an experiential value in the range from 0.3 to 0.6. The modified minutiae matching score is adopted when the mutual information can not be correctly evaluated if the overlapped area ratio is less than C_T . T_0 can be set to an experiential value while T_m is the maximum value of MI.

93.6 Experimental Results and Conclusion

The experiments reported in this Chapter have been implemented on Pattek fingerprint database in CASIA, in which the images were captured by the capacitor sensor of Veridicom company. The database contains 4000 fingerprint images of size 300×300 pixels captured at a resolution of 500 dpi, from 200 fingers with 20 impressions per finger. The captured images vary in quality, direction and position. Using the same minutiae sets, we conducted a set of experiments to compare our matching method with the minutia template matching method. The matching times of fingerprint images from the same finger are $200 \times C_{20}^2 = 38000$, while the matching times of those No.1 fingerprint images from the different fingers are

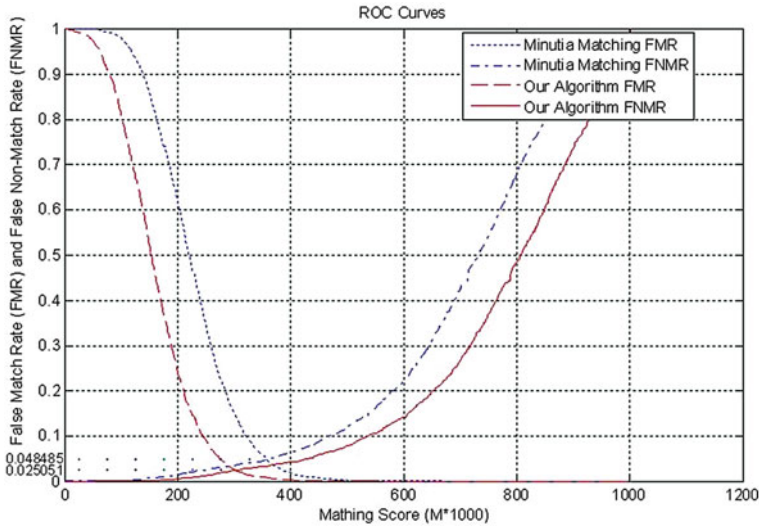


Fig. 93.3 ROC curves obtained with our algorithm and minutia matching algorithm

$C_{200}^2 = 19900$. From ROC curves, we can see that our algorithm outperforms the algorithm based on minutia matching obviously, as is shown in Fig. 93.3.

In conclusion, this Chapter presents a new fingerprint matching algorithm based on mutual information that is computed in a joint probability distribution of fingerprint orientation fields. The performance of the fingerprint recognition system is improved by integrating MI of orientation that describes global information with minutia based matching result.

References

1. Plum JPW, Maintz JBA, Viergever MA (2003) Mutual-information-based registration of medical images a survey. *IEEE Trans Med Imaging* 22(8):986–1004
2. Woods RP, Cherry SR, Mazziotta JC (1992) Rapid automated algorithm for aligning and reslicing PET images. *J Comput Assist Tomogr* 16(4):620–633
3. Woods RP, Mazziotta JC, Cherry SR (1993) MRI-PET registration with automated algorithm. *J Comput Assist Tomogr* 17(4):536–546
4. Collignon A, Vandermeulen D, Suetens P, Marchal G (1995) 3D multimodality medical image registration using feature space clustering. In: *Computer vision, virtual reality, and robotics in medicine*, vol 905. Springer, London, pp 195–204
5. Studholme C, Hill DLG, Hawkes DJ (1995) Multiresolution voxel similarity measures for MR-PET registration. In: *Information processing in medical imaging*. Kluwer, Dordrecht, pp 287–298
6. Collignon A, Maes F, Delaere D, Vandermeulen D, Suetens P, Marchal G (1995) Automated multi-modality image registration based on information theory. In: *Information processing in medical imaging*. Kluwer, Dordrecht, pp 263–274

7. Wells WM III, Viola P, Kikinis R (1995) Multi-modal volume registration by maximization of mutual information. In: *Medical robotics and computer assisted surgery*. Wiley, New York, pp 55–62
8. Cover TM, Thomas JA (1991) *Elements of information theory*. Wiley, New York
9. Studholme C, Hill DLG, Hawkes DJ (1999) An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognit* 32(1):71–86
10. Maes F, Vandermeulen D, Suetens P (1999) Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information. *Med Image Anal* 3(4):373–386
11. Liu L, Jiang T, Yang J, Zhu C (2006) Fingerprint registration by maximization of mutual information. *IEEE Trans Image Process* 15(5):1100–1110
12. Su H-T, Feng D-D, Wang XY, Zhao R-C (2003) Face recognition using hybrid feature. In: *Proceedings of the second international conference on machine learning and cybernetics, Xian, 2–5 Nov 2003*
13. Karu K, Jian AK (1996) Fingerprint classification. *Pattern Recognit* 29(3):389–404
14. Kass M, Witkin A (1984) Analyzing oriented patterns. *Comput Vis Graph Image Process* 37:263–285
15. Zhou J, Gu J (2004) A model-based method for the computation of fingerprints' orientation field. *IEEE Trans Image Process* 13(6):821–835
16. Nie D, Ma L, Xiao XZ, Xiao S (2005) Optimization based fingerprint direction field estimation. In: *Proceedings of 27th annual international conference of the engineering in medicine and biology society 2005 (IEEE-EMBS 2005)*, pp 6265–6268
17. Gamble F, Frye L, Grieser D (1992) Real-time fingerprint verification system. *Appl Opt* 31(5):652–655
18. Wilson C, Watson C, Paek E (1992) Effect of resolution and image quality on combined optical and neural network fingerprint matching. *Pattern Recognit* 33(2):317–331
19. Jain AK, Prabhakar S, Hong L, Pankanti S (2000) Filterbank-based fingerprint matching. *IEEE Trans Image Process* 9(5):846–859
20. Lee CJ, Wang SD (1999) Fingerprint feature extraction using gabor filters. *Electron Lett* 35(4):288–290

Chapter 94

Human Action Recognition Algorithm Based on Minimum Spanning Tree

Yi Ouyang and Jianguo Xing

Abstract Human pose tracking recognition algorithm for monocular video was proposed to model human part parameters using video features combination with 3D motion capture data. Firstly three-dimensional data projection constraint graph structure was defined. To simplify the reasoning process, a constraint graph of the spanning tree construction algorithm and the balancing algorithm were proposed. Combination with the proposed function mechanism, spanning tree of constraint graph and Metropolis–Hastings method, human motion under monocular video can be tracking and recognition, inferring the 3D motion parameters. By using data-driven (Markov chain Monte Carlo MCMC) and constrain map, human motion limb recognition algorithm is proposed, and the method can be applied to data-driven online human behavior recognition. Experimental results show that the proposed method can recognize human motion action automatically and accurately in monocular video.

Keywords Human action · Markov chain · Belief propagation · 3D human action estimation

94.1 Introduction

3D modeling of human motion is the mainly support technology for human–computer interaction, animation design, intelligent detection systems and security analysis application system. For non-linear, complex diversity, the lack of a clear classification structure and other characteristics some reasons, human motion

Y. Ouyang (✉) · J. Xing
Zhejiang GongShang University, 310018, Hangzhou, China
e-mail: oyy@mail.zjgsu.edu.cn

modeling is so complicated. At present, for tracking human motion in video image is divided into two categories: based on the Generative Models [1–3] of the human motion analysis and discriminate methods of analysis [4–7]. Only by observing the movement of the current state of time is often difficult to determine the category of the movement, in order to reduce the computational complexity, constant observation sequence is based on the assumption of conditional independence, the result does not reflect the dependence of the time series. This chapter presents an 3D human motion library based on data-driven Markov chain Monte Carlo (MCMC) method in monocular video to track human motion, the algorithm's basic idea is, using the human body motion capture data building the basic movement database, and at different perspective projection clustering the human silhouette; with [8, 9] method. Monocular video were detected in the human body, and the body can be accurately split the position of the body; finally, 3D human motion reasoning appearance model algorithm, using time constraints of the model to track the target and graph-driven MCMC and the combination of basic movements, is applied data-driven online actions recognition.

94.2 Human Motion Levels Model

Using only monocular video for 3D reconstruction of human motion is the kind of ill-posed problem. Mainly due to monocular video camera lack a lot of spatial information. In order to reconstruct each 3D human pose from frames of video images, the basic movement for the human body through the motion sensor to establish a database of basic movements, we use CMU database [17] combined with VOC image database [10] information.

94.2.1 Human Body and Texture Model

The human body model (HBTM) is constructed with the torso and limb, in which the location and motion parameters from the torso direction and angle of rotation between the limb composition. Through the latent variable, shape parameter describes the relationship between the torso and limbs, combined with the common physical description of the color histogram of human appearance. Human joints points is composed by 14 key parts, and pose represented by six-dimensional vector G , that is the global body's position and direction of rotation. J represents the angle of rotation between joint points. These parameters be modeled by the prior distribution $P(G)$ and $P(J)$. They can be composed by a set of training images, and assume that the probability distribution of approximately Gaussian distribution, and joint points of non-adjacent location parameters are independent. Skin texture model (ST) is composed by $ST = [C_1, C_2, C_3]^T$ the three parameters,

respectively, the hands, face and torso rectangle, the prior distribution $P(C)$ is learned from the histogram of training data

94.2.2 Cluster-Based Motion Model of the Relevant Action

As the larger dimension of human motion sequence, and there are large data redundancy, according to the basic model of human body motion data sequence [18] and minimum distance between limbs, we cluster these data at first. This chapter presents the Relevant Action Cluster (RPC) for human action analysis. All the input images M select a subset of nodes N in the cluster constraint, which have the largest cluster nodes. Each cluster node will be mapping to a set of images having the same 3D model. This will not only satisfy the action monocular camera image recognition, simply extend RPC node number of different angles of the projected image, you can improve recognition accuracy having more camera images.

Definition 1 The Relevant Action Cluster (RPC) has $\text{sim}(RPC_i, RPC_j) > \varepsilon$, the constraints of the RPC having similar shape, the difference data between the RPC less than $1 - \varepsilon$

Let $I = [I_1, I_2, \dots, I_M]$ to represent the image features of human limbs motion. Where $I_i = \{x_k, \theta_j, d, c\}$, M is the number of the RPC nodes, x is the limb center point, and the subscript k means limbs index, superscript j represents the horizontal projection angle index, d as the motion data frame number, c as the basic movement types. 2D projection image feature data is captured from the perspective of the level of 3D motion capture data at different angles. We experiment with 18 different angles, adjacent angles of 10 degrees intervals, and build 2D human motion graph model, where each node corresponds to a cluster RPC node.

Definition 2 RPC graph is $G = (V, E, W)$, where node set V for the RPC, E for the set of edges between v_i, v_j nodes, edges between two nodes which have weight $w_{i,j}$, that is, weighted graph G is undirected, S for the RPC node set size. When at least two nodes RPC images similar, there is a link between two nodes.

94.3 Spanning Tree Algorithm Based on RPC

Using traditional minimum cost spanning tree algorithm, lose some dependencies between nodes, while the uncertainty of tree depth will cause many computing problems, such as the time for determines the reasoning. In order to overcome these problems, we propose a weighted spanning tree construct algorithm based on RPC nodes, node merging algorithm idea is to reduce the size of the spanning tree

node, the node splitting to resolve connectivity issues, and make use of spanning tree balancing algorithm remain bounded spanning tree depth.

Definition 3 RPC spanning tree as $T = (V, E, W)$, V as node set, E for the edge set, W is the edge weight set; where each node can only have one parent node, can have many child node.

Node merge: When the same parent node, child node exists between the two sides, when the two sub-nodes associated with edge e intensity threshold intensity is less than Q can be considered an approximate property of the two child nodes, so the two nodes into one new node, its child nodes also point to the node; the two adjacent nodes and edges e deleted. The weight of parent node and the edge of the node is $w = \max(w_i, w_j)$.

Node splitting: When they find a child node also points to more than one parent node, remove the other nodes associated with the strength of the weak side of the parent, to ensure the dependencies between the nodes, will remove even the side of the parent node $N'_p(e)$ connected, that is according to formula (94.1) with the reservation side of the parent node to be updated.

$$N_p(e) = N_p(e) \cup (N_p(e) - N'_p(e)) \tag{94.1}$$

Spanning tree balance algorithm

Let Child (N), said child nodes of set N , the specific algorithm is as follows:

Step 1 Calculation of the depth of sub-tree and distance vector $Dist(i)$;

Step 2 For each sub-tree of length greater than H

From sub-tree node N , the child nodes of N increase the potential edge e' , and weight as

$$w' = w_p * \left(\frac{P(N_i)}{\sum_{j \in F_i/i} P(N_j)} \right) * w_c$$

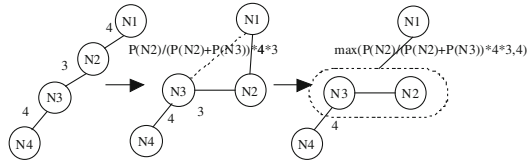
Step 3 Consolidate the node Child (N) and the Child (Child (N)).

Step 4 Modified sub-tree depth as $Dist(i) - 1$, Goto Step 1

For the sub-tree length L is longer than the combined number of sub-root nodes and root node

Assuming the tree depth of 4, when the depth of sub-tree is greater or equal to 4, it will be balanced, as shown in Fig. 94.1. Let $N1$ for the sub-root node, then the pair node $N2, N2, N3$ to merge child nodes, while increasing the edge, its weight $w' = \frac{P(N2)}{P(N2+N3)} * 3 * 4$, the combined new node $N1$ to the $N2$ and $N3$ the edge weight as $w_{1,3} = \max\left(\frac{P(N2)}{P(N2+N3)} * 3 * 4, 4\right)$, RPC node cluster as Fig. 94.1.

Fig. 94.1 Balancing for spanning tree of RPC graph



94.4 Human Actions Recognition Algorithm

For human action recognition, such as [3, 11] described the treatment effect directly through the video image is poor, by [3] inspired, we use segmented human motion recognition technology. We first through the HOG and deformable components of the human body detection method [8, 9] for human motion detection on the first stage; then the spanning tree node with RPC on human action recognition, through this method can be more accurate detection of the human range, and gives the location of the body; finally, the detection results were sent to the third phase of human space action reasoning.

3D human motion estimation, parameter estimation for the body center is particularly important, it is to consider the perspective of the relationship between operation characteristics and spatial reasoning.

94.4.1 The 2D Model Reasoning Based on the RPC

Let $\pi(\cdot)$ as tree node probability, the reasoning process as follow:

$$\begin{aligned} \pi(x_v|x_{-v}) &\propto \pi(x), \quad \text{and} \quad \pi(x) = \prod_{v \in V} \pi(x_v|x_{pa(v)}) \\ \pi(x_v|x_{-v}) &= \pi(x_v|x_{pa(v)}) \times \prod_{i,v \in pa(i)} \pi(x_i|x_{pa(i)}) \end{aligned} \tag{94.2}$$

where V is the set of RPC nodes, $pa(i)$ as the parent node of node i ; $-v$ means that all the nodes in V except v .

Prior distribution: Human model for the parameters of each component of the state vector are represented by $X_t^i = \{G,S,C,M\}$ in T frame, where G represents the global position and rotation parameters, S as the parameters that shape, C as that skin color parameters, t denotes the frame number. For simplicity, assuming these parameters independent, prior probability distribution as follow:

$$p(X_t) \propto p(G)p(S)p(C)p(M) \tag{94.3}$$

These prior probabilities will be combined with image constitutes a posteriori probability function.

94.4.2 The 3D Human Motion Reasoning

The evolution of 3D model of human motion is a known constraint from the start node, by traversing the spanning tree structure, by calculating the maximum a posteriori body image and body movements the best configuration. In order to analyze each type of action, we calculated for each type of action corresponding to the maximum probability, this value was used to measure the movement types of confidence.

Let T frame image sequence of the state vector expressed by $\{X_1, X_2, \dots, X_T\}$. On the shape of human body movement and state of motion relative to the color change more easily. Thus the shape parameters and dynamic parameters should be adjusted so that the object and image in the human motion is consistent, if only to observe the image associated with the current state of the condition. Image prior probability of the state of the human body can be decomposed into a state model with a series of conditions a priori probability of the product:

$$p(X) = \frac{1}{z} p(X_1) \prod_{t=1}^{T-1} p(X_{t+1}|X_t) \quad (94.4)$$

where Z is the normalization constant. The prior probability sample data can be learned by the training the motion capture data. Conditional probability can be approximated by normal distribution:

Dynamic model in which the covariance matrix, which consists of [18] obtained motion data learning for different types of human movement, such as its value is different. For the state parameters calculated by the probability function. Define the image of the probability function by four parts are: location relationship of limbs; background color difference; human skin color and feature matches specific probability function as [12] presented the definition.

94.4.3 Proposal Function

In [3, 13, 14] were used to top-down [15] method, and [16] method of reasoning to human action, often used for the Metropolis–Hastings MCMC algorithm, the algorithm is the key proposed function of choice, usually by way of random walk, using the proposed function is to generate candidate solutions state. In theory it can produce the entire state sequence of candidate solutions, but more loops in this way. The proposal function will determine the choice of the function convergence speed. This chapter may be considered state of the current state estimate, the previous state and next state and the model state and the image co-decision.

The previous state X_t^* : Human PRC generated by the current state of the dynamic model estimated parameters, the proposal function is:

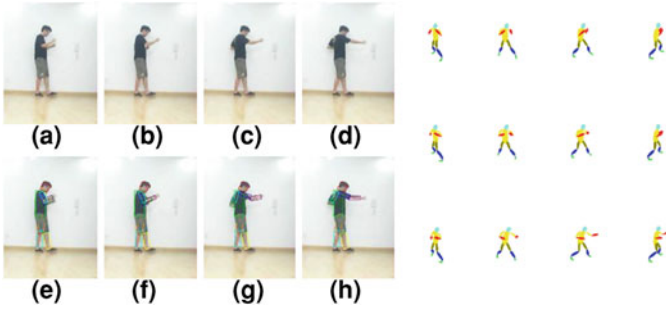


Fig. 94.2 Recognition for running motion sequence

$$q(X_t^*|X_{t-1}) = q(|X_t^* - X_{t-1}|) = s \tag{94.5}$$

where s is random number between $[0,1]$.

Using of back propagation for spanning tree of the RPC, to obtain another estimate of the current state, after propagation through, making the current state of the estimates take into account future trends.

The final proposal function as follow:

$$q(X_t^*|X_t, M^*, I_t, X_{t-1}, X_{t+1}) = w_1 \times q(X_t^*|X_{t-1}) + w_2 \times q(X_t^*|X_t, M^*, I_t) + w_3 \times q(X_t^*|X_{t+1}, M^*) \tag{94.6}$$

where is the weight factor, and the factor used to adjust the proportion between the various proposed components, experiments were taken 0.3, 0.4, 0.3.

94.5 Experiment Analysis

More than 200 were collected from a subset of the different motion sequences, each containing 10 sample subset of actions, each action of 18 different angles from the projection, one of the motion sequence in Figs. 94.2 and 94.3. We test 20 kinds of basic movement, which contains a variety of action walking, running, jumping, kicking, boxing. For different types of data movement, we estimated the depth of Z-axis parameters for the test identification error. A group for general walking, B group running, C group hand boxing, were set before the experiment the depth of the spanning tree when the RPC was 30, this method and Data-MCMC [3] methods and do not use RPC model MCMC method of reasoning directly compare the experimental error of measurement such as Table 94.1:.

The proposed action recognition algorithm based on RPC. We compared it with non-RPC structures methods and Data-MCMC method. Our method improved the recognition accuracy. The reason is considered a priori three-dimensional information of human basic movement, the results shown in Table 94.2.

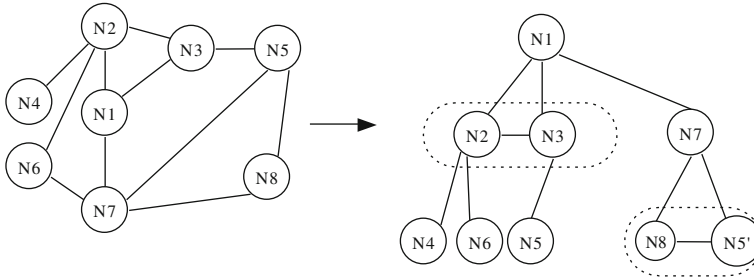


Fig. 94.3 Spanning Tree for RPC graph

Table 94.1 Recognition errors of three class motion using different algorithm

Weighted average errors	Walking	Running	Boxing
Non RPC model	25.35	30.18	35.54
Data-MCMC	24.47	27.61	26.53
Our method	21.36	23.15	24.14

Table 94.2 Errors of human parts for walking sequence

Error type	Torso	RUL	RLL	RSL	RUA	RLA
Center	16.21	17.24	15.21	15.12	13.52	15.75
Depth(Z)	11.2	14.2	12.5	10.32	12.1	13.4
Orient	7.58	5.12	3.12	4.14	6.78	4.34

Acknowledgments This work is supported by the Science and Technology Department of Zhejiang Province in China No. 2008C14100 and Department of Education of Zhejiang Province Project NO 1130KZ710019G

References

1. Agarwal A, Triggs B (2006) Recovering 3D Human pose from monocular images. *IEEE Trans Pattern Anal Mach Intell* 28(1):44–58
2. Sminchisescu C, Kanaujia A, Metaxas D (2006) Learning joint top-down and bottom-up processes for 3d visual inference. In: *IEEE Computer Society conference on computer vision and pattern recognition (CVPR'06)*, pp 1743–1752
3. Lee MW, Nevatia R (2008) Human pose tracking in monocular sequence using multilevel structured models. *IEEE Trans Pattern Anal Mach Intell* 31(1):27–38
4. Weiwei G, Patras I, Queen M (2009) Discriminative 3D human pose estimation from monocular images via topological preserving hierarchical affinity clustering. In: *IEEE 12th international conference on Computer vision workshops (ICCV Workshops), 2009*, pp 9–15
5. Agarwal A, Triggs B (2004) 3D human pose from silhouettes by relevance vector regression. *IEEE Comput Soc Con Comput Vis Pattern Recognit (CVPR'04) 2*, pp 882–888
6. Elgammal A, Lee CS (2004) Inferring 3D body pose from silhouettes using activity manifold learning. *IEEE Comput Soc Con Comput Vis Pattern Recognit (CVPR'04) 2*, pp 681–688

7. Lv F, Nevatia R (2007) Single view human action recognition using key pose matching and viterbi path searching. *IEEE Comput Soc Con Comput Vis Pattern Recognit (CVPR'04)* 4, pp 1–8
8. Felzenszwalb P, McAllester D, Ramanan D (2008) A discriminatively trained, multiscale, deformable part model. *IEEE Comput Soc Con Comput Vis Pattern Recognit (CVPR'08)*, pp 1–8
9. Felzenszwalb P, Girshick R, McAllester D, Ramanan D (2009) Object detection with discriminatively trained part-based models. *IEEE Trans Pattern Anal Mach Intell* 32(9): 1627–1645
10. Everingham M, Van-Gool L, Williams CKI, Winn J, Zisserman A (2008) EB/OL.2008. The PASCAL VOC 2008 Results. CMU, Cmu motion capture library. EB/OL.2007 <http://mocap.cs.cmu.edu>
11. Lee M, Nevatia R (2005) Dynamic human pose estimation using arkov chain Monte Carlo approach motion. *IEEE workshop motion video Comput 2005. WACV/MOTIONS'05* 2, pp 168–175
12. Lee M, Cohen I (2004) Proposal maps driven mcmc for estimating human body pose in static images. *IEEE Comput Soc Con Comput Vis Pattern Recognit (CVPR'04)* 2, pp 334–341
13. Tu ZW, Zhu SC (2002) Image segmentation by data-driven markov chain monte carlo. *IEEE Trans Pattern Anal Mach Intell* 24(5):657–672
14. Zhao T, Nevatia R (2004) Tracking multiple humans in crowded environment. *Proc IEEE Conf Comput Vis Pattern Recognit* 2:406–413
15. Zhu S, Zhang R, Tu Z (2000) Integrating bottom-up/top-down for object recognition by data driven markov chain monte carlo. *Proc IEEE Conf Comput Vis Pattern Recognit* 1, pp 738–745
16. Ramanan D, Forsyth DA, Zisserman A (2004) Strike a pose: tracking people by finding stylized poses. *IEEE Comput Soc Con Comput Vis Pattern Recognit (CVPR'04)* 1, pp 271–278
17. Wang Y, Zhang NL, Chen T (2008) Latent tree models and approximate inference in Bayesian networks. *J Artif Intell Res* 32(1):879–900

Chapter 95

Analysis of Selvi et al.'s Identity-Based Threshold Signcryption Scheme

Wei Yuan, Liang Hu, Hongtu Li and Jianfeng Chu

Abstract Signcryption can realize the function of encryption and signature in a reasonable logic step, which can lower computational costs and communication overheads. In 2008, Selvi et al. proposed an identity-based threshold signcryption scheme. In this chapter, we show that the threshold signcryption scheme of Selvi et al. is vulnerable if the attacker can replace the group public key. Then we point out that the receiver uses the senders' public key without any verification in the unencrypt stage cause this attack.

Keywords Identity-based · Signcryption · Bilinear pairing · Cryptanalysis · Attack

95.1 Introduction

Encryption and signature are the two basic cryptographic tools offered by public key cryptography for achieving confidentiality and authentication. Signcryption can realize the function of encryption and signature in a reasonable logic step

W. Yuan · L. Hu · H. Li · J. Chu (✉)

Department of Computer Science and Technology, Jilin University, Changchun, China
e-mail: chujf@jlu.edu.cn

W. Yuan
e-mail: yuanwei1@126.com

L. Hu
e-mail: hul@mails.jlu.edu.cn

H. Li
e-mail: li_hongtu@hotmail.com

which is proposed by Zheng [1] in 1997. Comparing to the traditional way of signature then encryption or encryption then signature, signcryption can lower the computational costs and communication overheads. As a result, a number of signcryption schemes [2–8] were proposed following Zheng’s work. The security notion for signcryption was first formally defined in 2002 by Baek et al. [9] against adaptive chosen ciphertext attack and adaptive chosen message attack. The same as signature and encryption, signcryption meets the attributes of confidentiality and unforgeability as well.

In 1984, Shamir [10] introduced identity-based public key cryptosystem, in which a user’s public key can be calculated from his identity and defined hash function, while the user’s private key can be calculated by a trusted party called Private Key Generator (PKG). The identity can be any binary string, such as an email address and needn’t to be authenticated by the certification authentication. As a result, the identity-based public key cryptosystem simplifies the program of key management to the conventional public key infrastructure. In 2001, Boneh and Franklin [11] found bilinear pairings positive in cryptography and proposed the first practical identity-based encryption protocol using bilinear pairings. Soon, many identity-based [12–15] and other relational [16–18] schemes were proposed and the bilinear pairings became important tools in constructing identity-based protocols.

Group-oriented cryptography [19] was introduced by Desmedt in 1987. Elaborating on this concept, Desmedt and Frankel [20] proposed a (t, n) threshold signature scheme based RSA system [21]. In such a (t, n) threshold signature scheme, any t out of n signers in the group can collaboratively sign messages on behalf of the group for sharing the signing capability.

Identity-based signcryption schemes combine the advantages of identity-based public key cryptosystem and Signcryption. The first identity-based threshold signature scheme was proposed by Baek and Zheng [22] in 2004. Then Duan et al. proposed an identity-based threshold signcryption scheme [23] in the same year by combining the concepts of identity based threshold signature and encryption together. However, in Duan et al.’s scheme, the master-key of the PKG is distributed to a number of other PKGs, which creates a bottleneck on the PKGs. In 2005, Peng and Li proposed an identity-based threshold signcryption scheme [24] based on Libert and Quisquater’s identity-based signcryption scheme [25]. However, Peng and Li’s scheme does not provide the forward security. In 2008, another scheme [26] was proposed by Fagen Li et al., which is more efficient comparing to previous scheme. However, Selvi et al. pointed out that Fagen Li et al.’s scheme is not equilibrium between the usual members and a dealer called clerk in Fagen Li et al.’s scheme and proposed an improved scheme [27].

In this chapter, we show that the threshold signcryption scheme of Selvi et al. is vulnerable if the attacker can replace the group public key. Then we point out that the receiver uses the senders’ public key without any verification in the unsigncrypt stage cause this attack.

95.2 Preliminaries

95.2.1 Bilinear Pairing

Let G_1 be a cyclic additive group generated by P , whose order is a prime q , and G_2 be a cyclic multiplicative group with the same order q . A bilinear pairing is a map $e : G_1 \times G_1 \rightarrow G_2$ with the following properties:

1. Bilinearity: $e(aP, bQ) = e(P, Q)^{ab}$ for all $P, Q \in G_1, a, b \in \mathbb{Z}_q$.
2. Non-degenerative: There exists $P, Q \in G_1$ such that $e(P, Q) \neq 1$.
3. Computable: There is an efficient algorithm to compute $e(P, Q)$ for all $P, Q \in G_1$.

95.2.2 Computational Problems

Let G_1 and G_2 be two groups of prime order q , let $e : G_1 \times G_1 \rightarrow G_2$ be a bilinear pairing and let P be a generator of G_1 .

- Discrete Logarithm Problem (DLP) Given $P, Q \in G_1$, find $n \in \mathbb{Z}_q$ such that $P = nQ$ whenever such n exists.
- Computational Diffie–Hellman Problem (CDHP) Given $(P, aP, bP) \in G_1$ for $a, b \in \mathbb{Z}_q^*$, find the element abP .
- Bilinear Diffie–Hellman Problem (BDHP) Given $(P, aP, bP, cP) \in G_1$ for $a, b, c \in \mathbb{Z}_q^*$, compute $e(P, P)^{xyz} \in G_2$
- Bilinear Diffie–Hellman Problem (DBDHP) Given $(P, aP, bP, cP, \tau) \in G_1^4 \times G_2$ for $a, b, c \in \mathbb{Z}_q^*$, decide whether $\tau = e(P, P)^{abc}$

95.2.3 Identity Based Threshold Signcryption

A generic identity-based threshold signcryption scheme with total n players and t threshold limit consists of the following five algorithms:

Setup: Given a security parameter k , the private key generator (PKG) generates the system's public parameters. Among the parameters produced by Setup is a public key P_{pub} . There is also a corresponding master key s that is kept secret by PKG.

Extract: Given a user's identity ID , the PKG will compute a public key Q_{ID} , generate the private key S_{ID} and transmit the private key to its owner in a secure way.

Keydis: Given a private key S_{ID} associated with an identity ID that stands for a group of users, the number of signcryption members n and a threshold parameter t ,

this algorithm generates n shares of S_{ID} and provides each one to the signcryption members M_1, M_2, \dots, M_n . It also generates a set of verification keys that can be used to check the validity of each shared private key. We denote the shared private keys and the matching verification keys by $\{S_i\}_{i=1, \dots, n}$ and $\{y_i\}_{i=1, \dots, n}$, respectively. Note that each (S_i, y_i) is sent to M_i , then M_i publishes y_i but keeps S_i secret.

Signcrypt: Give a message m , the private keys of t members $\{S_i\}_{i=1, \dots, t}$ in a sender group U_A , the receiver's public key Q_{ID_B} , the Signcrypt algorithm outputs an identity-based (t, n) threshold signcryption σ on the message m .

Decrypt: Give a ciphertext σ , the private key of the receiver S_{ID_B} , the public key the sender group Q_{ID_A} , it outputs the plain text m or \perp if σ is an invalid ciphertext between the group U_A and the receiver.

95.3 Review of Selvi et al.'s Identity-Based Threshold Signcryption Scheme

The scheme involves four roles: the PKG, a trust dealer, a sender group $U_A = \{M_1, M_2, \dots, M_n\}$ with identity ID_A and a receiver Bob with identity ID_B .

Setup: Given a security parameter k , the PKG chooses groups G_1 and G_2 of prime order q (with G_1 additive and G_2 multiplicative), a generator P of G_1 , a bilinear map $e : G_1 \times G_1 \rightarrow G_2$, a secure symmetric cipher (E, D) and hash functions $H_1 : \{0, 1\}^* \rightarrow G_1, H_2 : G_2 \rightarrow \{0, 1\}^{m_1}, H_3 : \{0, 1\}^* \rightarrow Z_q^*$. The PKG chooses a master-key $s \in_R Z_q^*$ and computes $P_{pub} = sP$. The PKG publishes system parameters $\{G_1, G_2, n_1, e, P, P_{pub}, E, D, H_1, H_2, H_3\}$ and keeps the master-key s secret.

Extract: Given an identity ID , the PKG computes $Q_{ID} = H_1(ID)$ and the private key $S_{ID} = sQ_{ID}$. Then PKG sends the private key to its owner in a secure way.

Keydis: Suppose that a threshold t and n satisfy $1 \leq t \leq n < q$. To share the private key S_{ID_A} among the group U_A , the trusted dealer performs the steps below.

- 1) Choose F_1, \dots, F_{t-1} uniformly at random from G_1^* , construct a polynomial $F(x) = S_{ID_A} + xF_1 + \dots + x^{t-1}F_{t-1}$
- 2) Compute $S_i = F(i)$ for $i = 0, \dots, n$. ($S_0 = S_{ID_A}$). Send S_i to member M_i for $i = 1, \dots, n$ secretly.
- 3) Broadcast $y_0 = e(S_{ID_A}, P)$ and $y_j = e(F_j, P)$ for $j = 1, \dots, t - 1$.
- 4) Each M_i then checks whether his share S_i is valid by computing $e(S_i, P) = \prod_{j=0}^{t-1} y_j^{ij}$. If S_i is not valid, M_i broadcasts an error and requests a valid one.

Signcrypt: Let M_1, \dots, M_t are the t members who want to cooperate to signcrypt a message m on behalf of the group U_A .

- 1) Each M_i chooses $x_i \in_R Z_q^*$, computes $R_{1i} = x_iP, R_{2i} = x_iP_{pub}, \tau_i = e(R_{2i}, Q_{ID_B})$ and sends (R_{1i}, τ) to the clerk C .

- 2) The clerk C (one among the t cooperating players) computes $R_1 = \prod_{i=1}^t R_{1i}$, $\tau = \prod_{i=1}^t \tau_i$, $k = H_2(\tau)$, $c = E_k(m)$, and $h = H_3(m, R_1, k)$.
- 3) Then the clerk C sends h to M_i for $i = 0, \dots, t$
- 4) Each M_i computes the partial signature $W_i = x_i P_{pub} + h \eta_i S_i$ and sends it to the clerk C, where $\eta = \prod_{j=1, j \neq i}^t -j(i-j)^{-1} \pmod{q}$.
- 5) Clerk C verifies the correctness of partial signatures by checking if the following equation holds:

$$e(P, W_i) = e(R_{1i}, P_{pub}) \left(\prod_{j=0}^{t-1} y_j^{h^j} \right)^{h m_i}$$

If all partial signatures are verified to be legal, the clerk C computes $W = \sum_{i=1}^t W_i$; otherwise rejects it and requests a valid one.

- 6) The final threshold signcryption is $\sigma = (c, R_1, W)$.

Unsigncrypt: When receiving σ , Bob follows the steps below.

- 1) Compute $\tau = e(R_1, S_{ID_B})$ and $k = H_2(\tau)$.
- 2) Recover $m = D_k(c)$
- 3) Compute $h = H_3(m, R_1, k)$ and accept σ if and only if the following equation holds:

$$e(P, W) = e(P_{pub}, R_1 + h Q_{ID_A})$$

95.4 Cryptanalysis of Selvi et al.'s Scheme

The two schemes are both insecure from the view of attack by a malicious attacker who can control the communication channel.

The attacker intercepts the ciphertext $\sigma = (c, R_1, W)$ from sender.

- 1) Randomly choose $\alpha, x \in Z_q^*$ and prepare a forged message m'
- 2) Compute $R'_1 = xP, R'_2 = xP_{pub}, \tau' = e(R'_2, Q_{ID_B}), k' = H_2(\tau), c' = E_{k'}(m'), h' = H_3(m', R'_1, k')$.
- 3) Compute $W' = \alpha P_{pub}$, set $Q'_A = (\alpha - x)P/h'$ as a public key of U_A
- 4) The final ciphertext is $\sigma' = (c', R'_1, W')$.
- 5) Attacker sends the forged ciphertext and the replaced public key to the receiver.

After receiving the ciphertext $\sigma' = (c', R'_1, W')$, the receiver

- 1) Compute $\tau = e(R'_1, S_{ID_B}) = e(R'_2, Q_{ID_B}) = \tau', k = H_2(\tau) = H_2(\tau') = k'$
- 2) Recover $m = D_k(c') = D_{k'}(c') = m', h = H_3(m', R'_1, k') = h'$.

3) Verify $e(P, W') \stackrel{?}{=} e(P_{pub}, R'_1 + hQ'_{ID_A})$

$$\because e(P_{pub}, R'_1 + hQ'_{ID_A}) = e(P_{pub}, xP + h \cdot (\alpha - x)P/h') = e(P_{pub}, \alpha P) = e(P, W')$$

\therefore The equation $e(P, W') = e(P_{pub}, R'_1 + hQ'_{ID_A})$ set.

Discussion:

In the view of the attacker, [27] can be simulated as following basic Sign-cryption scheme:

A sender “Alice” with key pairs $\{Q_A = H_1(Alice), S_A = sH_1(Alice)\}$

A receiver “Bob” with key pairs $\{Q_{Bob} = H_1(Bob), S_B = sH_1(Bob)\}$

Alice chooses $x \in Z_q^*, R_1 = xP, R_2 = xP_{pub}, \tau = e(R_2, Q_B), k = H_2(\tau), c = E_k(m), h = H_3(m, R_1, k), W = xP_{pub} + hS_A$ and sends $\sigma = (c, R_1, W)$ to Bob as the ciphertext of his message.

There is a small mistake of the definition $H_3 : \{0, 1\}^* \rightarrow Z_q^*$. We think the authors' real intention is $H_3 : \{0, 1\}^* \times G_1 \times \{0, 1\}^* \rightarrow Z_q^*$ to meet $h = H_3(m, R_1, k)$. In this hash function, any message about the sender is not contained. If an attacker Eve say “I am Alice” to Bob, Bob can not distinguish only with the hash value h. Our attack just utilizes this attribute of Li's scheme.

Suppose that H_3 is defined as $H_3 : \{0, 1\}^* \times G_1 \times \{0, 1\}^* \times G_1 \rightarrow Z_q^*$, and $h = H_3(m, R_1, k, Q_{ID_A})$. The attacker Eve intercepts the ciphertext $\sigma = (c, R_1, W)$ from sender Alice and she runs the algorithm of forging ciphertext like:

- 1) Randomly choose $\alpha, x \in Z_q^*$ and prepare a forged message m'
- 2) Compute $R'_1 = xP, R'_2 = xP_{pub}, \tau' = e(R'_2, Q_{ID_B}), k' = H_2(\tau), c' = E_{k'}(m'), h' = H_3(m', R'_1, k', Q'_A)$.
- 3) Compute $W' = \alpha P_{pub}$, set $Q'_A = (\alpha - x)P/h'$ as a public key of U_A
- 4) The final ciphertext is $\sigma' = (c', R'_1, W')$.
- 5) Send the forged ciphertext and the replaced public key to the receiver.

She will meet a hard problem that if she wants to compute h', Q'_A is necessary or if she wants to compute Q'_A, h' must be known. As a result, if she can succeed in forging the ciphertext, she must own the ability to solve the DL problem.

95.5 Conclusion

In this chapter, we show that the threshold sign-cryption scheme of Selvi et al. is vulnerable if the attacker can replace the group public key. Then we point out that the receiver uses the senders' public key without any verification in the sign-encrypt stage cause this attack.

References

1. Zheng Y (1997) Digital signcryption or how to achieve cost (signature and Encryption) \ll cost (signature) + cost (encryption), In: Proceedings of advances in CRYPTO'97, LNCS 1294. Springer, Berlin, pp 165–179
2. Bao F, Deng RH (1997) A signcryption scheme with signature directly verifiable by public key. In: PKC'98, LNCS, vol 1431. Springer, Berlin, pp 55–59
3. Chow SSM, Yiu SM, Hui LCK, Chow KP (2004) Efficient forward and provably secure ID-based signcryption scheme with public verifiability and public ciphertext authenticity. In: ICISC'03, LNCS, vol 2971. Springer, Berlin, pp 352–269
4. Boyen X (2003) Multipurpose identity based signcryption: a swiss army knife for identity based cryptography. In: CRYPT'03, LNCS, vol 2729. Springer, Berlin, pp 383–399
5. Mu Y, Varadharajan V (2000) Distributed signcryption. In: INDOCRYPT'00, LNCS, vol 1977. Springer, Berlin, pp 155–164
6. Yang G, Wong DS, Deng X (2005) Analysis and improvement of a signcryption scheme with key privacy. In: ISC'05, LNCS, vol 3650. Springer, Berlin, pp 218–232
7. SteinFeld R, Zheng Y (2000) A signcryption scheme based on integer factorization. In: ISW'00, LNCS, vol 1975. Springer, Berlin, pp 308–322
8. Libert B, Quisquater J (2004) Efficient signcryption with key prevacy from gap Diffie–Hellman groups. In: PKC'04, LNCS, vol 2947. Springer, Berlin, pp 187–200
9. Baek J, Steinfeld R, Zheng Y (2002) Formal proofs for the security of signcryption. In: PKC'02, LNCS, vol 2274. Springer, Berlin, pp 80–98
10. Shamir A (1984) Identity-based cryptosystems and signature schemes. In: CRYPTO'84, LNCS, vol 196. Springer, Berlin, pp 47–53
11. Boneh D, Franklin M (2001) Identity-based encryption from well pairing. In: CRYPTO'01, LNCS vol 2139. Springer, Berlin, pp 213–229
12. Barreto PSLM, Libert B, McCullagh N, Quisquater JJ (2005) Efficient and provably-secure identity-based signatures and signcryption from bilinear maps. In: ASIACRYPT'05, LNCS vol 3788. Springer, Berlin, pp 515–532
13. Fagen L, Hu X, Xuyun N (2009) A new multi-receiver ID-based signcryption scheme for group communications. In: ICCAS'2009. Springer, Berlin, pp 296–300
14. Yiliang H, Xiaolin G (2009) Multi-recipient signcryption for secure group communication. In: ICIEA. Springer, Berlin, pp 161–165
15. Jin Zhengping, Wen Qiaoyan, Hongzhen Du (2010) An improved semantically-secure identity-based signcryption scheme in the standard model. *Comput Electr Eng* 36(3):545–552
16. Huang X, Susilo W, Mu Y, Zhang E (2005) Identity-based ring signcryption schemes: cryptographic primitives for preserving privacy and authenticity in the ubiquitous world. In: 19th international conference on advanced information networking and applications, Taiwan, pp 649–654
17. Liu Z, Hu Y, Zhang X, Ma H (2010) Certificateless signcryption scheme in the standard model. *Inf Sci* 180(3):452–464
18. Yu Y, Yang B, Sun Y, Zhu S (2009) Identity based signcryption scheme without random oracles. *Comput Stand Interfaces* 31(1):56–62
19. Desmedt Y (1987) Society and group oriented cryptography: a new concept. In: CRYPTO'87, LNCS vol 293. Springer, Berlin, pp 120–127
20. Desmedt Y, Frankel Y (1991) Shared generation of authenticators and signatures. In: CRYPTO'91, LNCS, vol 576. Springer, Berlin, pp 457–469
21. Rivest RL, Shamir A, Adleman L (1978) A method for obtaining digital signatures and public-key cryptosystems. *Commun ACM* 21(2):120–126
22. Baek J, Zheng Y (2004) Identity-based threshold signature scheme from the bilinear pairings. In: International conference on information technology, Las Vegas, pp 124–128
23. Duan S, Cao Z, Lu R (2004) Robust ID-based threshold signcryption scheme from pairings. In: International conference on information security, Shanghai, pp 33–37

24. Peng C, Li X (2005) An identity-based threshold signcryption scheme with semantic security. *Computational intelligence and security, LNAI, vol 3902*. Springer, Berlin, pp 173–179
25. Libert B, Quisquater JJ (2003) A new identity based signcryption schemes from pairings. *IEEE information theory workshop, Paris*, pp 155–158
26. Fagen L, Yong Y (2008) An efficient and provably secure ID-based threshold signcryption scheme. In: *ICCCAS*, pp 488–492
27. Selvi SSD, Vivek SS, Rangan CP (2008) Cryptanalysis of Li et al.'s identity-based threshold signcryption scheme. *Embedded and Ubiquitous Computing*, pp 127–132

Chapter 96

Analysis of an Authenticated 3-Round Identity-Based Group Key Agreement Protocol

Wei Yuan, Liang Hu, Hongtu Li and Jianfeng Chu

Abstract In 2008, Gang Yao et al. proposed an authenticated 3-round identity-based group key agreement protocol, which is based on Burmester and Desmedt's protocol proposed at Eurocrypt 94. However, their protocol can only prevent passive attack. If the active attack is allowed, the protocol is vulnerable and an internal attacker can forge her neighbor's keying material. Thus their protocol does not achieve the aim of authentication.

Keywords Authentication · Identity-based · Key agreement · Bilinear pairing · Cryptanalysis · Attack

96.1 Introduction

Secure and reliable communications [1] have become critical in modern society. The centralized services such as file sharing, can be changed into distributed or collaborated system based on multiple systems and networks. Basic cryptographic

W. Yuan · L. Hu · H. Li · J. Chu (✉)

Department of Computer Science and Technology, Jilin University, Changchun, China
e-mail: chujf@jlu.edu.cn

W. Yuan
e-mail: yuanwei1@126.com

L. Hu
e-mail: hul@mails.jlu.edu.cn

H. Li
e-mail: li_hongtu@hotmail.com

functions such as data confidentiality, data integrity, and identity authentication are required to construct these secure systems.

Key agreement protocol [2–4] allows two or more participants, each of whom has a long-term key respectively, to exchange information over a public communication channel with each other. However, the participants can not ensure others' identity. Though Alice wants to consult a session key with Bob, Alice can not distinguish it if Eve pretends that she is Bob. The authenticated key agreement protocol overcomes this flaw and makes unfamiliar participants to ensure others' identities and consult a common session key in the public channel.

Shamir [5] introduced an identity-based public key cryptosystem in 1984, in which a user's public key can be calculated from his identity and defined hash function, while the user's private key can be calculated by a trusted party called Private Key Generator (PKG). The identity-based public key cryptosystem simplifies the program of key management and increases the efficiency. In 2001, Boneh and Franklin [6] found bilinear pairings positive applications in cryptography and proposed the first practical identity-based encryption protocol with bilinear pairings. Soon, the bilinear pairings become important tools in constructing identity-based protocols and a number of identity-based encryption or signature schemes [7–12] and authenticated key agreement protocols [13–17] were proposed.

In 2008, Gang Yao, Hongji Wang, and Qingshan Jiang [18] proposed an authenticated 3-round identity-based group key agreement protocol. The first round is for identity authentication, the second round is for key agreement, and the third round is for key confirmation. Their protocol is based on the protocol of Burmester and Desmedt [19] which was proposed at Eurocrypt 94. They declared the proposed protocol provably-secure in the random oracle model.

In this chapter, we show that an authenticated 3-round identity-based group key agreement protocol proposed by Gang Yao et al. is vulnerable: an internal attacker can forge her neighbors' keying material. Then we propose an improved provably-secure protocol based on Burmester and Desmedt's as well. At last, we summarize several security attributes of our improved authenticated identity-based group key agreement protocol.

96.2 Paper Preparation

96.2.1 Bilinear Pairing

Let P denote a generator of G_1 , where G_1 is an additive group of large order q and let G_2 be a multiplicative group with $|G_1| = |G_2|$. A bilinear pairing is a map $e : G_1 \times G_1 \rightarrow G_2$ which has the following properties:

1. Bilinearity: Given $Q, W, Z \in G_1$, $e(Q, W + Z) = e(Q, W) \cdot e(Q, Z)$ and $e(Q + W, Z) = e(Q, Z) \cdot e(W, Z)$. There for any $q, b \in Z_q$: $e(aQ, bW) = e(Q, W)^{ab} = e(abQ, W) = e(Q, abW) = e(bQ, W)^a$.
2. Non-degenerative: $e(P, P) \neq 1$, where 1 is the identity element of G_2 .
3. Computable: If $Q, W \in G_1$, one can compute $e(Q, W) \in G_2$ in polynomial time efficiently.

96.2.2 Computational Problems

Let G_1 and G_2 be two groups of prime order q , let $e : G_1 \times G_1 \rightarrow G_2$ be a bilinear pairing and let P be a generator of G_1 .

Discrete Logarithm Problem (DLP)

Given $P, Q \in G_1$, find $n \in Z_q$ such that $P = nQ$ whenever such n exists.

Computational Diffie–Hellman Problem (CDHP)

Given $(P, aP, bP) \in G_1$ for $a, b \in Z_q^*$, find the element abP

Bilinear Diffie–Hellman Problem (BDHP)

Given $(P, xP, yP, zP) \in G_1$ for $x, y, z \in Z_q^*$, compute $e(P, P)^{xyz} \in G_2$.

96.2.3 Introduction of BR Security Model

To describe the security model for entity authentication and key agreement aims, Bellare and Rogaway proposed the BR93 model [13] for two-party authenticated key agreement protocol in 1993 and the BR95 model [14] for three-party authenticated key agreement protocol in 1995. In BR model, the adversary can control the communication channel and interact with a set of Π_{U_x, U_y}^i oracles, which specify the behavior between the honest players U_x and U_y in their i th instantiation. The predefined oracle queries are described informally as follows:

Send (U_x, U_y, i, m) : The adversary sends message m to the oracle Π_{U_x, U_y}^i . The oracle Π_{U_x, U_y}^i will return the session key if the conversation has been accepted by U_x and U_y or terminate and tell the adversary.

Reveal (U_x, U_y, i) : It allows the adversary to expose an old session key that has been previously accepted. After receiving this query, Π_{U_x, U_y}^i will send this session key to the adversary, if it has accepted and holds some session key.

Corrupt (U_x, K) : The adversary corrupts U_x and learns all the internal state of U_x . The corrupt query also allows the adversary to overwrite the long-term key of corrupted principal with any other value K .

Test (U_x, U_y, i): It is the only oracle query that does not correspond to any of the adversary's abilities. If Π_{U_x, U_y}^i has accepted with some session key and is being asked a **Test** (U_x, U_y, i) query, then depending on a randomly chosen bit b , the adversary is given either the actual session key or a session key drawn randomly from the session key distribution.

Freshness: The notion is used to identify the session keys about which adversary should not know anything because she has not revealed any oracles that have accepted the key and has not corrupted any principals knowing the key. Oracle $\Pi_{A,B}^i$ is fresh at the end of execution, if, and only if, oracle $\Pi_{A,B}^i$ has accepted with or without a partner oracle $\Pi_{B,A}^i$, both oracle $\Pi_{A,B}^i$ and its partner oracle $\Pi_{B,A}^i$ have not been sent a **Reveal** query, and the principals A and B of oracles $\Pi_{A,B}^i$ and $\Pi_{B,A}^i$ (if such a partner exists) have not been sent a **Corrupt** query.

Security is defined using the game G, played between a malicious adversary and a collection of Π_{U_x, U_y}^i oracles and instances. The adversary runs the game simulation G, whose setting is as follows.

Phase 1: Adversary is able to send any **Send**, **Reveal**, and **Corrupt** oracle queries at will in the game simulation G.

Phase 2: At some point during G, adversary will choose a fresh session on which to be tested and send a **Test** query to the fresh oracle associated with the test session. Note that the test session chosen must be fresh. Depending on a randomly chosen bit b , adversary is given either the actual session key or a session key drawn randomly from the session key distribution.

Phase 3: Adversary continues making any **Send**, **Reveal**, and **Corrupt** oracle queries of its choice.

Finally, adversary terminates the game simulation and outputs a bit b' , which is its guess of the value of b . Success of adversary in G is measured in terms of adversary's advantage in distinguishing whether adversary receives the real key or a random value. A wins if, after asking a **Test** (U_x, U_y, i) query, where Π_{U_x, U_y}^i is fresh and has accepted, adversary's guess bit b' equals the bit b selected during the **Test** (U_x, U_y, i) query.

A protocol is secure in the BR model if both the validity and indistinguishability requirements are satisfied:

Validity. When the protocol is run between two oracles in the absence of a malicious adversary, the two oracles accept the same key.

Indistinguishability. For all probabilistic, polynomial-time (PPT) adversaries A, $\text{Adv}_A(k)$ is negligible.

96.3 Review of Gang Yao et al.'s Protocol

Let U_1, \dots, U_n be n participants, and PKG be the private key generator. Let ID_i be the identity of U_i . Suppose that G_1 and G_2 are two cyclic groups of order q for

some large prime q . G_1 is a cyclic additive group and G_2 is a cyclic multiplicative group. Let P be an arbitrary generator of G_1 , and $e : G_1 \times G_1 \rightarrow G_2$ be a bilinear pairing.

In Gang Yao et al.'s protocol, the following two steps prepare the system parameters:

Setup:

The PKG chooses a random number $s \in \mathbb{Z}_q^*$ and set $R = sP$. The PKG also chooses $H_0 : \{0, 1\}^* \rightarrow G_1^*$ to be a Map-to-Point hash function, and H is a cryptographic hash function. Then the PKG publishes system parameters $\{q, G_1, G_2, e, P, R, H_0, H\}$, and keeps s as its master key.

Extract:

Given a public identity $ID \in \{0, 1\}^*$, the PKG computes the public key $Q = H_0(ID) \in G_1$ and generates the associated private key $S = sQ$. The PKG passed S as the private key to the user via some secure channel.

Let n users U_1, \dots, U_n with respective public keys $Q_i = H_0(ID_i) (1 \leq i \leq n)$ decide to agree upon a common secret key. $S_i = sQ_i$ is the long term secret key of U_i sent by the PKG on submitting U_i 's public identity $(1 \leq i \leq n)$. Let U denote $U_1 || \dots || U_n$.

We assume that U_1 is the protocol initiator. The protocol may be performed in three rounds as follows:

Round 1: Identity Authentication

Every participant U_i generates a random number $r_i \in \mathbb{Z}_q^*$, computes

$$E_i = r_i P, \quad F_i = H(U, e(E_i, R)) S_i + r_i R$$

And broadcasts E_i and F_i .

After receiving every E_j and $F_j (1 \leq j \leq n, j \neq i)$, U_i verifies that none of them equals 1. If the check succeeds, U_i verifies whether

$$e\left(\sum_{j \neq i} F_j, P\right) = e\left(\sum_{j \neq i} H(U, e(E_j, R)) Q_j + E_j, R\right)$$

holds or not. If the verification succeeds, U_i continues with the next round. Otherwise, the protocol execution is terminated and a notification of failure will be broadcasted.

Round 2: Key Agreement

U_i computes $T = H_0(ID_1 || E_1 || \dots || ID_n || E_n)$, then he computes Y_i, X_i as follows.

$$Y_i = r_i T, \quad X_i = r_i (E_{i+1} - E_{i-1} + T),$$

And broadcasts X_i and Y_i .

After receiving every X_i and $Y_i (1 \leq j \leq n, j \neq i)$, U_i verifies whether

$$e\left(\sum_{j \neq i} Y_j, P\right) = e\left(\sum_{j \neq i} E_j, T\right)$$

holds or not. IF the verification succeeds, U_i continues with the next round. Otherwise, the protocol execution is terminated and a notification of failure will be broadcasted.

Round 3: Key Confirmation

U_i computes the keying material Z_i as

$$Z_i = e\left(nr_i E_{i-1} + \sum_{j=1}^{n-1} (n-1-j)(X_{i+j} - Y_{i+j}), R\right),$$

then he computes

$$C_i = H(i \| U \| E_1 \| \cdots \| E_n \| X_1 \| \cdots \| X_n \| Y_1 \| \cdots \| Y_n \| Z_i)$$

and broadcasts C_i .

After receiving every $C_j (1 \leq j \leq n, j \neq i)$, U_i computes the session key as

$$K_i = H(U \| E_1 \| \cdots \| E_n \| X_1 \| \cdots \| X_n \| Y_1 \| \cdots \| Y_n \| Z_i \| C_1 \| \cdots \| C_n).$$

Otherwise, U_i terminates the protocol execution and a notification of failure will be broadcasted.

96.4 Cryptanalysis of Gang Yao et al.'s protocol

In our view, two important principles should be attention: Before we want to protect a message, we should know whether it really needs to be protected. After finishing our protocol, we should ensure all the valuable messages have been protected. In Gang Yao et al.'s protocol, to derive the session key, $K_i = H(U \| E_1 \| \cdots \| E_n \| X_1 \| \cdots \| X_n \| Y_1 \| \cdots \| Y_n \| Z_i \| C_1 \| \cdots \| C_n)$, all the parameters except Z_i can be gained from the broadcast messages. It is important to ensure the transmitted messages are not modified, forged, or deleted by attackers. In the round 1, $E_i (1 \leq i \leq n)$ should be protected and in the round 2, X_i and Y_i should be protected. However, only E_i and Y_i here are protected in Gang Yao et al.'s protocol. Because the equation $e\left(\sum_{j \neq i} Y_j, P\right) = e\left(\sum_{j \neq i} E_j, T\right)$, where $T = H_0(ID_1 \| E_1 \| \cdots \| ID_n \| E_n)$, user U_i can ensure that Y_j is not modified or forged but X_i is transmitted without any verification. Actually, both X_{i+j} and Y_{i+j} are needed in the equation $Z_i = e\left(nr_i E_{i-1} + \sum_{j=1}^{n-1} (n-1-j)(X_{i+j} - Y_{i+j}), R\right)$ in the round 3. Thus, X_i can be replaced by $X_i = Y_i$. Then Z_i can be expressed as $Z_i = e(nr_i E_{i-1}, R)$. Due to the characteristic of bilinear pairing: $Z_i = e(nr_i E_{i-1}, R) =$

$e(nr_i r_{i-1} P, sP) = e(r_{i-1} sP, r_i P)^n = e(r_{i-1} R, E_i)^n$ That is, with the random number r_i , any user U_i can generate U_{i+1} 's keying material Z_{i+1} .

As a result, an attacker who can control the communication channel has the ability to intercept and forge all the X_i 's. If a malicious user U_e wants to forge U_{e+1} 's keying material C_{e+1} , she can compute $Z_{e+1} = e(r_e R, E_{e+1})^n$ and $C'_{e+1} = H(e + 1 || U || E_1 || \dots || E_n || X_1 || \dots || X_n || Y_1 || \dots || Y_n || Z_{e+1})$. Finally, she can broadcast C'_{e+1} to replace C_{e+1} .

96.5 Conclusions

In this chapter, we show that an authenticated 3-round identity-based group key agreement protocol proposed by Gang Yao et al. is vulnerable: an internal attacker can forge her neighbors' keying material. Then we propose an improved provably-secure protocol based on Burmester and Desmedt's protocol as well. At last, we summarize some security attributes of our improved authenticated identity-based group key agreement protocol.

Acknowledgments The authors would like to thank the editors and anonymous reviewers for their valuable comments. This work is supported by the National Natural Science Foundation of China under Grant No. 60873235 and 60473099, the National Grand Fundamental Research 973 Program of China (Grant No. 2009CB320706), Scientific and Technological Developing Scheme of Jilin Province (20080318), and Program of New Century Excellent Talents in University (NCET-06-0300).

References

1. Kulkarni SS, Bruhadeshwar B (2010) Key-update distribution in secure group communication. *Comput Commun* 33:689–705
2. Diffie W, Hellman ME (1976) New directions in cryptography. *IEEE Trans Inf Theory* IT-22:644–654
3. Diffie W (1988) The first ten years of public-key cryptography. *Proc IEEE* 76(5):560–577
4. Zhao J, Gu D, Li Y (2010) An efficient fault-tolerant group key agreement protocol. *Comput Commun* 33:890–895
5. Shamir A (1984) Identity-based cryptosystems signature schemes. *Advances in cryptology, CRYPTO'84, LNCS 196*, Springer, Berlin, pp 47–53
6. Boneh D, Franklin M (2001) Identity-based encryption from the weil pairing, *advances in cryptology, CRYPTO'2001, LNCS 2139*, Springer, Berlin, pp 213–229
7. Chin J-J, Heng S-H, Goi B-M (2008) An efficient and provable secure identity-based identification scheme in the standard model, *LNCS 5057*, Springer, Berlin, pp 60–73
8. Liu Z, Hu Y, Zhang X, Ma H (2010) Certificateless signcryption scheme in the standard model. *Inf Sci* 180:452–464
9. Zhang J, Yang Y, Niu X, Gao S, Chen H, Geng Q (2009) An improved secure identity-based on-line/off-line signature scheme, *ISA 2009, LNCS 5576*, Springer, Berlin, pp 588–597

10. Chang TY (2009) An ID-based group-oriented decryption scheme secure against adaptive chosen-ciphertext attacks. *Comput Commun* 32:1829–1836
11. Kiayias A, Zhou H-S (2007) Hidden identity-based signatures, LNCS 4886, Springer, Berlin, pp 134–147
12. Li C-T (2010) On the security enhancement of an efficient and secure event signature protocol for P2P MMOGs, ICCSA 2010, LNCS 6016, pp 599–609
13. Lu R, Cao Z (2005) A new deniable authentication protocol from bilinear pairings. *Appl Math Comput* 168:954–961
14. Lu R, Cao Z, Wang S, Bao H (2007) A new ID-based deniable authentication protocol. *Informatics* 18(1):67–78
15. Cao T, Lin D, Xue R (2005) An efficient ID-based deniable authentication protocol from pairings, AINA'05, 388–391
16. Chou JS, Chen YL, Huang JC, (2006) An ID-based deniable authentication protocol on pairings. *Cryptol ePrint Arch*: Rep 335
17. Hwang JY, Choi KY, Lee DH (2008) Security weakness in an authenticated group key agreement protocol in two rounds. *Comput Commun* 31:3719–3724
18. Yao G, Wang H, Jiang Q (2008) An authenticated 3-round identity-based group key agreement protocol. *The third international conference on availability, reliability, and security*, ACM, pp 538–543
19. Burmester M, Desmedt Y (1994) A secure and efficient conference key distribution system, EUROCRYPT'94, LNCS 950, Springer, Berlin, pp 275–286

Chapter 97

Research on ECC Digital Certificate in ATN

Wu Zhijun, Lang Jihai and Huang Jun

Abstract Aeronautic Telecommunication Network (ATN) is the infrastructure network for next generation air traffic management system. This Chapter discusses the application of digital certificate in ATN air-ground subnet and the structure of X.509 digital certificate, which are based on ECC, and are generated by OpenSSL. Test results show that ECC compression certificate can enhance the security of ATN air-ground data link by the premises of saving bandwidth and guarantee of the computing performance.

Keywords ATN · ECC · X.509 · Compressed certificate

97.1 Introduction

ATN air-ground data link is a kind of wireless network which is more vulnerable to be attacked compared with the ground-ground wired network.

The main threats to the ATN air-ground network are: data leakage, physical disguise, information tampering and Denial of Service (DoS) attack. This chapter responds the International Civil Aviation Organization (ICAO) suggestion [1] by

W. Zhijun (✉) · L. Jihai · H. Jun
School of Electronics and Information Engineering, Civil Aviation University of China,
300300 Tianjin, China
e-mail: zjwu@cauc.edu.cn

L. Jihai
e-mail: lanjinglinghuang@163.com

H. Jun
e-mail: huangjun_cauc@163.com

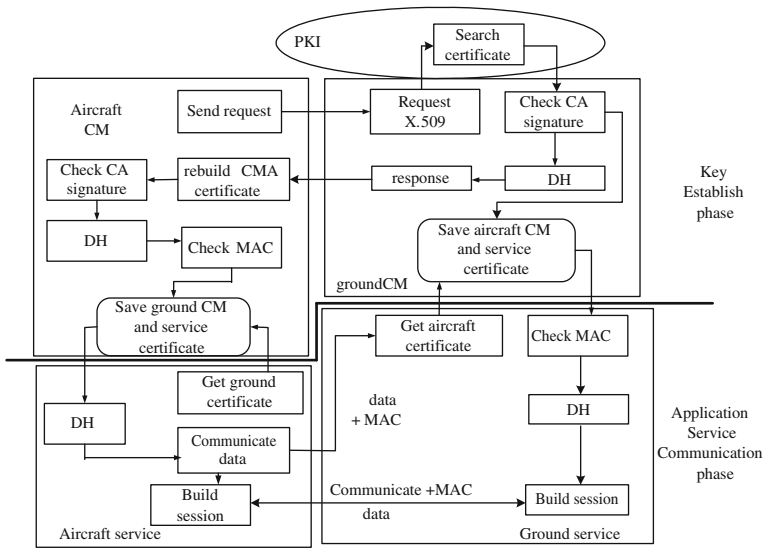


Fig. 97.1 ATN application service for security communication

adapting ECC-based asymmetric and symmetric encryption algorithm and digital signature to protect the security of ATN data link.

97.2 The Application of Digital certificates in the ATN Air-Ground Communication

By using Public Key Infrastructure (PKI) on the application layer, we can provide a secure communication mechanism on ATN [2]. Controller pilot data link communications (CPDLC) is a typical ATN application service. As shown in Fig. 97.1, CPDLC is taken as an example of the digital certificates application in the ATN air-ground system to secure communication. ATN terminal communications are divided into two phases which are key establishment and application services [2].

97.2.1 Key Establishment Phase

Aircraft context management (CM) sends a login request message to the ground CM, it will obtain the CM’s public key certificate and the list of applications which is supported from the Certificate Authority (CA)’s certificate directory. The ground CM will produce a shared secret parameters Z , and then use of Diffie-Hellman (DH) key agreement mechanism to generate a temporary shared key KCM. After

that, the ground CM sends a landing response message, which contains the certificate and the compression of the ground CM's application support services, and replies the Message Authentication Code (MAC) treatment by utilizing KCM. After receiving the response, the aircraft will re-compose of the compression standard certificate and decode into X.509 certificates, verifying that the CA signature and public key certificate can ensure the authenticity of the ground CM. Then the ground CM's shared key and the aircraft's key are combined to co-produced parameter Z and the KCM.

97.2.2 Application Service Communication Phase

After the first phase completed, aircraft Air Traffic Services (ATS) application services need to communicate with the ground specified application services. Assuming the ground as the initiator, who checks the parameter Z and aircraft's public key certificate by using the ground CM, and then click the key application services with the ground through the DH algorithm to generate the temporary shared key KAP1, which generates communication information with the MAC and the information, should be sent out together. When aircraft receives communication message sent by the application information of ground, it will obtain corresponding public key certificate from the ground aircraft CM, while generate the same KAP1, verifying the communication of information MAC. If it is valid, the two sides have established communication.

ATS of aircraft application service uses the same aircraft CM's shared key pair, but the temporary shared key generated by ground application varies from the public key, such as the KAP1, KAP2. So the aircraft CM login information is issued only once, but more than one secure session can be created between application service and the ground.

97.3 A ATN-ECC Structure Digital Certificate

The certificate format used in the ATN's public key infrastructure is based on X.509 certificate format [1]. X.509 is the most widely used certificate form. Figure 97.2 shows the X.509 public certificate structure. X.509v3 no-compressed format uses distinguished encoding rules (DER) rules of ISO/IEC 8825-1 identification for its encoding [3].

However, the size of a full certificate is thousands of bytes, the limited bandwidth resources is a heavy burden for the ATN air-ground network, so the certificate compression is a good strategy. The principle is that, amongst the communication entities, some of the same certificate in the default field values can be pre-stored in the communication entities, such as the certificate version number v3, then omitted from the certificate in the compression of these domains;

Fig. 97.2 X.509 public certificate structure

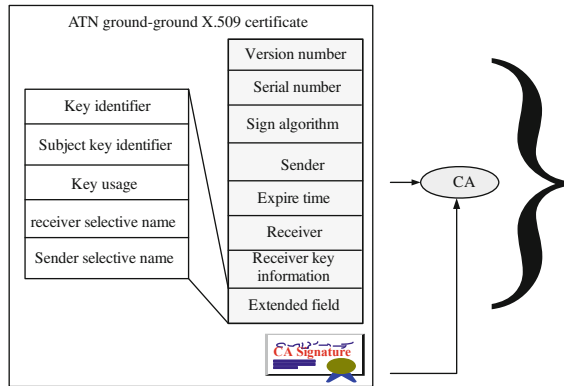
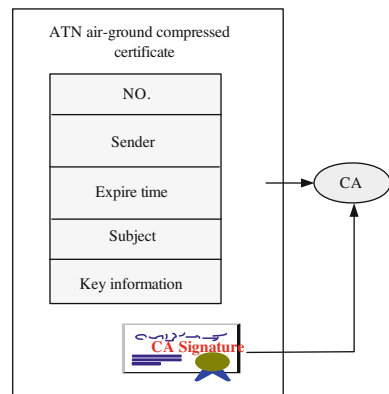


Fig. 97.3 Compressed certificates structure



when entity received the compressed certificate, then use pre-stored information to restore the compressed certificate, the omitted domain will be re-assigned and create standard X.509v3 certificate. Compressed certificate is shown in Fig. 97.3.

ATN compressed certificate can be used as ISO/IEC 8825-2 compressed encoding rules specified in the packed encoding rules (PER) encoding rules to code. After compressing, the size of certificate is much smaller than standard certificate.

97.4 Certificate Requests, Issue and Compression

97.4.1 Certificate Request

An X.509v3 certificate request mainly includes request information and a digital signature of request information. OpenSSL provides relevant programming interface library, using the OpenSSL API to construct an ATN-ECC certificate request includes the following steps:

- (1) Select elliptic curve generated using the specified key parameter file and generate a key;
- (2) In the construct of the main configuration file, the “req” information field which contains the subject name, subject public key and a set of optional extended attributes;
- (3) Use the subject’s private key to sign the certificate request information, select the SHA-1 as digest algorithm.

97.4.2 Certificate Issue

The policy field in the configuration file can be set for different user groups, such as CAs, ATN ATS appentities, ATN Intermediate System by using different matching strategies. Modify the extended field “v3_ca” and “usr_cert” which contains key purposes of authority, subject alternative name and the issuer.

According to the certificate request and object of certificate requester, the version number, serial number of the certificate, the certificate start and end time, the certificate subject name, public key information from the certificate, the certificate issuer information, extended projects and signature on the certificate can be set. Finally encoded and stored by using DER [4]. Certificate request and issue process is shown in Fig. 97.4.

97.4.3 Compression Certificate

Because ATN-ECC certificates are based on the X.509v3, the CA is fixed and known, and the default signature algorithm is ecdsa-with-SHA1 algorithm, similar fixed-field values can be omitted from the certificate as to achieve the purpose of reducing the certificate volume. ATN-ECC compression certificate agreed on that some of the certificate in the default field values between the aircraft and ground can pre-store that field in the media. After aircraft receiving the compressed certificate, the aircraft uses pre-stored information to restore the compressed certificate, and then standard X.509 certificates can be generated. The process of transfer between compressed and standard certificate is shown in Fig. 97.5.

ATN-ECC certificate compressing process omitted the certificate number, signature algorithm, issuer, subject information, institutions and subject key identifier; while serial number, expiration date, subject public key information, key algorithm, the optional name and the name of the issuer were preserved.

The revert of compressed certificate can use the serial number, which is provided by the compressed certificate, to find the corresponding issuer, subject information, key identifier key institutions in the local storage media and issue a list of the main body of the information list. Finally, according to the compression

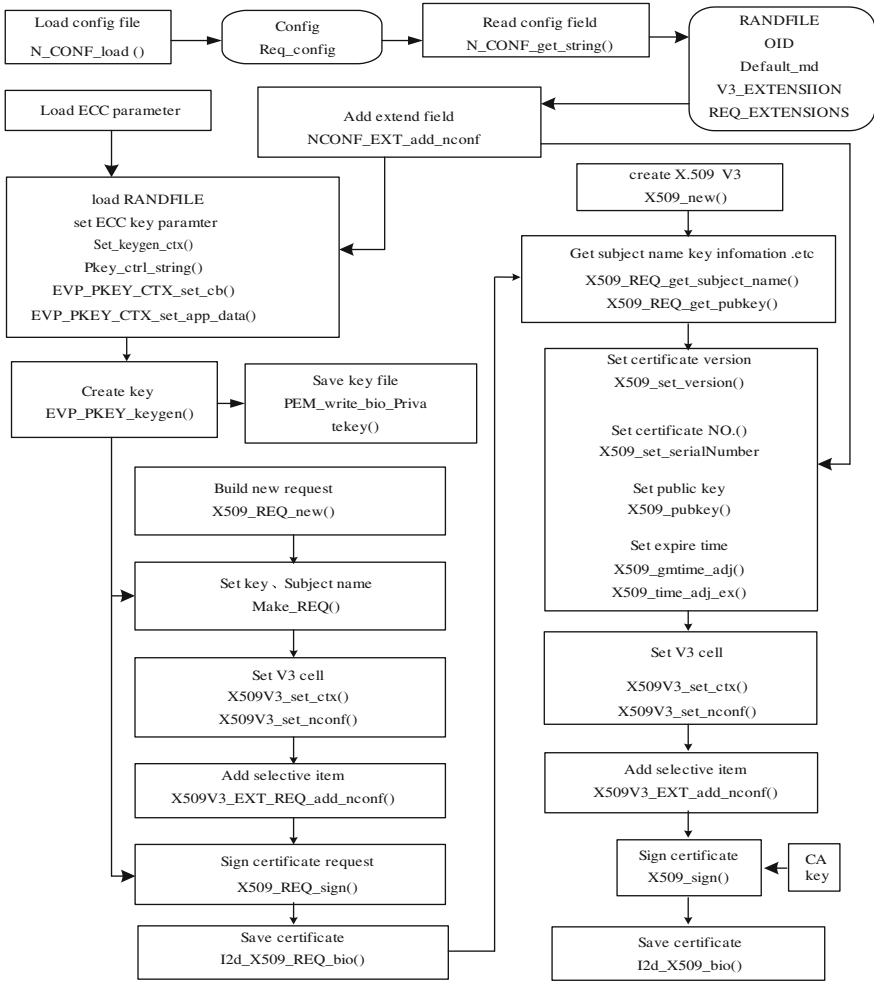


Fig. 97.4 Process of certificate request and issue

key identifier of main export, public key re-generates the standard ANT-ECC X.509v3 certificate.

97.5 Performance Analysis

97.5.1 Size of the Key and Certificate

The security of ECC algorithm depends on the group defined by the elliptic curve discrete logarithm problem [5, 6]. The algorithm requires a shorter key than the

Fig. 97.5 Process of transfer between compressed and standard certificate

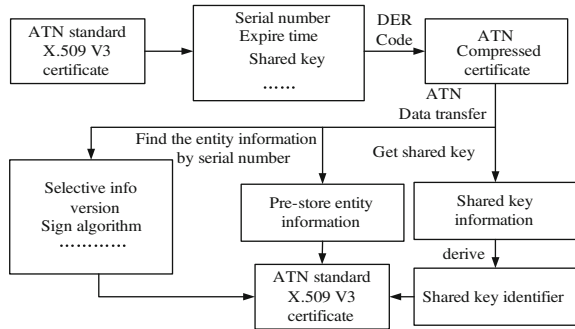


Table 97.1 ECC, RSA key size of equivalent size compared intensity

Key attack time (MPIS year)	RSA private key Size (bit)	ECC private key size (bit)	RSA/ECC Key size ratio
1,011	1,024	160	7:1
1,020	2,048	210	10:1

RSA algorithm, and large integer factorization problem based on RSA algorithm in the same security level. As shown in Table 97.1.

Table 97.2 shows that such intensity ECC key size is much less than RSA key size. The size of ECC certificate is also less than that of RSA certificate. If the compression technology is used, then the size of ECC certificates compression can decrease further. ATN-ECC not only guarantees the security but also saves the ATN deficient resources.

97.5.2 Computing Time

Standard and compressed certificates are created by using the OpenSSL API, the conversion time between them is tested. The results are shown in Table 97.3. Hardware platform for test is AMD Athlon 64X2 7750 with 2 GB RAM, while OS is Linux Fedora Core9.

As Table 97.3 shows, ECC algorithm consumes 1 ms more than RSA algorithm in generating standard certificate, because the ECC encryption algorithm requires two point multiplications. The transfer between standard and compressed certificate consumes the same time, because the certificate needs threshold compression options, exports the private key and compresses the standard certificate size, the appropriate subject key identifier default entry should be rebuild and found.

The generating procedure of standard and compressed certificates will take milliseconds to switch, which has little effect on information transmission. The size of the compressed ATN certificate is much smaller than that of standard

Table 97.2 Comparison between RSA and ECC digital certificate key size (byte)

Certificate	RSA	ECC	ECC compressed	RSA/ECC/ECC compressed ratio
ATN-CA	1,127	642	321	3.5:2:1
ATN User	839	553	236	3.55:2.34:1

Table 97.3 Test of RSA, ECC certificate compress

Time cost (ms)	RSA certificate - 1024	ECC certificate - 163	ECC certificate to ATN compressed certificate	ATN Compressed certificate to ECC certificate
First time	3.503	4.418	1.203	2.452
Second time	3.472	4.339	1.186	2.365
Third time	2.375	4.410	1.240	2.387
Average time	3.450	4.389	1.210	2.401

ECC and RSA certificates. Take all these reasons into consideration, ECC compressed certificate is more suitable for ATN air-ground network environment.

97.6 Conclusions

This chapter analyzed the X.509 certificates in ATN network system, built the PKI environment for the X.509 standard certificate generating and compressing. ECC and RSA certificate’s safe level, size and production efficiency were also compared. The ground-ground data link which uses the ECC standard certificate and ground-air data link based on compressed X.509 certificate, conversion time and efficiency are tested. Analysis showed that use the ECC in the ATN network which lacks of resources can ensure the security of network transmission performance. By compressing the size of the certificate and no significant increase in computing costs, we can reduce the network resource consumption effectively. This chapter relates only one level of CA, and does not consider multi-level and cross-certification, etc. In the future, we will build multi-level CA in the ATN environment for further research.

References

1. ICAO. Doc9705-AN/956 (2002) Manual of technical provisions for the Aeronautical Telecommunication Network (ATN) 3rd edn [S.n]. ICAO
2. McParland T, Patel V, Hughes WJ (2001) Securing air-ground communications [Cn]. Digital avionics systems conference (DASC). The 20th Conference Vol 2, 14-18 Oct 2001, pp 7A7/1- 7A7/9

3. R. Housley (1999) RFC 2459—Internet X.509 public key infrastructure certificate and CRL profile [S.n]. ISOC January
4. Chenyun M, Yan W (2008) PKI network security and authentication technology [M]. Posts&Telecom Press, pp 39–64
5. Yang W (2006) The research on ECC digital signature [D]. Wuhan University of Technology
6. Stallings W (2006) Cryptography and network security principles and practices [M], 4th edn. Publishing house of electronics industry, pp 188–210

Chapter 98

A Robust Method Based on Static Hand Gesture Recognition for Human–Computer Interaction Under Complex Background

Xingbao Meng, Jing Lin and Yingchun Ding

Abstract An appearance-based approach needs usually a perfect segmentation. However, it is a difficult task especially under complex background. As a result, it limits the robustness for application. In this chapter, we design a new method for static hand gesture recognition in complex background for human–computer interface (HCI). In this method, we do not need perfect segmentation or hand tracking. The Hu invariant moment features are extracted from a binary image after simple segmentation and served as the input of our classifier, which is constructed beforehand based on support vector machines (SVM) algorithm. Furthermore, a Euclidean distance is calculated to combine with SVM model for avoiding the non-hand gestures. Tests on the testing dataset show the proposed method exhibits a recognition rate near 100%. Experimental results on a simple HCI system on real-time demonstrated the effectiveness, speediness and robustness of the system under cluttered background.

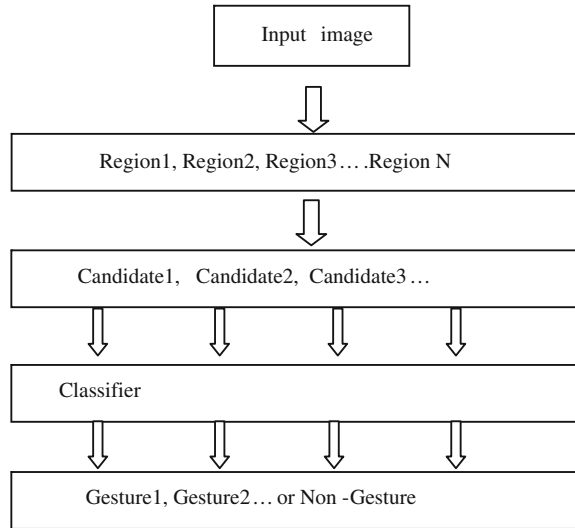
Keywords Hand gesture recognition · Support vector machine (SVM) · Hu invariant moments · Human–computer interaction (HCI)

98.1 Introduction

Vision-based hand gesture analysis has greatly increased so far and many different methods have been proposed [1–3]. These methods can be subdivided into two main categories: model-based and appearance-based approach. Conventional model-based approach for image analysis attempts to infer the pose of the palm and finger joints. This method searches for the kinematic parameters which map

X. Meng · J. Lin · Y. Ding (✉)
Department of Physics, University of Chemical and Technology,
Beijing 100029, China
e-mail: dingyc@mail.buct.edu.cn

Fig. 98.1 The flow of proposed method



the 2D projection images to 3D hand model. 3D hand tracking and gesture estimation are main issues of this subject. However, the majority of appearance-based methods depend on the parameters extracted from images, which include contours, edges, image moments and image eigenvectors [4]. These approaches have the advantages of low complexity and implementing easily. Given the robustness and speediness of a HCI system, these approaches are more preferable.

The method proposed in this chapter belongs to appearance-based methods. It is effective and robust against complex background. It has also a fast processing speed. Figure 98.1 shows the flow of the proposed method. Firstly, we segment the input image into a collection of disjoint regions by Otsu algorithm. Then remove the noise regions by their area. As a result, there are only several regions conserved in binary image. By this step, we can segment the hand regions from the scene, but we do not know which regions are the true hand regions. We serve these rest regions as hand region candidates and extract their Hu invariant moments respectively. These Hu invariant moments compose a seven-dimensional vector. By a classifier we constructed preliminarily, we can classify whether these candidates are defined hand gestures or not. In this method, the key steps are hand regions segmentation from complex background and classifier construction. The classifier is constructed based on SVM algorithm and a calculated Euclidean distance. The proposed method have been experimented on a simple HCI application: hand gesture controlling MP3 player.

The rest of the chapter is organized as follow. The segmentation of hand image and extracting features are addressed in the next section. The problem of hand samples training by SVM and constructing classifier are then presented in Sect. 3. Section 4 shows the experimental results of the proposed method in detail. Finally, conclusions are given in Sect. 5.

98.2 Hand Region Segmentation and Extracting Features

Hand image segmentation separates the hand image from the background. It is the first important step in every hand gesture recognition system. However, Vision-based approaches all share the problem related to the vagaries of low-level segmentation, so it is a difficult task to get a perfect hand regions especially under complex environment. Hand tracking techniques, such as Cam Shift, Kalman filter [5] and Particle filter [6], can help to avoid this problem, but these methods are known to be hard tasks. In this chapter, we do not require both perfect segmentation and hand tracking. Our target is only to get near-complete hand regions in the binary image after segmentation. The details are given in the following subsections.

98.2.1 Image Binaryzation by Otsu Algorithm

Skin color is the most commonly used cue for segmenting image parts from the hand or face in gesture analysis. However, approaches based on predefined skin color model suffer from sensitivity with respect to changing illumination conditions [7]. Otsu method which suggested by Otsu [8] is used to automatically perform histogram shape-based image thresholding or the reduction of a gray-level image to a binary image. Compared with skin color model, Otsu algorithm is more efficient for our method, so it is considered in this chapter. Figure 98.2a and c show the segmentation of hand image under simple background and under complex background respectively by Otsu algorithm. Furthermore, Fig. 98.2d shows the contours of the candidates. Figure 98.2e shows the input image with the bounding rectangles of candidates superimposed.

98.2.2 Find Contours and Calculate Their Area

This step is in order to remove the unnecessary regions. By using the finding contours interface which provided in OpenCV, we can find all the contours of the binary image. Areas of the contours are then calculated as follows:

$$Area = \left| \frac{1}{2} \sum_{i=0}^{N-2} (x_i^* y_{i+1} - x_{i+1}^* y_i) \right| \quad (98.1)$$

Where $N-1$ is total amount of all the points of each contour. x_i and y_i are the X coordinate and Y coordinate of the points respectively. By the areas, we can set a threshold T . The region whose area is below T is then removed. This approach is able to avoid most of unnecessary regions.

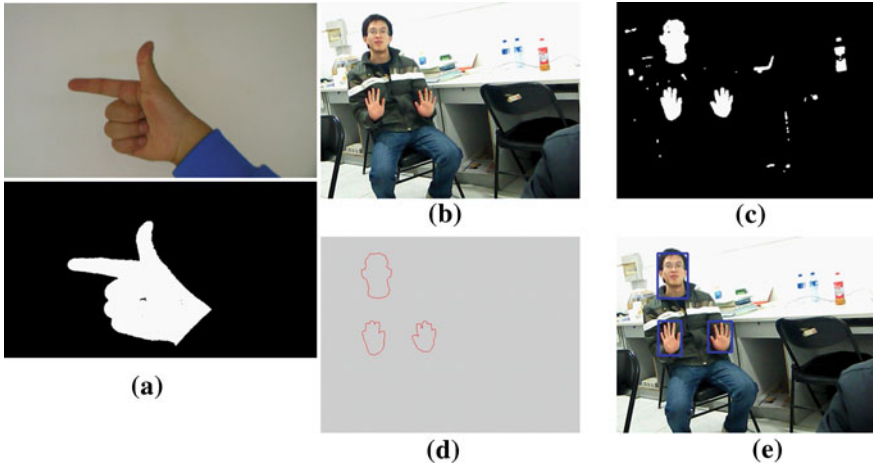


Fig. 98.2 **a** Segmentation by Otsu method under simple background. **b** The RGB color image. **c** Binary image after segmented by Otsu. **d** Contours of the rest regions, or “candidates”. **e** the input image with the bounding rectangles of candidates superimposed

98.2.3 Extracting Features

Hand segmentation is followed by feature extraction. Image moments, which are fast to compute, provide average coarse summary of global averages of orientation and position [9]. Hu invariant moments which proposed by Hu [10] are useful to describe objects after segmentation. They are linear combination of the central moments. The Hu set of invariant moment is invariant to scale, rotation, and reflection.

As mentioned before, we could find the contour of relative candidate in the binary image. Then, the bounding rectangles of each contour are employed to set interest regions of image, also called region of interest (ROI), as shown in Fig. 98.2e. The moments of each ROI are then computed for characterizing the candidate.

98.3 Classifier Construction

Classifier is crucial to our method. The classifier should achieve two basic tasks. One is that it can classify the defined gestures rapidly and accurately. Another is that it must avoid the non-hand gestures effectively. Recently, there are some promising techniques such as SVM, recurrent neural network (RNN), and dynamic time warping (DTW). Since SVM shows superior performance in classifying with good generalization properties especially for limited samples [11, 12], it is

Fig. 98.3 The binary images of defined gestures



employed to build a classifier in our method. In this chapter, the first work is implemented by a SVM classifier. Euclidean distance of each hand gesture is then calculated to give a threshold (T_i) for avoiding the non-hand gestures.

98.3.1 Support Vector Machines

In this chapter, the directed acyclic graph (DAGSVM) algorithm presented by Platt [13] is employed to solve a 5-class classification problem. This algorithm is based on the idea which use a number of SVM-based binary classifiers combined to tackle the multi-class classification problems. For K -classes classification problem, it builds $K(K - 1)/2$ classifiers each of which is trained on data from two classes. Then it employs a rooted binary directed acyclic graph (DAG) which has $K(K - 1)/2$ internal nodes and K leaves. Each node is a binary SVM of i th and j th classes. Given a test sample x , starting at the root node, the binary decision function is evaluated. The sample is then classified. By the DAG, we can find the best result finally.

98.3.2 Data Set Training

Given a set of training samples, each marked as belonging to different categories, SVM training algorithm can build a model that predicts whether a new sample falls into which category. As mentioned in Sect. 2.3, Hu invariant moments extracted from candidate window of the binary image compose a seven dimensional feature vector $V[h_1, h_2, h_3, h_4, h_5, h_6, h_7]$. This feature vector represents a training sample. In our work, five gestures have been defined. Figure 98.3 shows the binary images of the defined hand gestures. The training examples include 2,500 hand binary images which are from the color hand images under simple background. Each color hand image has only one hand pose in the scene.

In fact, the scale of the moment values is too small. In order to avoid the complicated calculation of these values, we scale the training data set into the range $[0, 10]$. The RBF kernel function is chosen for training.

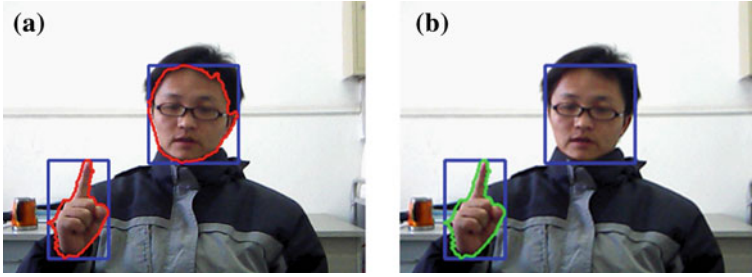


Fig. 98.4 Recognition results based on decision functions and proposed method. The contours superimposed with *red* or *green* color signify the objects are served as true hand posture

98.3.3 Classifying the Hand Gesture and Non-Hand Gestures

After training step, we can build a SVM model to classify an unknown sample effectively. However, it is unable to classify the non-hand gestures correctly. For avoiding the non-hand gestures, a conventional method is dependent on setting a threshold by the SVM decision function

$$f(x) = \sum_{i=1}^N a_i y_i k(x, x_i) + b \tag{98.2}$$

where $k(x, x_i)$ is the kernel function, x_i represents training example, $y_i \in \{ -1, +1 \}$ represents class labels. For N classes classification problem, there should be $N*(N - 1)/2$ threshold values. By these threshold values, we could avoid the non-hand gestures. However, on the test time, we have found that this method is not enough effective. Figure 98.4a shows that the head region is mis-recognized as a true hand gesture based on the conventional method.

For addressing this problem, we propose a novel method using the distance of new sample and the center of each class. Firstly, we compute the center of each category as follows:

$$S_{avg}(i) = \frac{1}{N} \sum_{j=0}^N V_{i,j}; i \in (1, 2, 3, 4, 5), \quad (j = 0, 1, 2, 3 \dots N) \tag{98.3}$$

Where N is the total amount of training samples, $V_{i,j}$ is the feature vector of sample, $S_{avg}(i)$ is the center of corresponding category. A Euclidean distance then is calculated as follows:

$$D_i = |V_{i,j} - S_{avg}(i)|; i \in (1, 2, 3, 4, 5), \quad (j = 0, 1, 2, 3 \dots N) \tag{98.4}$$

Where D_i represents the Euclidean distance between category i and corresponding testing samples. Finally, we set a threshold T_i according to the distance. Given a new sample, it is classified by SVM model firstly. The Distance of the new

Table 98.1 Parameters of the Unify 6100 web camera

Price	Image resolution	Interface	Frame speed	Sensor
¥49.0	5 megapixels 640 × 480	USB 2.0	30 fram/s	CMOS

Table 98.2 Recognition rates of test samples

Test samples	Gesture1	Gesture2	Gesture3	Gesture4	Gesture5
Recognition rate	99.6%	99.3%	100%	98.9%	98.4%

sample is then calculated. If the distance is below T_i , the sample is then seen hand gesture, vice versa. Figure 98.4b shows our method can avoids the non-gesture effectively compared with the former method.

98.4 Experimental Results

98.4.1 Experiments on Testing Samples

We have trained our algorithm on a set of 2,500 training data and tested on 2,500 test samples respectively. The dataset are from 5000 RGB image collected by a web camera. Table 98.1 shows the web camera parameters. A set of 2,500 binary images, which compose the testing samples, are cut from 2,500 RGB images with one hand in scene under complicated background and different lighting condition. Note that the size of all the samples is not uniform. The testing samples are then divided into five groups for five gestures averagely.

Tests on the testing samples show that we achieve recognition rates demonstrated in Table 98.2. The results validate the effectiveness of this proposed method.

98.4.2 Tests on Real-Time

Our method has also conducted on a simple HCI system on real-time. It is developed using the OpenCV library and implemented on a PC with 1.73G processor and 2.0G memory. It takes 0.04 s for recognizing one image. On real-time test, our method has been embedded in a MP3 player. On this system, we use only one hand for control. In fact, we can use both hands to control. When the recognition of one frame is combination of different gestures, it is able to achieve different functions. Tests on the MP3 player demonstrate the effectiveness, speediness and robustness of the system.

98.5 Conclusion

In this chapter, we describe a robust method based on static hand gesture recognition for HCI. In terms of performance, our method is fairly robust against complicated background. It is also simple, fast, and effective. Experiments on the test data set show the recognition accuracy is near 100%. For HCI application, we can use both hands, even several hands from several persons, for controlling and manipulating the system by this method. However, a fatal weakness of our method is that it is difficult to get a complete hand region, when the hand and some bright color (such as red and yellow) objects overlap. As a result, the system cannot recognize effectively. Furthermore, our method is also limited to static gestures currently. Dynamic gesture recognition will be extended in the future.

Acknowledgment This work was supported by National Natural Science Foundation of China under the Grants 60978006.

References

1. Sharma R, Huang TS (1997) Visual interpretation of hand gestures for human–computer interaction: a review. *IEEE Trans Pattern Anal Mach Intell* 19(7):677–695
2. Mitra S, Acharya T (2007) Gesture recognition: a survey. *IEEE Trans Syst Man Cyberne Part C Appl Rev* 37(3):311–324
3. Erol A, Bebis G, Niculescu M (2007) Vision-based hand pose estimation: a review. *Comput Vis Imag Underst* 108:52–73
4. Ge SS, Yang Y, Lee TH (2008) Hand gesture recognition and tracking based on distributed locally linear embedding. *Image Vis Comput* 26:1607–1620
5. Dankers A, Barnes N, Zelinsky A (2007) MAP ZDF segmentation and tracking using active stereo vision: hand tracking case study. *Comput Vis Imag Underst* 108:74–86
6. Shan C, Tan T, Wei Y (2007) Real-time hand tracking using a mean shift embedded particle filter. *Pattern Recogn* 40:1958–1970
7. Yang R, Sarkar S (2009) Coupled grouping and matching for sign and gesture recognition. *Comput Vis Imag Underst* 113:663–681
8. Otsu N (1979) A threshold selection method from gray level histogram, *IEEE SMC-9* (1):62–66
9. Tin H, Maung H (2009) Real-time hand tracking and gesture recognition system using neural networks. In: *Proceedings of World Academy of Science, Engineering and Technology*, p 50
10. Hu MK (1962) Visual pattern recognition by moment invariants. *IRE Trans Inform Theor* 8:179–187
11. Bradski G, Kaehler A (2008) *Machine learning. In: learning OpenCV: computer vision with the OpenCV library*, O'Reilly Media USA pp 495–517
12. Jianjun Y, Hongxun Y, Feng J (2004) Based on HMM and SVM multilayer architecture classifier for Chinese sign language recognition with large vocabulary. In: *Proceedings of the third international conference on image and graphic*
13. Platt JC, Cristianini N, Shawe-Taylor J (2000) Large margin DAGs for multiclass classification. *Adv Neur Inform Process Syst* 12:547–553

Chapter 99

Research on Privacy Preserving Based on K-Anonymity

Xiao-ling Zhu and Ting-gui Chen

Abstract K-anonymity is a highlighted topic of privacy preservation research in recent years, for it can effectively prevent privacy leaks caused by link attacks; so far K-anonymity has been widely used in all fields. In this chapter, based on the existing K-anonymity privacy protection of the basic ideas and concepts, K-anonymity model, and enhanced the K-anonymity model has been studied, finally, the future directions in this field are discussed.

Keywords Privacy preservation · K-anonymity · Generalization and suppression · The enhanced K-anonymity

99.1 Introduction

With the rapid development of information technology and networks, the emergence of data mining make it possible for people to get useful information from a large database, data mining has been widely used in retail, health care, education, insurance, banking and other fields. Data Release, as an effective means of information exchange and data sharing, has brought great convenience for data retrieval and use. However, the leakage of sensitive information are also increasingly prominent

X. Zhu
College of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, Zhejiang, China
e-mail: lynn_525@sina.com

T. Chen (✉)
Contemporary Business and Trade Research Center, Zhejiang Gongshang University, Hangzhou 310018, Zhejiang, China
e-mail: ctgsimon@gmail.com

during the data released, data Mining providing people with a strong knowledge discovery, while it also posing a threat of personal privacy, privacy protection has become a hot topic in database security research. A reasonable and effective method of protection, which can protect the user's privacy and keep the data available, is the trend of developments in information security.

The existing data mining methods are: heuristic-based privacy protection technology, cryptography-based privacy preserving techniques, and privacy protection technology based on the reconstruction, for different methods they applied well in related fields, and can protect user's privacy information to some extent. Now the commonly used method is the K-anonymity, aiming at the existing K-anonymity to summarize the main ideas and models and analysis, and the future direction of development are discussed.

99.2 K-Anonymity Model

In real life, some agencies often should publish some relevant data, such as for the need of research on population statistics, medical and health. Although the published data has been hidden personal identifiable information, such as name, ID number, telephone number and other attributes. The attacker get data through other channels to obtain the data link operations to infer the privacy of data, resulting in privacy leak, this process is called link attacks, which mainly work in quasi-identifier. Currently, to avoid link attacks, and protecting private information from being leaked, K-anonymous method is the most common.

K-anonymity was raised in 1998 by Samarati and Sweeney, it requires the published data exists a certain number (at least for the K) whose records cannot be distinguished, so that an attacker cannot distinguish the respective privacy information of a specific individuals, thereby it prevents the leakage of personal privacy. User can specify a parameter K for the greatest risk of information leakage that they can receive in K-anonymous. It protects the privacy of individuals to some extent, while it also reduces the availability of data; the work of K-anonymity focuses mainly on the protection of private information and increase their availability. Since the proposed, K-anonymity has been the general concern of academia; many scholars at home and abroad research and develop the technology in different way (Fig. 99.1).

99.3 K-Anonymity Model for the Main Algorithm

99.3.1 *Generalization and Suppression*

The aim of the current data mining is focused on how to set the original data by anonymous effectively data, and at the same time to achieve the best anonymity, the maximum data availability, the minimum spending of time and space.

Fig. 99.1 Link attacks

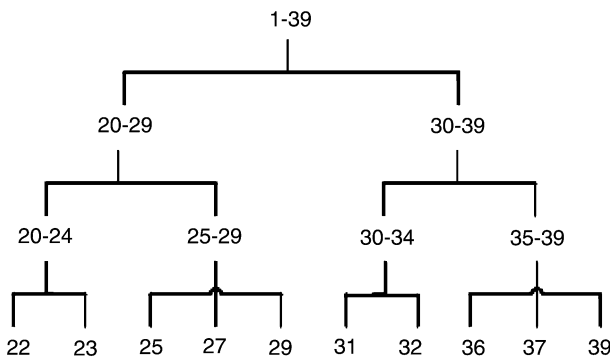
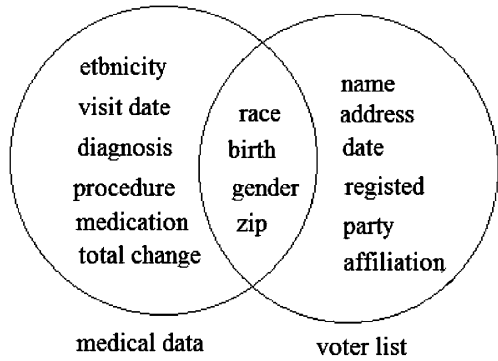


Fig. 99.2 Generalization of age

Different with distortion of data interference methods such as distortion, disturbance and randomization, K-anonymity can maintain the authenticity of the data. The common use to achieve K-anonymity is through generalization and suppression. Generalization can be divided into value generalization and domain generalization, Fig. 99.2 describes the generalization of age. Suppression is to delete or hide some of the attributes in the data table directly to protect patient privacy.

Table 99.1 shows a medical information form, after inhibition, it does not include sensitive information such as their names, Medicare numbers, home address, and ID numbers and so on. But there exist quasi-identifier information such as gender, age, zip code, etc. Through the collection of these attributes, attackers can tell the personal information indirectly, so it is possible to disclose patient medical information.

To prevent the disclosure of private information and to protect patient privacy, Table 99.2 generalized the data, the patient information in the table specified the age in a large range, zip code in the "*" indicates any number, the rationale is as follows in Table 99.2.

Table 99.1 Original medical information

ID	Quasi-identifier			Other	Sensitive information
	Gender	Age	Zip		
1	Male	25	644625	Gastric ulcer
2	Male	27	650500	AIDS
3	Male	22	671060	Flu
4	Male	29	675067	Neurasthenia
5	Female	34	671060	Flu
6	Female	31	671060	Hepatitis
7	Female	37	650500	Neurasthenia
8	Female	36	650500	Flu

99.3.2 The Enhanced K-Anonymity Model

The original K-anonymity is to prevent identity disclosure, but through the property will still bring the disclosure of information. In order to solve the shortcomings in K-anonymity, Machanavajjhala proposed K-anonymity model for the two attack methods, homogeneity attack and the background knowledge attack. Homogeneity attack is the attacker derived K-anonymous table information of a sensitive individual; background knowledge inference attacks is that attackers use some additional information in advance to carry out attacks. The two attacks will result in disclosure of sensitive property in K-anonymity.

99.3.2.1 L-Diversity Model

Machanavajjhala, who gives L-diversity model to reduce the privacy, leaks by increasing the diversity of an anonymous in group of sensitive properties. In a published table, a K-anonymous group contains at least L sensitive properties that on behalf of a good sense of representative. For example, in Table 99.2, tuples whose ID number are 1, 2, 3, 4, 6 form a group with 5 species diversity, their frequency was 12.5, 12.5, 37.5, 25 and 12.5% in value, and no one has predominant function, so it can be set by L-diversity model.

However, in this model, it is difficult to speculate how much background knowledge the attacker knows about, any posted data will become unsafe if the other knows a lot about the patient’s background knowledge, there does not have a good way to set the parameters in L-diversity model.

99.3.2.2 (α , k)-Anonymous Model

In (α , k)-anonymous model, property with higher degree of sensitivity has been better protected, by constraining the frequency of anonymity property values in the

Table 99.2 Generalization of the data

ID	Quasi-identifier			Other	Sensitive information
	Gender	Age	Zip		
1	Male	25	644***	Gastric ulcer
2	Male	27	650***	AIDS
3	Male	22	671***	Flu
4	Male	29	675***	Neurasthenia
5	Female	34	671***	Flu
6	Female	31	671***	Hepatitis
7	Female	37	650***	Neurasthenia
8	Female	36	650***	Flu

sensitive group less than a given parameter α , so it avoids the situation that the frequency of certain sensitive information too high, increases the diversity of sensitive values, and prevents the consistency attack. For example, in a data sheet of a hospital medical records, some patients' illness are more sensitive and need protection, such as AIDS; while there are many diseases that are very common and not need to protect, such as flu, so under such a circumstance that a data sheet with different protection needs that the higher ones need to be protected, a more suitable idea comes out, that is (α, k) -anonymous mode.

In (α, k) -anonymous model, only the relevant and sensitive property values are necessary to protect, so only consider the sensitive attribute value, such as AIDS. This model allows inference between credibility to the sensitive lower than α , it is simple and effective way to prevent sensitive to the value for the homogeneity attack (Table 99.3).

Table 99.3 provides a (α, k) -anonymous form. In the table, flu and neurasthenia is not considered as sensitive information, from the (female 30–39, 671***) to the reasoning of depressive neurosis credibility value is 25%.

The attackers could not see the value of property of higher degree of sensitivity after processing by (α, k) -anonymous model, so it can protect the security of this kind of information effectively.

While (α, k) -anonymous model only considers the sensitive attribute the highest level-sensitive property value, there is no other level of sensitive property to process property values, and does not take into account the sensitivity of the same property value, so the same level of sensitive attributes property values or the existence of other levels of privacy disclosure.

99.3.2.3 (α, L) -Diversification K-Anonymity Model

(α, L) -anonymity model considers only the highest level of the sensitive property value, but neither other level of sensitive property values, nor take into account the sensitivity property values of the same issues. And sometimes it is hard to

Table 99.3 (0.25, 3)-Anonymous

ID	Quasi-identifier			Other	Sensitive information
	Gender	Age	Zip		
1	Male	25	644***	Flu
2	Male	27	650***	Neurasthenia
3	Male	22	671***	Flu
4	Male	29	675***	Neurasthenia
5	Female	34	671***	Flu
6	Female	31	671***	Flu
7	Female	37	650***	Neurasthenia
8	Female	36	650***	Flu

Table 99.4 Health categories

ID	Value	Sid
1	AIDS, hepatitis	1
2	Gastric ulcer	2
3	Neurasthenia, flu	3

determine whether an illness is the higher degree of sensitivity that needs to be protected, or lower degree of sensitivity that not need to be protected

K-anonymity model in (α, L) -diversification K-anonymity model can determine flexibility to protect the privacy or not according to the extent of protection. At the same time, have special treatment to the high-level sensitive property values of privacy protection, and with better privacy protect effect. Table 99.4 is a Health categories table (where Sid is sensitive to the privacy level of property value) is Sid the data in Table 99.2 for the classification.

(α, L) -diversification K-anonymity model is a data table to meet K-anonymity, α -distribution, and the number of Sid in the sensitive group is no less than L at the same time. α -distribution constraint is that all of the sensitive attributes of privacy frequency of Sid from equivalent class less than α which is a given data, that is sensitive to all the anonymous group of private property when the degree of the frequency Sid $\leq \alpha$, where a is user-defined number, and $0 < \alpha < 1$.

Assuming the Sid need to protect equals to 1, privacy protection level is as classification in Table 99.4, while Table 99.5 is a constraint to meet the 0.5 distribution of a data set. In Table 99.5, there are two anonymous group: {1, 2, 3, 4} and {5, 6, 7, 8}, in the first anonymous group, the frequency is 0.25 when Sid equals to 1, and in the second one the frequency is 0.25, so for all anonymous groups when Sid equals to 1 the frequency of Sid ≤ 0.25 . Then it meets to (0.25, 3)-diversification 4-anonymous as shown in Table 99.5.

Table 99.5 provides a data table satisfying (0.25, 3)-diversification 4-anonymous model, according to the foregoing, the distribution of this data sheet meet to the constraints of 0.25, and the number of each anonymous tuple is no less than 4,

Table 99.5 (0.25, 3)-Diversification 4-anonymous

ID	Quasi-identifier			Other	Sensitive information
	Gender	Age	Zip		
1	Male	25	644***	2
2	Male	27	650***	1
3	Male	22	671***	3
4	Male	29	675***	3
5	Female	34	671***	3
6	Female	31	671***	1
7	Female	37	650***	3
8	Female	36	650***	3

the different number of the Sid values in the table equals to 3, so Table 99.4 satisfy(0.25, 3)-diversification 4-anonymous model.

Construct (α, L) -diversification K-anonymity model algorithm as follow:

Input: data set T;

Output: data table T * that meet (α, L) -diversification K-anonymity model.

- (1) According to the health status in Table 99.5, the value of the sensitive property in table T was replaced by Sid, who represent the sensitive level, then table T turns into table T1.
- (2) Construct a data table T2 that consistent with (α, K) data tables anonymous model, in which Sid is regarded as the sensitive property, and the generalization in accordance with top-down algorithm.
- (3) For each anonymous group, check the privacy level Sid for the number of different values.
- (4) IF $(3L)$.
- (5) Return the final table T*.
- (6) Else
- (7) Of all the anonymous groups that does not meet the requirement, have them further generalization or exchange tuples to make sure that the value of Sid is greater than L.
- (8) Returns the final data table T*.

First, the value of the sensitive property in table T was replaced by Sid who represent the sensitive level, according to the health status in Table 99.5, then table T turns into table T1, and then turn T1 into anonymity, so as to meet K-anonymous and α -distribution, in this step, top-down local generalization algorithm has been used. And then check whether the generalization of the data sheet meets (α, L) -diversification K-anonymity model conditions, that is the privacy degree of different values greater than or equal to L. If all anonymous groups met for the condition, the entire data table is the final meet (α, L) -diversification K-anonymity model data tables, and if not, have further generalization or suppression to make sure that the different values of privacy degree number is greater than L.

In (α, L) -diversification K-anonymity model, the parameter α can be set by users themselves according to their privacy protection needs. It provides an effective solution to the problem of imbalance distribution of sensitive attributes, divides the attribute values on the sensitive level of privacy protection, and protects the privacy effectively.

The enhanced K-anonymity models are mainly based on K-anonymity and to make the information security. L-diversity model in which properties are divided into groups, by increasing the variety in groups, it can prevent attackers from locating the information; (α, k) -anonymous model, by processing to the higher level of sensitive attribute and make its feasible degree smaller than α , can effectively protect sensitive information of higher degree; (α, L) -diversification K-anonymity model divides the properties according to the level of sensitive information which is determined by the users themselves flexibility, and for the sensitive attributes with higher degree value require special treatment.

99.4 Development Trend and Summary

Because K-anonymity can prevent users' private information from being leaked in the released environment, ensure the authenticity of the published data, it applications widely in the industry and attracted widespread attention. However, nowadays the majority of K-anonymity algorithms are based on static data sets, and in the real world, data is constantly changing, including changes in forms of data, attribute changes, adding new data, and deleting the old data. Besides, the data between data sets are likely to be interrelated, how to achieve privacy protection in a much more complex environment with dynamic data, still need further study.

Acknowledgments This research is supported by Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20103326120001), Zhejiang Provincial Natural Science Foundation of China (No. Y7100673), Zhejiang Provincial Social Science Foundation of China (Grant No. 10JDSM03YB), and the Contemporary Business and Trade Research Center of Zhejiang Gongshang University (No. 1130KUSM09013 and 1130KU110021) as well as Research Project of Department of Education of Zhejiang Province (No. Y200907458). We also gratefully acknowledge the support of Science and Technology Innovative project (No. 1130XJ1710215).

References

1. Sweeney L (2002) Achieving k-anonymity privacy protection using generalization and suppression [J]. Intern J Uncertain Fuzziness Knowl-based Syst 10(5):571–588
2. Sweeney L (2002) k-Anonymity: a model for protecting privacy [J]. Intern J Uncertain Fuzziness Knowl-based Syst 10(7):557–570
3. Sweeney L (2001) Computational disclosure control: a primer on data privacy protection, Ph. D. thesis Massachusetts Institute of Technology, pp 67–82

4. Samarati P, Sweeney L (1998) Generalizing data to provide anonymity when disclosing information (abstract). In: Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems, Seattle, p 188
5. Long Q (2010) Privacy protection based on k-anonymity [J]. *Sci Technol Assoc Forum* 3:41–43
6. Qin X, Men A, Zou Y (2010) Privacy protection based on Kanonymity algorithms [J]. *J Chifeng Univ* 26(5):14–16
7. Wang P, Wang J (2010) Progress of research on K-anonymity privacy2preserving techniques. *J Chi-feng Univ (Nat Sci Ed)* [J] 27(6):2016–2019
8. Kan Y, Cao T (2010) Enhanced privacy preserving K-anonymity model (α , L)-diversity K-anonymity. *Comput Eng Appl* 46(21):148–151
9. Cen T, Han J, Wang J, Li X (2008) Survey of K-anonymity research on privacy preservation. *Comput Eng Appl* 44(4):130–134
10. Wong R, Li J, Fu A et al (2006) (α , k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing [C]. In: International conference on knowledge discovery and data mining, vol 123. pp 754–759
11. Aggarwal G, Feder T, Kenthapadi K et al (2005) Anonymizing tables[C]. In: Proceedings of the 10th international conference on database theory (ICDT05), Edinburgh, Scotland, pp 246–258
12. Machanavajjhala A, Gehrke J, Kifer D (2007) l-Diversity: privacy beyond k-anonymity [J]. *ACM Trans Knowl Discov Data* 1(1):24–35
13. Li N, Li T, Venkatasubramanian S (2007) T-closeness: privacy beyond k-anonymity and l-diversity [C]. *ICDE*, pp 106–115
14. Fung B, Wang K, Yu P (2005) Top-down specialization for information and privacy preservation [C]. In: Proceedings of the 21st international conference on data engineering (ICDE05), Tokyo, Japan
15. Meyerson A, Williams R (2004) On the complexity of optimal K-anonymity [C]. In: Proceedings of the 23rd ACM-SIGMOD-SIGACTSIGART symposium on the principles of database systems, Paris, France, pp 223–228
16. LeFevre K, DeWitt DJ, Ramakrishnan R (2005) Incognito: efficient full-domain k-anonymity. In: SIGMOD' 05 proceedings of the 2005 ACM SIGMOD international conference on management of data, pp 49–60
17. Bayardo R, Agrawal R (2005) Data privacy through optimal K-anonymity. In: ICDE05: the 21st international conference on data engineering, pp 217–228

Chapter 100

Existential Forgery Attack Against One Strong Signature Scheme

Zhengjun Cao and Qian Sha

Abstract The GMY signature is based on the intractability of factorization. In the scheme, messages will be signed in a bit-by-bit fashion. It claims that an adversary cannot forge a new signature, even after seeing a large number of genuine signed messages. In this chapter, we present an existential forgery attack against it. It shows the adversary succeeds in forging the signature of one message, not necessarily of his choice.

Keywords Signature · Existential forgery · Known-signature attack

100.1 Introduction

Digital signature is the most significant contribution provided by public-key cryptography, which is introduced by Diffie and Hellman [1]. Digital signature [2] has many applications in information security, including authentication, data integrity, and nonrepudiation. The signature is authentic: it convinces the document's recipient that the signer deliberately signed the document. It must be verifiable: if a dispute arises about the origin of a signed document (caused by either a lying signer trying to repudiate his signature, or a fraudulent claimant), an unbiased third party should be able to resolve the matter equitably, without requiring access to the signer's secret information (private key) [3]. There are various signature models [4]. We refer to [5] for different authentication properties of various signatures.

Z. Cao (✉) · Q. Sha
Department of Mathematics, Shanghai University, Shanghai, China
e-mail: caozhj@yahoo.cn

The following basic definitions related to digital signature can be found in Ref. [3].

1. A digital signature is a data string which associates a message (in digital form) with some originating entity.
2. A digital signature generation algorithm (or signature generation algorithm) is a method for producing a digital signature.
3. A digital signature verification algorithm (or verification algorithm) is a method for verifying that a digital signature is authentic (i.e., was indeed created by the specified entity).
4. A digital signature scheme (or mechanism) consists of a signature generation algorithm and an associated verification algorithm.

A theoretical treatment of digital signature security can be found in Ref. [3, 6–8]. Usually, attacks against digital signatures are classified as follows.

Key-only attack. In this attack, the adversary knows only the public key of the signer and therefore only has the capability of checking the validity of signatures of messages given to him.

Known signature attack. The adversary knows the public key of the signer and has seen message/signature pairs chosen and produced by the legal signer.

Chosen message attack. The adversary is allowed to ask the signer to sign a number of message of the adversary's choice.

Similarly, several levels of success for an adversary are classified:

Existential forgery. The adversary succeeds in forging the signature of one message, not necessarily of his choice.

Selective forgery. The adversary succeeds in forging the signature of some message of his choice.

Universal forgery. The adversary, although unable to find the secret key of the signer, is able to forge the signature of any message.

Total break. The adversary can compute the signer's secret key.

The level of security required in a digital signature scheme may vary according to the application. For example, in situations where an adversary is only capable of mounting a key-only attack, it may suffice to design the scheme to prevent the adversary from being successful at selective forgery. In situations where the adversary is capable of a message attack, it is likely necessary to guard against the possibility of existential forgery [3].

It is usual to an adversary that trying to produce a value $\sigma(M)$ for a given message M such that they consist of a valid signature

$$\text{Sig}(M) = (M, \sigma(M))$$

not to generate a message \bar{M} for a given value $\sigma(?)$ such that they consist of a valid signature

$$\text{Sig}(\bar{M}) = (\bar{M}, \sigma(?))$$

Most of signature schemes don't admit the possibility to forge in both sides.

The GMY signature scheme [8] is based on the intractability of factorization. Its security has been widely accepted in the past decades. In the scheme, messages will be signed in a bit-by-bit fashion. The authors claimed that an adversary cannot forge a new signature, even after seeing a large number of genuine signed messages. In this paper, we will present an existential forgery attack against it.

100.2 Preliminary

Let p be a prime. For any integer a , the Legendre symbol of $a \in \mathbb{Z}_p^*$ is defined as:

$$\left(\frac{a}{p}\right) = a^{\frac{p-1}{2}} \pmod p$$

The equation

$$a \equiv x^2 \pmod p$$

has solutions if and only if $\left(\frac{a}{p}\right) = 1$.

A quadratic residue mod N is an element $a \in \mathbb{Z}_N^*$, such that the equation $a \equiv x^2 \pmod N$ has solutions in \mathbb{Z}_N^* .

Let N be the product of two distinct primes, $N = pq$, then every quadratic residue, $a, \pmod N$ has four square roots mod $N : x, -x, y, -y$. The Jacobi symbol is a function from the integers into $\{0, 1, -1\}$. For $x \in \mathbb{Z}_N^*$, the Jacobi symbol of x is denoted by $\left(\frac{x}{N}\right)$, where

$$\left(\frac{x}{N}\right) = \left(\frac{x}{p}\right) \left(\frac{x}{q}\right)$$

We will call an integer, N , a Blum integer, if $N = pq$ such that $p \equiv q \equiv 3 \pmod 4$. If N is a Blum integer, we then have the following facts:

- (1) For all $x \in \mathbb{Z}_N^*$, $\left(\frac{x}{N}\right) = \left(\frac{-x}{N}\right)$
- (2) For all $x, y \in \mathbb{Z}_N^*$, if $x^2 \equiv y^2 \pmod N$ and $x \not\equiv \pm y$ then $\left(\frac{x}{N}\right) = -\left(\frac{y}{N}\right)$

Let N be a Blum integer and

$$\mathcal{Q}_1^N := \{x \in \mathbb{Z}_N^* : \left(\frac{x}{p}\right) = 1 \text{ and } \left(\frac{x}{q}\right) = 1\}$$

$$\mathcal{Q}_2^N := \{x \in \mathbb{Z}_N^* : \left(\frac{x}{p}\right) = -1 \text{ and } \left(\frac{x}{q}\right) = -1\}$$

$$\mathcal{Q}_3^N := \{x \in \mathbb{Z}_N^* : \left(\frac{x}{p}\right) = 1 \text{ and } \left(\frac{x}{q}\right) = -1\}$$

$$\mathcal{Q}_4^N := \{x \in \mathbb{Z}_N^* : \left(\frac{x}{p}\right) = -1 \text{ and } \left(\frac{x}{q}\right) = 1\}$$

Q_1^N, Q_2^N, Q_3^N and Q_4^N partition \mathbb{Z}_N^* in equinumerous classes. Q_1^N is the set of quadratic residues mod N . Let y_i be a given element in Q_i^N for $1 < i < 4$. We have [8]:

Lemma 1 For all $x \in \mathbb{Z}_N^*$, there exists a unique $1 < i < 4$ such that $xy_i \pmod{Q_1^N}$.

Lemma 2 Let y in Q_3^N . Then $-y$ is in Q_4^N , and for all $x \in Q_1^N$ either $xy^{-1} \pmod{N} \leq \frac{N}{2}$ or $x(-y)^{-1} \pmod{N} \leq \frac{N}{2}$, but not both.

Lemma 3 Let x be a quadratic residue mod N . Then, x has four square roots: x_1, x_2, x_3 and x_4 such that $x_1 \in Q_1^N, x_2 \in Q_2^N, x_3 \in Q_3^N$ and $x_4 \in Q_4^N$.

100.3 Review of the GMY Strong Signature

The signature scheme [8] can be described as follows.

100.3.1 Setup

(1) User A chooses two random numbers p, q such that

$$p \equiv q \equiv 3 \pmod{4}$$

and computes $n = pq$.

(2) Choose $a \in_R \mathbb{Z}_n^*$ such that Jacobi symbol

$$\left(\frac{a}{n}\right) = -1$$

(3) Choose $x_0 \in_R \mathbb{Z}_n^*$, publish n, a, x_0 as his public key, keep p, q in secret.

100.3.2 Signing

To sign a message $M = (b_1, b_2, \dots, b_k) \in \{0, 1\}^k$, user A executes the following steps:

(1) Pick out the unique number among

$$\{x_0, -x_0, x_0a, -x_0a\}$$

denoted by c_1 such that it is a square residue (mod n).

(2) Solve

$$x^2 \equiv c_1 \pmod{n}$$

Denote two solutions ($\in (0, n/2)$) as x_1^*, x_1^{**} . They satisfy

$$\left(\frac{x_1^*}{n}\right) = -\left(\frac{x_1^{**}}{n}\right)$$

Without loss of generality, let $\left(\frac{x_1^*}{n}\right) = 1$, $\left(\frac{x_1^{**}}{n}\right) = -1$ and

$$x_1 = \begin{cases} x_1^*, & \text{if } b_1 = 0, \\ x_1^{**}, & \text{if } b_1 = 1. \end{cases}$$

- (3) $x_0 \leftarrow x_1$. Repeating this procedure recursively, user A obtains x_2, x_3, \dots, x_k . The signature is

$$(M, x_k)$$

100.3.3 Verification

Given a signature (M, x_k) , where $M = (b_1, b_2, \dots, b_k) \in \{0, 1\}^k$, and the public key (n, a, x_0) of user A, verifier executes the following steps:

- (1) Check

$$\frac{1 - \left(\frac{x_k}{n}\right)}{2} = b_k$$

If it does not hold, then he reject the signature and stop this protocol.

- (2) Let

$$c_k := x_k^2 \pmod{n}$$

Pick out the unique number in

$$\{c_k, a^{-1}c_k\}$$

denoted by x_{k-1} such that

$$\left(\frac{x_{k-1}}{n}\right) = 1 - 2b_{k-1}$$

(Note that $n = pq, p \equiv q \equiv 3 \pmod{4}$, $\left(\frac{c_k}{n}\right) = \left(\frac{-c_k}{n}\right)$)

- (3) $c_{k-1} \leftarrow x_{k-1}^2 \pmod{n}$. Repeating this procedure recursively, verifier obtains

$$x_{k-2}, x_{k-3}, \dots, x_1$$

- (4) Check

$$x_0 \in \{x_1^2, -x_1^2, a^{-1}x_1^2, -a^{-1}x_1^2\}$$

If it holds, verifier accepts the signature, otherwise rejects it.

100.4 An Existential Forgery to the GMY Strong Signature

100.4.1 The Basic Idea

The security of GMY signature has been widely accepted in the past decades. In this section, we will present an existential forgery attack against it, which shows that the GMY signature is insecure. Our strategy is to produce a value \bar{x}_k given three valid signatures

$$S(M^{(i)}) = (M^{(i)}, x_k^{(i)}) (i = 1, 2, 3)$$

of the same length k bits, then generate a message \bar{M} such that they consist of a valid signature

$$S(\bar{M}) = (\bar{M}, \bar{x}_k)$$

It is easy to see the relations between x_i and $x_{i-1} (i = k, \dots, 1)$, which can be described as follows:

$$\begin{aligned} x_{k-1} &\in \{x_k^2, a^{-1}x_k^2\} \\ x_{k-2} &\in \{x_{k-1}^2, a^{-1}x_{k-1}^2\} \\ &\dots\dots \\ x_1 &\in \{x_2^2, a^{-1}x_2^2\} \\ x_0 &\in \{x_1^2, -x_1^2, a^{-1}x_1^2, -a^{-1}x_1^2\} \end{aligned}$$

By the signing phase, we know

$$\begin{aligned} x_0 &\equiv (-1)^\alpha (a^{-1})^{\beta_1} x_1^2 \pmod{n} \\ x_1 &\equiv (a^{-1})^{\beta_2} x_2^2 \pmod{n} \\ &\dots\dots \\ x_{k-1} &\equiv (a^{-1})^{\beta_k} x_k^2 \pmod{n} \end{aligned}$$

where $\alpha, \beta_i \in \{0, 1\} (i = 1, \dots, k)$. Therefore,

$$x_0 \equiv (-1)^\alpha x_k^{2^k} (a^{-1})^\Delta \pmod{n}$$

where

$$\Delta = \beta_1 + \beta_2 \times 2 + \beta_3 \times 2^2 + \dots + \beta_k \times 2^{k-1}$$

Remark 1 By the verification, we know $\beta_i (i = 1, 2, \dots, k)$ are accessible to the adversary.

100.4.2 Attack

For convenience, we now suppose that we have three valid signatures $(M^{(i)}, x_k^{(i)}) (i = 1, 2, 3)$ signed by user A. Hence, we have

$$x_0 \equiv (-1)^{\lambda_1} \left(x_k^{(1)}\right)^{2^k} (a^{-1})^{\Delta^{(1)}} \pmod{n}$$

$$\Delta^{(1)} = \beta_1^{(1)} + \beta_2^{(1)} \times 2 + \beta_3^{(1)} \times 2^2 + \cdots + \beta_k^{(1)} \times 2^{k-1}$$

where $\lambda_1 \in \{0, 1\}$, $\beta_i^{(1)} \in \{0, 1\}$, $i = 1, 2, \dots, k$.

$$x_0 \equiv (-1)^{\lambda_2} \left(x_k^{(2)}\right)^{2^k} (a^{-1})^{\Delta^{(2)}} \pmod{n}$$

$$\Delta^{(2)} = \beta_1^{(2)} + \beta_2^{(2)} \times 2 + \beta_3^{(2)} \times 2^2 + \cdots + \beta_k^{(2)} \times 2^{k-1}$$

where $\lambda_2 \in \{0, 1\}$, $\beta_i^{(2)} \in \{0, 1\}$, $i = 1, 2, \dots, k$.

$$x_0 \equiv (-1)^{\lambda_3} \left(x_k^{(3)}\right)^{2^k} (a^{-1})^{\Delta^{(3)}} \pmod{n}$$

$$\Delta^{(3)} = \beta_1^{(3)} + \beta_2^{(3)} \times 2 + \beta_3^{(3)} \times 2^2 + \cdots + \beta_k^{(3)} \times 2^{k-1}$$

where $\lambda_3 \in \{0, 1\}$, $\beta_i^{(3)} \in \{0, 1\}$, $i = 1, 2, \dots, k$.

Since $\beta_i^{(j)}, \lambda_j (i = 1, 2, \dots, k; j = 1, 2, 3)$ are accessible to the adversary, without loss of generality, we assume that

$$\Delta^{(1)} > \Delta^{(2)} > \Delta^{(3)}$$

and set

$$\Delta' := \Delta^{(1)} - \Delta^{(2)} + \Delta^{(3)}$$

Since $\Delta^{(1)} > \Delta' > 0$, there exist

$$\bar{\beta}_i \in \{0, 1\}, i = 1, \dots, k$$

such that

$$\Delta' = \bar{\beta}_1 + \bar{\beta}_2 \times 2 + \cdots + \bar{\beta}_k \times 2^{k-1}$$

Set

$$\bar{x}_k := x_k^{(1)} \left(x_k^{(2)}\right)^{-1} x_k^{(3)} \pmod{n}$$

We have

$$x_0 \equiv (-1)^{\lambda_1 - \lambda_2 + \lambda_3} (\bar{x}_k)^{2^k} (a^{-1})^{\Delta'} \pmod{n}$$

Let

$$\bar{b}_k := \frac{1 - \frac{\bar{x}_k}{n}}{2}$$

$$\bar{b}_{k-i} := \frac{1 - \left(\frac{\bar{x}_k}{n} \right)^{2^i} (a^{-1})^{\sum_{j=1}^i \bar{\beta}_{k+1-j} \times 2^{i-j}}}{2}$$

where $i = 1, \dots, k - 1$. Set

$$\bar{M} := (\bar{b}_1, \bar{b}_2, \dots, \bar{b}_k).$$

Finally, we obtain a forged message \bar{M} such that signature (\bar{M}, \bar{x}_k) is valid.

100.4.3 Correctness

In fact, by the forgery procedure, we have

$$\begin{aligned} \bar{x}_{k-1} &= (\bar{x}_k)^2 (a^{-1})^{\bar{\beta}_k} \\ &\dots\dots\dots \\ \bar{x}_{k-i} &= (\bar{x}_k)^{2^i} (a^{-1})^{\sum_{j=1}^i \bar{\beta}_{k+1-j} \times 2^{i-j}} \\ &\dots\dots\dots \\ \bar{x}_1 &= (\bar{x}_k)^{2^{k-1}} (a^{-1})^{\bar{\beta}_k \times 2^{k-2} + \bar{\beta}_{k-1} \times 2^{k-3} + \dots + \bar{\beta}_2 \times 2} \end{aligned}$$

Hence,

$$(\bar{x}_1)^2 = (\bar{x}_k)^{2^k} (a^{-1})^{\bar{\beta}_k \times 2^{k-1} + \bar{\beta}_{k-1} \times 2^{k-2} + \dots + \bar{\beta}_2 \times 2}$$

Therefore,

$$x_0 \equiv (-1)^{\lambda_1 - \lambda_2 + \lambda_3} (\bar{x}_k)^{2^k} (a^{-1})^{\Delta'}$$

$$\in \{(\bar{x}_1)^2, -(\bar{x}_1)^2, (\bar{x}_1)^2 a^{-1}, -(\bar{x}_1)^2 a^{-1}\}$$

since $\Delta' = \bar{\beta}_1 + \bar{\beta}_2 \times 2 + \dots + \bar{\beta}_k \times 2^{k-1}$ and $\bar{\beta}_1 \in \{0, 1\}$. This means the forged message and signature pair, (\bar{M}, \bar{x}_k) , is valid.

100.5 Conclusion

In this chapter, we present an existential forgery attack against the GMY strong signature. It shows the adversary succeeds in forging the signature of one message, not necessarily of his choice. We think the technique developed in the paper will be helpful to investigate other signature schemes.

Acknowledgements The authors would like to thank the National Natural Science Foundation of China (Project 60873227), the Key Disciplines of Shanghai Municipality (S30104).

References

1. Diffie W, Hellman M (1976) New directions in cryptography. *IEEE Trans Inf Theory* 22:644–654
2. Rivest R, Shamir A, Adleman L (1978) A method for obtaining digital signatures and public-key cryptosystems. *Commun ACM* 21(2):120–126
3. Menezes A, Oorschot P, Vanstone S (1996) *Handbook of applied cryptography*. CRC Press, Ottawa
4. Cao ZJ, Liu ML (2008) Classification of signature-only signature models. *Sci China Ser F-Inf Sci* 51(8):1083–1095
5. Cao ZJ, Markowitch O (2009) Different authentication properties and a signcryption scheme revisited. In: 2009 International conference on computational intelligence and security. *IEEE CS*, pp 395–399
6. Dwork C, Naor M (1994) An efficient existentially unforgeable signature scheme and its applications. In: Desmedt Yvo (eds) *CRYPTO'94 LNCS vol 839*. pp 234–246 Springer, Heidelberg
7. Goldwasser S, Micali S, Rivest R (1988) A digital signature scheme secure against adaptive chosen-message attacks. *SIAM J Computing* 17(20):281–308
8. Goldwasser S, Micali S, Yao A (1983) Strong signature schemes. In: *Proceedings of 15th ACM symposium on theory of computing*, ACM, Boston, pp 431–439

Chapter 101

Improved Min-Sum Decoding Algorithm for Moderate Length Low Density Parity Check Codes

Waheed Ullah, Jiangtao and Yang FengFan

Abstract In this chapter, a new technique to improve the min-sum decoding algorithm for the low density parity check (LDPC) code has been proposed. This technique is based on the magnitude overestimation correction of the variable message by using two normalized factors in all iterations. The variable message is modified with a normalized factor when there is a sign change and with another normalized factor when there is no sign change during any two consecutive iterations. In this way, the algorithm gives a more optimum approximation to the min-sum decoding algorithm. This new technique outperforms for medium and short length codes and for small number of iterations, which make it suitable for practical applications and hardware implementation.

Keywords LDPC codes · Sum product algorithm · Min-sum · Belief propagation · Parity check matrix · Tanner graph

101.1 Introduction

Low density parity check (LDPC) codes, also known as Gallager codes, are a type of linear block codes, first proposed by Gallagar [1] and were scarcely considered in the three decades that followed due to its computational complexity and the limited computational ability of receivers at that time. LDPC was reinvented by

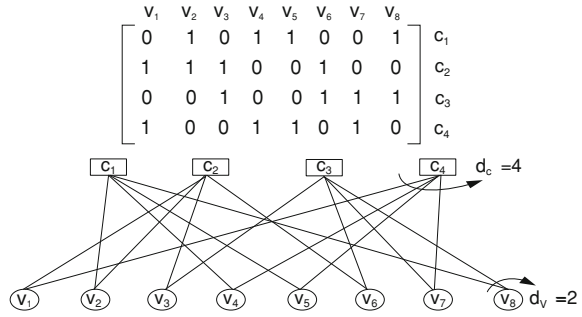
W. Ullah (✉) · Jiangtao · Y. FengFan
College of Electronics and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, 210016, China
e-mail: uet_waheed@yahoo.com

Mackay and Neal [2] and have taken considerable attention recently due to its powerful error correcting capabilities and their Shannon limit performance [3–5], with belief propagation decoding algorithm. The name comes from the characteristic of their parity check matrix [6] which contains only a few ones (1s) as compared to the number of zeros (0s). The Belief Propagation (BP) or the sum-product decoding algorithm (SPA) is the best performing algorithm with very high computation complexity and also has dependence on the noise variance. The other popular iterative decoding algorithms which offer extremely low hardware complexity with little performance degradation, is the min-sum algorithm (MSA) [7]. The min-sum algorithm reduces the decoding complexity and is free of noise variance as well. Based on these properties of MSA, several approaches have been made to keep the performance close to SPA while still the decoding complexity is less. Different methods are used to bring the simplified form of the algorithm close in performance to the original BP or sum product algorithm. Some well know approaches are the normalized min-sum (Normalized MSA) and the offset min-sum (Offset MSA) [8]. Density evolution [9, 10], is used to analyze the performance of these decoding algorithms for determining the optimum values of the key parameters either as normalized or offset values. Check message is modified during the iteration process to avoid it from over estimation which brings the min-sum algorithm close in performance to the standard SPA [11] and make them suitable for practical applications and hardware implementation [12]. Due to min-sum, which is a reduced complexity algorithm, LDPC has gained popularity in a wide area of practical applications like local area networks, satellite and inter-satellite communications, deep sea communications etc. Some practical considerations and implementations have been have been proposed in [13, 14].

The choice of the scaling factor in the normalized and offset types of min-sum algorithm is not fixed. The suitable parameters can only be chosen by simulation prior to the implementation. This chapter is focused on the performance improvement to the min-sum decoding algorithm [15] by using two hardware friendly scaling factors. The performance of min-sum decoding algorithm is improved further by adding one additional scaling factor. As in normalized min-sum [9], there is only one scaling factor used which is found by exhaustive search algorithm for better performance.

The proposed improved min-sum algorithm (IMSA) modifies the variable message during the two consecutive iterations. When the signs of the present and previous messages are different then it is modified with one scaling factor, otherwise the message is altered by another scaling factor. The results show that the proposed IMSA is better than normalized MSA and even novel modified min-sum algorithm both in performance and complexity for the normalization factors. The results are also compared with SPA to further validate the significant improvement in performance.

Fig. 101.1 Tanner graph representation of parity check matrix



101.2 Representation of LDPC Codes

101.2.1 Algebraic Representation

LDPC code can be denoted in general as (N, d_v, d_c) , where N is the length of the code equal to the number of columns in the parity check matrix, d_c is the number of ones (1s) in a column of a parity check matrix, d_v is the number of ones (1s) in a row of a parity check matrix. LDPC codes can be regular and irregular. If the number of ones (1s) in each row and column of a parity check matrix are the same, it is called regular; and if the number of ones (1s) in each row and column are not the same, it is called irregular. For a regular code, following condition applies

$$M.d_c = N.d_v \tag{101.1}$$

where M and N are the rows and columns of a parity check matrix respectively. The code is valid only if $H \cdot \text{code}^T = 0$, where H is the sparse parity check matrix.

101.2.2 Tanner Graph Representation

The sparse parity check matrix [6] is best represented by a bipartite graphs know as Tanner graphs [14]. Each row of the parity check matrix represents the variable node and each column represents the check node. The one in each row or column shows the connectivity between variable and check nodes. The set of bit nodes connecting to check node m is denoted by $N(m) = \{n|h_{mn} = 1\}$ and the set of check nodes connecting to bit node n is by $M(m) = \{m|h_{mn} = 1\}$. A typical Tanner graph is shown in the Fig. 101.1. This graph is for $(6, 2, 4)$ regular LDPC code.

101.2.3 LDPC Min-Sum Decoding Algorithm

Let $X = \{x_1, x_2, \dots, x_n\}$ be the transmitted code after binary phase shift keying (BPSK). It is transmitted over an additive white Gaussian noisy (AWGN) channel.

$$Y = X + n \tag{101.2}$$

where n is an AWGN and $Y = \{y_1, y_2, \dots, y_n\}$.

Now LDPC min-sum decoding [7, 10, 13], can be stated in the following steps for a parity check matrix H_{mn} , where m is the number of rows and n is the number of columns.

Step 1 Initialization: Set $L_n = Y$ as initial log likelihood ratio (LLR) and for each $(m, n) \in \{(m, n) | h_{mn} = 1\}$

$$V_{mn}^0 = L_n \tag{101.3}$$

Set the maximum number of iterations (I_{max}) as $i = 0$ to I_{max} .

Step 2 Row processing: bit nodes to check nodes

For $m = 0$ to $M-1$, update C_{mn}^i for each $n \in N(m)$

$$C_{mn}^i = \prod_{\substack{n' \in N(m) \\ n' \neq n}} \text{sign}(V_{mn'}^{i-1}) \cdot \min_{\substack{n' \in N(m) \\ n' \neq n}} |V_{mn'}^{i-1}| \tag{101.4}$$

Step 3 Column processing: check nodes to bit nodes

For $n = 0$ to $N-1$, update

$$\tilde{L}_n^i = L_n + \sum_{m \in M(n)} C_{mn}^i \tag{101.5}$$

Now updating V_{mn}^i for each $m \in M(n)$

$$V_{mn}^i = L_n^i - C_{mn}^i \tag{101.6}$$

Step 4 Hard Decision:

$$\hat{x}_n = \begin{cases} 0, & \text{for } \tilde{L}_n > 0 \\ 1, & \text{for } \tilde{L}_n \leq 0 \end{cases} \tag{101.7}$$

Step 5 Stop condition: If the parity check equation is satisfied i.e.

$$H \cdot (\hat{x}_1 \hat{x}_2 \dots \hat{x}_n)^T = 0 \tag{101.8}$$

Or maximum iteration (I_{max}) is reached then terminate the decoding or otherwise $i = i + 1$ and go back to step 2.

The message passing between check nodes and variable nodes in Steps 2 and 3 can also be represented in a graphical way as shown in Figs. 101.2 and 101.3.

Fig. 101.2 Step 2. Check node update

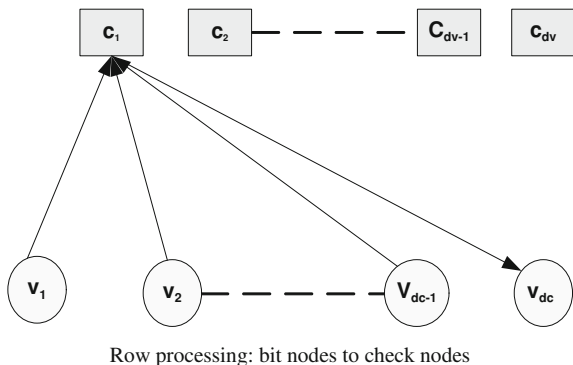
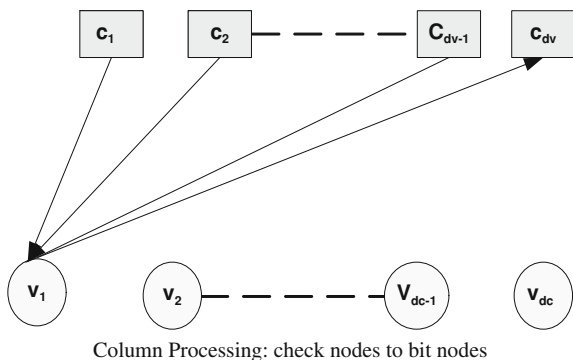


Fig. 101.3 Step 3. Variable node update



101.3 Modified Min-Sum Decoding Algorithm

101.3.1 Variable Message Update Conditions

Min sum algorithm (MSA) greatly reduces the complexity of SPA at the cost of performance degradation and the bit error ratio is significantly higher than SPA. All efforts are made to make the MSA close to SPA in performance while keep it simple. The Offset and Normalized MSA alter the inaccurate magnitude for the check node update calculated in step 2. The offset and normalized min-sum algorithms increase the hardware complexity very less but improve the performance significantly for MSA and bring it close to SPA.

The modified min-sum algorithm (MMSA) [15] takes into account the following two conditions for two consecutive iterations:

- (i) The sign of the present and previous variable messages are the same then the magnitude increase is comparatively small.
- (ii) The sign of the present variable message and the previous variable messages are different. In this case, the magnitude increase is large and the variable message needs to be corrected to avoid overestimation.

Based on the these two fundamental facts, the newly proposed technique modifies the variable message at the check node processing with two different scaling factors either at a sign change or at no sign change. The choice of the scaling factors is obviously dependent on the magnitude increase and its hardware implementation complexity. The range for both the scaling factors s is such that; $0 < s < 1$. For the first condition, as the magnitude increase is less, the scaling factor is chosen in the range 0.5–0.9, and for the second, as the magnitude increase is large so it is modified by a factor in the range 0.1–0.5.

101.3.2 Method for Variable Message Correction

In the step 3, Eq. 101.6, the variable message is calculated at the i th iteration but before using for update, it is stored temporarily as $V_{mn}^{i,\text{tmp}}$.

$$V_{mn}^{i,\text{tmp}} = L_n^i - C_{mn}^i \quad (101.9)$$

Next the signs of the present message $V_{mn}^{i,\text{tmp}}$ and previous message V_{mn}^{i-1} are compared as; If

$$\text{sign}(V_{mn}^{i,\text{tmp}}) == \text{sign}(V_{mn}^{i-1}) \quad (101.10)$$

Then update the message as;

$$V_{mn}^i = sf_1(V_{mn}^{i,\text{tmp}}) \quad (101.11)$$

Else if

$$\text{sign}(V_{mn}^{i,\text{tmp}}) \neq \text{sign}(V_{mn}^{i-1}) \quad (101.12)$$

Then update the message as;

$$V_{mn}^i = sf_2(V_{mn}^{i,\text{tmp}}) \quad (101.13)$$

The scaling factors sf_1 and sf_2 are chosen in such a way that these could be conveniently implemented in hardware, and at the same time provide good approximation to the error performance. Now if the signs are different then the change in magnitude is large and is modified with small factor to reduce the overestimation effect. The scaling factors set for the simulation are $sf_1 = 0.5$ and $sf_2 = 0.25$. Now Eqs. 101.11 and 101.13 can be re-written as;

$$V_{mn}^i = 0.5(V_{mn}^{i,\text{tmp}}) \quad (101.11a)$$

$$V_{mn}^i = 0.25(V_{mn}^{i,\text{tmp}}) \quad (101.13a)$$

This brings further improvement to the MSA in both lower and upper region of SNR by using two scaling factors. The complexity is very less both in hardware and software but the performance achieved is far better.

101.3.3 Hardware Implementation and Complexity Analysis

We see that both scaling factors are easily implemented in hardware as shift registers. The scaling factors chosen are the division by 2 and 4 which are simply implemented in hardware as shift registers as data is shifted by one and two respectively. The hardware complexity is less than the min-sum algorithm in Haiyang et al. [15] as there is no adder needed. The sign comparator decides which input to select for assigning to the current message V_{mm}^i through multiplexer (Mux) unit. The shift register is fast and easy to implement. So the hardware complexity does not increase reasonably while the performance achievement is better. If we compare the hardware complexity with normalized min sum and offset min-sum decoding algorithms, then the complexity is increased just by a sign comparator which is comparatively very less but the BER advantage is significant

Also this is a reliable way for updating the variable message. Instead of uniform modification to all the variable messages, it gives the flexibility to update the messages in two ways which gives the advantage of better performance.

101.4 Simulation Results

Two types of codes are selected for the validation of the performance results for the proposed improved MSA through computer simulations. Regular medium and short length LDPC codes (1024, 512) and (648, 324) are chosen for all the decoding algorithms to evaluate the performance improvement. After encoding process and binary phase shift keying (BPSK) modulation, the desired code is passed through AWGN for a range of signal to noise ratio (SNR) values. The maximum allowable number of iteration is kept as 10. In the Fig. 101.4, we see clearly that the improved min-sum decoding algorithm outperform than MMSA, and even from SPA for the selected length of codes. The outperformance of improved min sum decoding algorithm than standard sum product algorithms is due to the fact that SPA depends on large sparse parity check matrix while the parity check matrix selected here is moderate length.

101.5 Comparison and Analysis

Simulation is performed for validating the performance of the proposed technique for medium and short length codes which best suits for most of the practical applications. The performance is tested at 10 maximum number of iteration for all types of the decoding algorithms. The results obtained are compared with MMSA and SPA. Also, simulations are carried out for two different code lengths. We see that for the (1024, 512) LPDC code at 4.5db SNR in Fig. 101.5, the BER for IMSA, MMSA and SPA are 0.000606, 0.001553 and 0.002029 respectively.

Fig. 101.4 Implementation for variable message correction

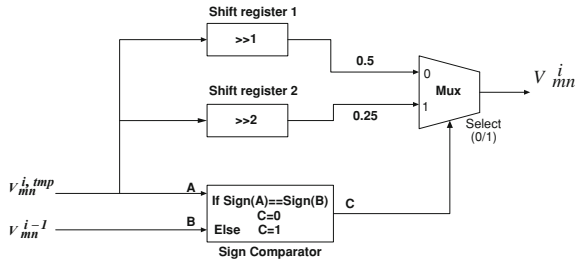


Fig. 101.5 BER performance comparison (1024, 512) LDPC code, maximum number of decoding iterations = 10

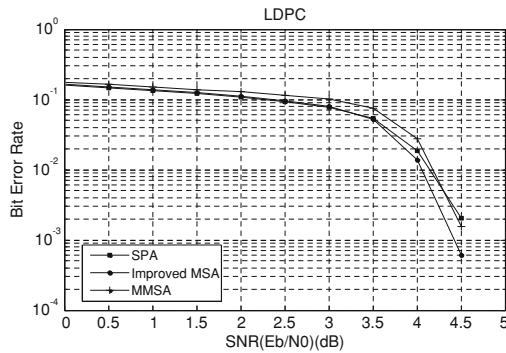
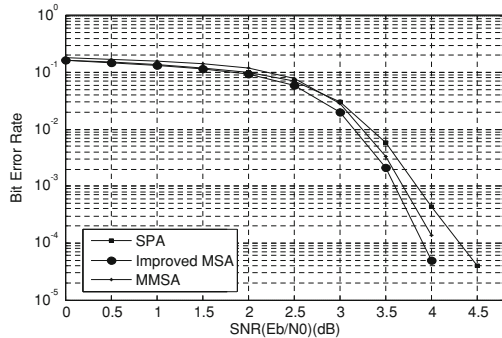


Fig. 101.6 BER performance comparison (648, 324) LDPC code, maximum number of decoding iteration = 10



Similarly for the (648, 324) LDPC code at SNR 4 in Fig. 101.6, the BER for IMSA, MMSA, SPA is 0.000050, 0.000137 and 0.000437, respectively. In both types of the codes, it has been observed that the proposed technique yields better results than other min-sum decoding algorithms.

101.6 Conclusion

In this chapter a totally new approach is used with some of the existing methods to improve the performance of the min-sum decoding algorithm for medium and short length codes which can be easily applied to practical systems. The proposed

method is an efficient way for modifying the variable message during the vertical process in all iterations and hence the overestimation is corrected optimally. Due to its two way normalization to correct the variable message magnitude, this algorithm has an inherent capability of improved performance. Furthermore, it is simple to implement, and the hardware cost is less.

Acknowledgments Supported by Science and Technology on Avionics Integration Laboratory and National Foundation of Aeronautical Science and Research under contract No. 20105552031.

References

1. Gallager RG (1963) Low-density parity-check code. MIT Press, Cambridge
2. Mackay D, Neal R (1996) Near Shannon limit performance of low density parity check codes. *Electron Lett* 32(18):1645–1646
3. Chung S, Forney GD, Richardson JJ, Urbanke R (2001) On the design of low-density parity-check codes within 0.0045 db of the Shannon limit. *IEEE Commun Lett* 5:58–60
4. Richardson T, Shokrollahi MA, Urbanke RL (2001) Design of capacity-approaching irregular low-density parity-check codes. *IEEE Trans Inf Theory* 47:619–637
5. Richardson T, Urbanke R (2000) The capacity of low-density parity check codes under message-passing decoding. *IEEE Trans Inf Theory* 47:599–618
6. Mackay D (1999) Good error correcting codes on very sparse matrices. *IEEE Trans Inf Theory* 45(2):399–431
7. Fossorier MPC, Mihaljevic M, Imai H (1999) Reduced complexity iterative decoding of low density parity check codes based on belief propagation. *IEEE Trans Commun* 47(5):673–680
8. Chen J, Dholakia A, Eleftheriou E et al (2005) Reduced-complexity decoding of LDPC codes. *IEEE Trans Commun* 53(8):1288–1299
9. Chen J, Fossorier MPC (2002) Density evolution for two improved BP-based decoding algorithms of LDPC codes. *IEEE Commun Lett* 6(5):208–210
10. Heo J (2003) Analysis of scaling soft information on low density parity check code. *Electron Lett* 55:219–221
11. Chen J, Fossorier MPC (2002) Near optimum universal belief propagation based decoding of low-density parity check codes. *IEEE Trans Commun* 50(3):406–414
12. Zhao J, Zarkeshvari F, Banihashemi AH (2005) On implementation of min-sum algorithm and its modifications for decoding low-density parity-check (LDPC) codes. *IEEE Trans Commun* 53:549
13. Chandrasetty VA, Aziz SM (2010) FPGA implementation of high performance LDPC decoder using modified 2-bit min-sum algorithm. Proceedings of the 2nd international conference on computer research and development, Kuala Lumpur, 7–10 May 2010, pp 881–885
14. Aziz SM, Pham MD (2010) Implementation of low density parity check decoders using a new high level design methodology. *J Comput* 5(1):0234–0237
15. Hai-yang L, Wen-ze QU, Bin L, Jiang-peng L, Shi-dong L, Jie C (2010) Novel modified min-sum decoding algorithm for low-density parity-check codes, www.sciencedirect.com. *J China Univ Posts Telecommun* 17:1–5

Chapter 102

Cryptanalysis of an Authentication Protocol for Session Initiation Protocol

Qi Xie, Mengjie Bao and Haijun Lu

Abstract The session initiation protocol (SIP) is a challenge-response based authentication protocol which is used in 3G mobile networks. In 2010, Chen et al. proposed an authentication protocol for SIP using elliptic curves cryptography. However, this paper show that the Chen et al. authentication scheme is vulnerable to off-line password guessing attack, and does not provide perfect forward secrecy. To overcome their weaknesses, a security-enhanced authentication scheme for SIP is proposed, which not only defends against the off-line password guessing attack, but also supports perfect forward secrecy.

Keywords Authentication · Key agreement · Session initiation protocol · Password guessing attack · Perfect forward secrecy

102.1 Introduction

The session initiation protocol (SIP) is an application layer control protocol for creating, modifying, and terminating multimedia sessions between multiple users in the IP based telephony environments, which was proposed by internet engineering task force network working group in 1999 [1, 2]. Because SIP is based on the challenge-response mechanism protocol, therefore, the original SIP cannot

Q. Xie (✉) · M. Bao · H. Lu
School of Information Science and Engineering, Hangzhou Normal University,
Hangzhou 310036, People's Republic of China
e-mail: qixie68@yahoo.com.cn

provide security at an acceptable level. For example, the original SIP cannot resist the off-line guessing attack and the spoofing attack.

In order to overcome the weaknesses of the original SIP, many kinds of authentication schemes for SIP have been proposed. In 2005, Yang et al. [3] proposed an authentication scheme for SIP based on the Diffie–Hellman key exchange algorithm. Unfortunately, it cannot resist the replay attack and off-line password guessing attack. Based on elliptic curve cryptosystem (ECC), Durlanik et al. [4] proposed a new SIP scheme, which is more efficient than Yang et al.’s scheme. In 2009, Wu et al. [5] proposed a provably secure SIP scheme in Canetti–Krawczyk (CK) model. They claimed the scheme can resist replay attacks, off-line password guessing attacks, man-in-the-middle attacks, and server spoofing attacks. However, their scheme is still vulnerable to forgery attack if an adversary gets the secret information from smart card. In 2006, Ring et al. [6] proposed a secure SIP authentication scheme based on identity-based cryptography (IBC), but it is not efficient due to heavy computation. In 2008, Wang et al. [7] proposed a new authentication scheme for SIP based on certificateless public key cryptography (CL-PKC), but it still remains the heavy computation. In 2010, Liao et al. [8] proposed a new authentication scheme for SIP using self-certified public keys on elliptic curves. In order to solve the heavy computation problem, Chen et al. [9] proposed a scheme based on ECC using smart cards, and Tsai [10] proposed an efficient authentication scheme for SIP based on hash function. Yoon et al. [11] pointed out that Tsai’s scheme cannot resist off-line password guessing attacks, Denning-Sacco attack, stolen-verifier attacks, and does not provide forward security. Then, they proposed a new scheme with perfect forward security. However, Yoon et al.’s scheme is still vulnerable to stolen-verifier attack, and may also suffer from off-line password guessing attack.

In this chapter, we show that Chen et al.’s scheme [9] cannot resist off-line password guessing attacks and does not provide forward secrecy. To overcome the weaknesses, we propose an improvement scheme.

102.2 Review of Chen et al.’s Scheme

Chen et al.’s scheme consists of two phases: setup phase, mutual authentication with key agreement phase. G_p denotes a cyclic addition group of p , $E(F_p)$ is an elliptic curve over a finite field F_p , $h()$, $h'()$ and $H()$ denote one-way hash functions, where $h()$ and $h'()$ take as input an arbitrary block of data or points on elliptic curve and returns a bit string, $H()$ takes as input an arbitrary block of data or points on elliptic curve and returns a point on elliptic curve.

102.2.1 Setup Phase

In this phase, user U_A and the remote server S perform the following steps:

- (1) U_A chooses his/her ID_A , password pw_A and randomly chooses a large number b , computes

$$PWB = h(pw_A \oplus b). \quad (102.1)$$

Then, U_A submits ID_A and PWB to S .

- (2) S first selects a base point P with the order n over $E(F_p)$, computes

$$K_{ID_A} = q_s \cdot H(ID_A) \in G_p, \quad (102.2)$$

where $(q_s, Q_s = q_s \cdot P)$ is S 's private/public key pair. Then S computes

$$B_A = h(ID_A \oplus PWB), \quad (102.3)$$

$$W_A = h(PWB || ID_A) \oplus K_{ID_A}, \quad (102.4)$$

stores $[B_A, W_A, h(), H()]$ into a smart card and sends the smart card to U_A over a secure channel.

- (3) Upon receiving the smart card, U_A stores the random number b into the smart card. So the smart card contains $[B_A, W_A, b, h(), H()]$.

102.2.2 Mutual Authentication with Key Agreement Phase

Assume that U_A is about to logon to the remote server S , they perform the following steps:

- (1) U_A enters his/her ID_A and pw_A , the smart card calculates PWB and B_A' by Eqs. 102.1 and 102.3, and checks if $B_A' = B_A$. If it holds, the smart card calculates

$$Q = h(PWB || ID_A), \quad (102.5)$$

$$K_{ID_A} = W_A \oplus Q. \quad (102.6)$$

Then chooses a random point $R_A = (x_A, y_A) \in E(F_p)$, computes $t_1 = h(T_1)$,

$$M_A = R_A + t_1 \cdot K_{ID_A}, \quad (102.7)$$

$$\bar{R}_A = x_A \cdot P, \quad (102.8)$$

at the timestamp T_1 . Finally, the smart card sends REQUEST message $m_1 = [T_1, ID_A, M_A, \bar{R}_A]$ to S .

(2) After receiving m_1 , S computes $t_1 = h(T_1)$,

$$Q_{ID_A} = H(ID_A) = (x_Q, y_Q), \quad (102.9)$$

$$R'_A = M_A - q_s \cdot t_1 \cdot Q_{ID_A} = (x'_A, y'_A), \quad (102.10)$$

and verifies if $\overline{R_A} = x'_A \cdot P$. If so, S chooses a random point $R_s = (x_s, y_s) \in E(F_p)$, computes $t_2 = h(T_2)$, $M_s = R_s + t_2 \cdot q_s \cdot Q_{ID_A}$, session key

$$k = h(x_Q, x'_A, x_s) \quad (102.11)$$

and $M_k = (k + x_s) \cdot P$ at the timestamp T_2 . Finally, S sends CHALLENGE message $m_2 = [T_2, M_s, M_k, realm]$ to U_A .

(3) After receiving m_2 , U_A computes Q_{ID_A} from Eq. 102.9, $t_2 = h(T_2)$,

$$R'_s = M_s - t_2 \cdot K_{ID_A} = (x'_s, y'_s), \quad (102.12)$$

$$k' = h(x_Q, x_A, x'_s). \quad (102.13)$$

$$M'_k = (k' + x'_s) \cdot P \quad (102.14)$$

and verifies if $M_k = M'_k$. If it holds, U_A computes $response = h(k || realm || ID_A)$ and sends the RESPONSE message $\{realm, ID_A, response\}$ to the server.

(4) After receiving RESPONSE message, S computes and verifies whether $h(k || realm || ID_A) = response$. If so, U_A is authenticated, and they share k .

102.3 Weaknesses of Chen et al.'s Scheme

In this section, we show that Chen et al.'s scheme suffers from off-line password guessing attack, and does not support Perfect Forward Secrecy (PFS).

102.3.1 Off-line Password Guessing Attack and Forgery Attack

Since the smart card contains $[B_A, W_A, b, h(), H()]$, if an adversary obtain the information from user's smart card, then he can guess the password $pw_{A'}$ and computes $PWB' = h(pw_{A'} \oplus b)$, $B_{A'} = h(ID_A \oplus PWB')$, and checks if $B_A = B_{A'}$. If so, then the guessed $pw_{A'}$ is right. Otherwise, he chooses another password and tries again. Since user's password is supposed to be short and the password space is small, so this attack is feasible.

After an adversary obtains the user's password, he can get K_{ID_A} from Eq. 102.4, and can impersonate the user pass through the authentication of the server.

102.3.2 Perfect Forward Secrecy

Since the session key is $k = h(x_Q, x_A, x_s)$. Therefore, if an adversary obtains the server's secret key q_s , then he/she can get R_A , R_s and Q_{ID_A} from Eqs. 102.9, 102.10 and 102.12, as ID_A , M_A , T_1 and T_2 are communication messages in insecure networks. Thus, he/she can get the session key.

102.3.3 Other Weakness

As the authentication mechanism of SIP is similar to the HTTP authentication, which is a challenge-response based authentication protocol. So the user authenticates the server in challenge step, while the server authenticates the user in response step. In Chen et al.'s scheme, the server authenticates the user both in request step and in response step. Therefore, it is unnecessary that the server authenticates the user in request step.

102.4 The Proposed Scheme for SIP

102.4.1 Setup Phase

In this phase, user U_A and the remote server S perform the following steps:

- (1) U_A chooses his/her ID_A , password pw_A and randomly chooses a large number b , computes

$$PWB = H(pw_A \oplus b \oplus ID_A). \quad (102.15)$$

Then, U_A submits ID_A and PWB to S .

- (2) S first selects a base point P with the order n over $E(F_p)$, computes

$$K_{ID_A} = q_s \cdot H(ID_A) \in G_p. \quad (102.16)$$

where $(q_s, Q_s = q_s \cdot P)$ is S 's private/public key pair. Then S computes

$$W_A = PWB \oplus K_{ID_A}, \quad (102.17)$$

stores $[ID_A, W_A, h(), h'(), H()]$ into a smart card and sends the smart card to U_A over a secure channel.

- (3) Upon receiving the smart card, U_A stores the random number b and ID_A into the smart card. So the smart card contains $[ID_A, W_A, b, h(), h'(), H()]$.

102.4.2 Mutual Authentication with Key Agreement Phase

Assume that U_A is about to logon to the remote server S , they perform the following steps:

- (1) U_A enters his/her password pw_A , the smart card calculates PWB and K_{ID_A} from Eqs. 102.16 and 102.17. Then chooses a nonce c , computes $t_1 = h(T_1)$,

$$R_A = c \cdot P, \quad (102.18)$$

$$M_A = R_A + t_1 \cdot K_{ID_A}, \quad (102.19)$$

at the timestamp T_1 . Finally, the smart card sends REQUEST message $m_1 = [ID_A, T_1, M_A]$ to S .

- (2) After receiving m_1 , S computes $t_1 = h(T_1)$,

$$Q_{ID_A} = H(ID_A), \quad (102.20)$$

$$R'_A = M_A - q_s \cdot t_1 \cdot Q_{ID_A}, \quad (102.21)$$

chooses a nonce d , computes $t_2 = h(T_2)$, where T_2 is the timestamp, and

$$R_s = d \cdot P, \quad (102.22)$$

$$R = d \cdot R_A = c \cdot d \cdot P, \quad (102.23)$$

$$M_s = R_s + t_2 \cdot q_s \cdot Q_{ID_A}, \quad (102.24)$$

$$k = h'(R || R'_A || R_s), \quad (102.25)$$

$$M_k = h'(k || R_s). \quad (102.26)$$

Finally, S sends CHALLENGE message $m_2 = [T_2, M_s, M_k, realm]$ to U_A .

- (3) After receiving m_2 , U_A computes Q_{ID_A} from Equation (102.20), $t_2 = h(T_2)$,

$$R'_s = M_s - t_2 \times K_{ID_A}, \quad (102.27)$$

$$k = h'(R || R_A || R'_s) \quad (102.28)$$

$$M'_k = h'(k || R'_s) \quad (102.29)$$

and verifies if $M_k = M'_k$. If it holds, U_A computes $response = h(k + 1 || realm || ID_A)$ and sends the RESPONSE message $\{realm, ID_A, response\}$ to the server.

- (4) S computes $response' = h(k + 1 || realm || ID_A)$ and verifies whether $response' = response$. If so, U_A is authenticated by S , they share the session key $k = h'(R || R_A || R_s)$.

102.5 Security Analysis

102.5.1 Password Guessing Attack and Smart Card Lost Attack

If an adversary (may be a legal user) get all information from the user U_A 's smart card, and wants to mount password guessing attack. Obviously, he can get all communication messages m_1 and m_2 . However, he cannot get the right password, because M_A and M_s include unknown R_A and R_s , and $R = dR_A = cR_s = cdP$, the adversary cannot compute R from R_A and R_s due to the intractability of the Computational Diffie–Hellman (CDH) problem.

102.5.2 Replay Attack and Forgery Attack

In the proposed scheme, we use nonce and timestamp to against the replay attack. If an adversary impersonates the user and replays the REQUEST message m_1 , it can be detected by the server according to the timestamp T_1 . If an adversary use new timestamp T_1' , computes $M'_A = t_1^{-1}R_A + t'_1 \cdot K_{ID_A}$ and send the REQUEST message $m_1' = [ID_A, T_1', M'_A]$ to the server. The server can still send the right CHALLENGE message $m_2 = [T_2, M_s, M_k, realm]$ to the adversary, but the adversary is unable to verify if the CHALLENGE message is right or not due to unknown $R = dt_1^{-1}R_A = cdt_1^{-1}P$, and cannot get the right session key, so his RESPONSE message cannot pass through the authentication of the server.

102.5.3 Denning–Sacco Attack

If an adversary gets the session key $k = h'(R^{\parallel}R_A^{\parallel}R_s)$, he may want to obtain the user's password or the server's secret key. However, an adversary cannot get R_A or R_s from k which is protected by a hash function, so he is not able to get he user's password or the server's secret key from M_A and M_s .

102.5.4 Known-Key Security and Session Key Security

Since session key is $k = h'(R^{\parallel}R_A^{\parallel}R_s)$, where $R = dR_A = cR_s = cdP$. So an adversary cannot know k without knowing random nonces c or d , and it is infeasible if he wants to get cdP from cP and dP under the assumption of CDH problem.

102.5.5 Perfect Forward Security

If an adversary knows the server's secret key and the user's password, then he can get R_A and R_s , but he cannot know cdP from R_A and R_s since it is an elliptic curve Diffie–Hellman problem. Therefore, he is not able to get the session key.

102.6 Conclusions

In this chapter, we have shown that Chen et al.'s authenticated key agreement scheme for SIP is vulnerable to off-line password guessing attack and does not support Perfect Forward Secrecy. To overcome the weaknesses in their scheme, we proposed a security enhancement scheme for SIP.

Acknowledgment This research was supported by National Natural Science Foundation of China (No.61070153).

References

1. Rosenberg J, Schulzrinne H, Camarillo G, Johnston A, Peterson J, Sparks R, Handley M, Schooler E (2002) SIP session initiation protocol. RFC 3261. Internet Engineering Task Force
2. Franks J, Hallam-Baker P, Hostetler J, Lawrence S, Leach P, Luotonen A, Stewart L (1999) HTTP authentication: basic and digest access authentication. IETF RFC 2617
3. Yang CC, Wang RC, Liu WT (2005) Secure authentication scheme for session initiation protocol. *Comput Secur* 24:381–386
4. Durlanik A, Sogukpinar I (2005) SIP authentication scheme using ECDH. *World Enformatika Soc Trans Eng Comput Technol* 8:350–353
5. Wu L, Zhang Y, Wang F (2009) A new provably secure authentication and key agreement protocol for SIP using ECC. *Comput Stand Interfaces* 31(2):286–291
6. Ring J, Choo K, Foo E, Looi MA (2006) New authentication mechanism and key agreement protocol for SIP using identity based cryptography. In: *Proceedings of AusCert R&D Stream*, pp 61–72
7. Wang F, Zhang Y (2008) A new provably secure authentication and key agreement mechanism for SIP using certificateless public-key cryptography. *Comput Commun* 31:2142–2149
8. Liao Y, Wang S (2010) A new secure password authenticated key agreement scheme for SIP using self-certified public keys on elliptic curves. *Comput Commun* 33:372–380
9. Chen T, Yeh H, Liu P (2010) A secured authentication protocol for SIP using elliptic curves cryptography. *PGCN2010, Part1, CCIS119*. Springer, Berlin, pp 46–55
10. Tsai JL (2009) Efficient nonce-based authentication scheme for session initiation protocol. *Int J Netw Security* 8(3):312–316
11. Yoon E, Shin Y, Jeon I, Yoo K (2010) Robust mutual authentication with a key agreement scheme for the session initiation protocol. *IETE Tech Rev* 27(3):203–213

Chapter 103

A Cognitive Model in Biomimetic Pattern Recognition and Its Applications

Xiao Xiao, Xianbao Wang, Sunyuan Shen and Shoujue Wang

Abstract A cognitive model in Biomimetic Pattern is proposed to deal with the issue of pattern recognition. Its mathematical basis is placed on topological analysis of sample set in high dimensional feature space. The fundamental idea of this model is based on the fact of the continuity in the feature space of any one of the certain kinds of samples. The cognitive model and its neural network conformation algorithm are given in this chapter. The algorithm is applied to face database and speech database and its effectiveness is proved by the experiments.

Keywords Biomimetic informatics · High dimensional space · Neural networks · Pattern recognition

103.1 Introduction

Pattern Recognition has been developed for dozens of years. Classification is the main problem of pattern recognition. In this paper, pattern recognition is based on “pattern cognition” instead of “pattern classification”. First the traditional two methods of pattern recognition are listed in the follow.

X. Xiao (✉) · X. Wang · S. Shen · S. Wang
Information College, Zhejiang University of Technology, Hangzhou 310014, China
e-mail: xiaoxiao800412@163.com

S. Wang
Institute of Semiconductors, Chinese Academy of Science, Beijing 100083, China

103.1.1 Face Recognition

By extracting the facial features face recognition technology is used to recognize the identity. Face recognition can be used for identification of criminals and the public security system, such as driver's license check, passport check, customs control systems, etc. In recent years, it has become the research focus of computer vision and pattern recognition. However, the facial expression changes and light changes make the face recognition difficult.

In the development of face recognition technology, the principal component analysis (PCA) [1], linear discriminate analysis (LDA) [2], locality preserving projections (LPP) [3] are proposed, and they all achieves good results in face recognition.

103.1.2 Keywords Recognition

Keyword Spotting, usually called keyword confirmation or word confirmation, aims at finding the special words from the speaker's continuous speech, which usually contains other words or other sounds such as coughing, breathing, razzing, background noise and transmit noise etc. [4].

Early keyword spotting systems were all based on Dynamic Time Wrap (DTW) while current systems are usually based on Hidden Markov Model (HMM) or its improved methods. These methods commonly need endpoint detection and division of the continuous speech. Methods without endpoint detection and division were proposed in some literatures. However, too many constraints have influenced their application. The correct recognition rate of traditional methods excessively depends on the endpoint detection and division, so they can't get results with good robustness [5].

103.2 Cognitive Algorithm

The traditional method of pattern recognition is to find the best classification hyper-plane which will be trapped when it is used to deal with high-dimensional data set with low-dimensional geometry structure. How to follow the external discipline of the data distribution is a very important problem. The follow cognitive model can solve the problem [6–8].

103.2.1 Cognitive Model

We uses $Z = f [P(X, Th)]$ neuron shortly called Z neuron to cover the training samples (namely construct the training net), where X is the input vector, Th represents the threshold value, the excitation function f is

$$f(X) = \begin{cases} 1, X \in P \\ -1, X \notin P \end{cases}$$

While P is a closed geometry which is the topology product of a two dimensional simplex θ_2 and a hypersphere with a radius Th . P is formulated as $P = \{X | \rho_{X\theta_2} \leq Th\}$, while $\rho_{X\theta_2} = \min \rho(X, Z), \forall Z \in \theta_2$, $\rho(X, Z)$ is the Euclid distance between two points.

103.2.2 Cognitive Algorithm

After feature extraction we get a point set of the training samples are as follow steps:

Step 1: Suppose a type of keyword training samples' point set is $\vec{X} = \{\vec{x}_1, \dots, \vec{x}_T\}$, $\vec{x}_i \in R^n$, n represents the dimension of the training samples' space, T is the number of the points. Calculate the distance between every two points of all the training samples, and find out two points with minimum distance, record them as A and B. Working out the sum of the distances from other points to these two points. Find out the shortest distance and record the point which is not inline with A, B as C.

Constitute the first two dimensional simplex

$$\theta_2 = \theta_{(ABC)} = \left\{ Y | Y = a_1A + a_2B + a_3C, \sum_{i=1}^3 a_i = 1, a_i \geq 0 \right\}$$

We use one neuron z_1 to cover it.

Step 2: Exclude the points which have been covered by z_1 , and find the point which is with the shortest sum of distance to the three vertexes (A, B, C) of the first two dimensional simplex from points remainder, record it as D. Then find out the two points whose distance to D are shorter among A, B and C, and record them as E and F. Then D, E and F constitute the second two dimensional simplex $\theta_{(DEF)}$, and cover it with a neuron z_2 .

Step 3: Repeat the above steps until all the training points have been covered. Finally we can create m neurons, and the area this type of sample points has covered is a union of these neurons' covering area: $Z = \bigcup_{j=1}^m z_j$

Use the same method to the other kinds of samples and finally we get t covering areas of all the sample points: $Z = \bigcup_{j=1}^m z_j \quad i = 1, \dots, t$

103.2.3 Recognition Algorithm

For each sample point x to be recognized, calculate the distance ρ between x and the neuron U_j

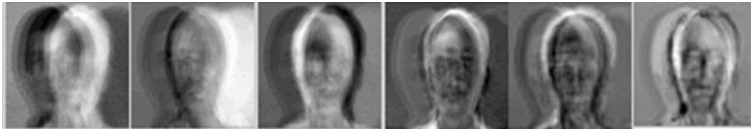


Fig. 103.1 Parts of Eigen faces

Table 103.1 Result of UMIST database

Number of training samples	240
Number of testing samples	324
Errors	8
Correct recognition rate (%)	97.5

Table 103.2 Result of YALE database

Number of training samples	60
Number of testing samples	105
Errors	15
Correct recognition rate (%)	85.7

$$\rho_{xU_j} = \|x - U_j\|$$

We use the minimum distance method to judge if x is belonged to U_j or not [9].

103.3 Application in Face Recognition

103.3.1 Experiments and Results

UMIST face database consists of 564 facial images of 20 persons, and all images are grayscale image. In order to improve the computing speed, we process the images to 64×64 pixels. We take 12 images of each person total 240 images to make PCA. 240 Eigen faces are gotten. Parts of Eigen face are shows in Fig. 103.1. We test the face sample as the algorithm described in Sect. 103.2. And the experimental results can be seen in Table 103.1.

YALE face database consists of 165 facial images of 15 persons, and all images are grayscale image. We also process the images to 64×64 pixels. We take 4 images of each person total 60 images to make PCA. 60 Eigen faces are gotten similarly. And the experimental results can be seen in Table 103.2.

Table 103.3 Training parameters of every type of keywords

Keywords	Baba	Mama	Piqiu	Renmen	Zhongguo
Number of neurons	79	127	79	44	133
Threshold:Th	90	120	80	40	100
Number of training samples	81	134	82	46	136

Table 103.4 The results of five kinds of keywords under certain threshold

Keywords	Baba	Mama	Piqiu	Renmen	Zhongguo
Number of test sentences	43	66	44	22	22
Total number of keywords	64	108	66	44	110
Correct recognition rate	90.6	86.1	95.5	75.0	97.27
False alarming rate	7.52	4.55	8.09	7.21	6.44
Threshold:TH	186	196.5	186	242	220

103.3.2 Results Analysis

From the experimental results can see: (1) A face recognition system based on this cognition algorithm trains each type of samples respectively, therefore, a new type of samples will not affect the trained ones. (2) The training samples in the experiments are continuous, which is the requirement for a face recognition system based on Biomimetic Pattern Recognition. (3) Method of covering of complicated geometrical shapes in high-dimension space is adopted to construct the sample space in Biomimetic Pattern Recognition.

103.4 Application in Keywords Recognition

103.4.1 Experiments and Results

The method of keywords future extraction is not detailed in this chapter, the keywords and the training parameters are showed in Table 103.3 [10].

We use the remaining continuous speeches which havenot been trained as our test sentences. Then feature extracted them as the training part and test them follow the recognition algorithm in Sect. 103.2. The results can be seen in Table 103.4.

103.4.2 Results Analysis

We take a sentence (‘ni shi zhong guo ren, wo shi zhong guo ren, ta shi zhong guo ren, wo men dou shi zhong guo ren, wo men dou ai zhong guo’) from a untrained

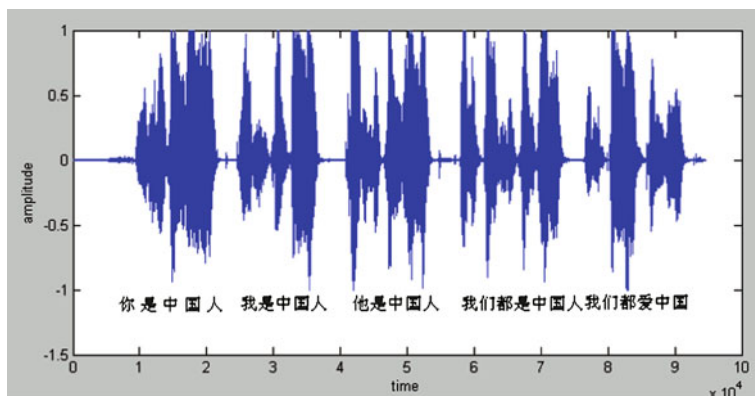


Fig. 103.2 Wave of the test speech

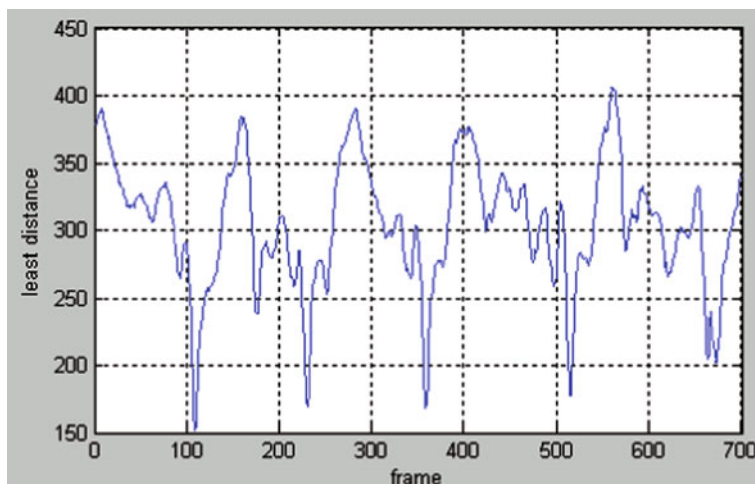


Fig. 103.3 The minimum distance curve between every point of the whole sentence to the keyword net 'zhong guo'

person as an example to explain the recognition process of the keyword 'zhong guo'. Figure 103.2 is the original speech wave, namely the test wave; Fig. 103.3 shows the minimum distance curve of the whole sentence to the keyword net 'zhong guo'; Fig. 103.4 shows the square wave after we ascertain the threshold ($TH = 220$), from it we can see the number of the vales. Finally, we show the ROC curve, which depicts the relationship between the correct recognition rate and the false alarming rate, from all the test speeches to the keyword net 'zhong guo' in Fig. 103.5.

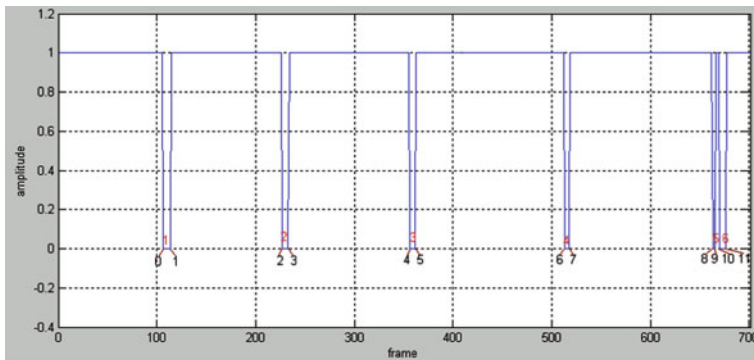
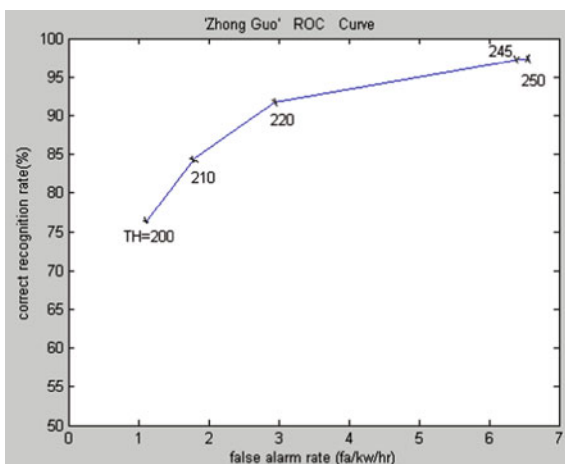


Fig. 103.4 The square wave and the number of vales after we ascertain the threshold 220

Fig. 103.5 The ROC curve of 'zhong guo'



From the experiment results we can see that when the false alarming rate is under 10, the highest recognition rate reaches 97.27%. Commonly, the correct recognition rate is higher than 86%. When the training samples are fewer, namely not contain enough information, the recognition rate is 75%.

103.5 Conclusion

This chapter presents a research to the cognitive algorithm based on biomimetic pattern recognition theory. For proving the effectiveness of the algorithm we applied it in face recognition and keywords recognition. The results show that the cognitive model can be used in many areas of pattern recognition. And the algorithm should be improved to get better efficiency in the future.

References

1. Turk M, Pentland A (1991) Eigenfaces for recognition [J]. *Cogn Neurosci* 3(1):71–86
2. Belhumeur PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. fisherfaces: recognition using class specific linear projection [J]. *IEEE Trans* 19(7):711–720 PAMI2
3. He X, Yan S et al (2005) Face recognition using Laplacianfaces [J]. *IEEE Trans* 27(3):328–340 PAMI2
4. Hang-jun Y, Hui-sheng C et al (1995) *Speech signal digital processing* [M]. Publishing House of Electronics Industry, Beijing
5. Junkawitsch L, Hoge NH et al (1996) A new keyword spotting algorithm with re-calculated optimal thresholds [A]. *The fourth international conference on spoken language processing* [C], vol 4, pp 2067–2070
6. Shoujue W (2002) Biomimetic pattern recognition—a new model of pattern recognition theory and its applications [J]. *Chin J Electron* 30(10):10
7. Shoujue W (2008) *First step to multi-dimensional space biomimetic informatics* [M]. National Defense Industry Press, Beijing
8. Shoujue W (2004) Face recognition: biomimetic pattern recognition vs traditional pattern recognition [J]. *Chin J Electron* 32(7):1057–1106
9. Shou-jue W, Bai-nan W (2002) Analysis and theory of high-dimension space geometry for artificial neural networks. *Acta Electron Sin* 30(1):1–4
10. Shoujue W, Sunyuan S, Wenming C (2006) Research on continuous speech recognition with small vocabulary based on biomimetic pattern recognition. [J]. *Chin J Electron* 15(4A): 845–848

Chapter 104

One Way to Enhance the Security of Mobile Payment System Based on Smart TF Card

Xiaoning Jiang, Xiping Wang and Yihao Huang

Abstract One smart TF card based method to enhance the security of mobile payment is proposed in this chapter. Its main purpose is to solve the problems of identity authentication and information leakage. In the chapter, we provide the design and implementation of mobile phone client system in detail. It adopts double factor authentication of user identity to enhance the security of legal user's identity. At the same time, it uses the traditional shared key protocol for reference to introduce a two-way authentication protocol. What's more, the non-reputation, confidentiality and integrity of the transaction can be guaranteed by the signature.

Keywords Mobile payment · Security · Smart card · Identity authentication · Authentication protocol

This work was supported by the Innovation Fund for Technology Based Firms of China (09C26223304071), the Important Science and Technology Special Project of Zhejiang Province (2010C11G2050012) and the graduate student Research Innovation Fund of Zhejiang Gongshang University (1120XJ1510116).

X. Jiang (✉) · X. Wang · Y. Huang
College of Information and Electronic Engineering, Zhejiang Gongshang
University, Hangzhou 310018, China
e-mail: jiangxiaoning@mail.zjgsu.edu.cn

X. Wang
e-mail: wxp421421@163.com

Y. Huang
e-mail: hyhlfq@live.cn

104.1 Introduction

Along with the issuance of 3G license, the problem of network bandwidth of mobile equipment has been solved basically. The motion applications come one after another incessantly, but the electronic commerce applications using mobile equipment, like the handset paying, actually could not obtain widespread promotion. The key factor obstructing its development is the security problem. The security challenges in mobile commerce are related to but not limited to the mobile devices, the radio interface, the network operator infrastructure and the type of mobile commerce application [1]. The primary is the identity authentication of mobile equipment and user. The safety level of mobile equipment is not high enough, and the identification of equipment is hard to implement; In the next place, because of the wireless transmission, for the general mobile equipment which is lack of cipher processing capacity, the information leakage occurs easily and the integrity cannot be effectively guaranteed in the process of payment. How to practically ensure the security of the mobile payment and let general users accept the principle that mobile payment is safe, is a burning question which various manufactures of industry chain confront.

104.2 Related Works

We know that the primary security problems in current mobile payment are the identity authentication and the information leakage in wireless transmission by the analysis above. Next we will compare the strengths and weaknesses of current identity authentication method and data encryption in transmission.

104.2.1 Identity Authentication Method in Mobile Payment Process

The most simple mechanism to discriminate the authenticity of sender's identity is interactive proof [2], which uses the information both sides share to verify the identity by a series of questions and answers. The main identity authentication technique in mobile payment can be divided into three types: static password based authentication, dynamic password based authentication, and the association of the above two.

(1) Static password based authentication

Static password based authentication is a method that the user directly passes the password to the server in the form of plaintext, while the server directly

keeps the user's password for the purpose of verification. This method is simple. However, there are severe safety risks of this method [3]:

- (a) It cannot resist password-guessing attack and replay attack;
- (b) It can only carry out unilateralism authentication, namely only the system can verify the user while the user can not verify the system. The attacker may masquerade to be the system to defraud the user of this static password;
- (c) Since all users' passwords are stored in the form of file at the authenticator, the attacker can initiate offline dictionary attacks once he steals the password file.

So, in terms of safety, this is a polar unsafe identity authentication method.

(2) Dynamic password based authentication

Dynamic password, also names for One-Time-Password (OTP) [4]. It makes the password change dynamically and continuously. It adopts a kind of special hardware called Dynamic Token to generate current password according to current time or used times. The authentication server takes the same operation. Dynamic password technique can resist the replay attack, but it still has following defects:

- (a) The implementation is relatively complicated, and the security cannot be guaranteed once the dynamic password generator is lost.
- (b) There still exists password-guessing attack;
- (c) It can only carry out unilateralism authentication, namely only the system can verify the user, so it cannot avoid the attack from the server end.

104.2.2 Mechanism of Data Encryption

The data encryption is implemented by handset software based on CPU currently. Because of the limited operation capability, it can only take some simple operation. What's more, the process of signature makes a high claim for processing capacity and memory. Strong data encryption cannot come true in current conditions.

In the view of the existing technology's insufficiency, we propose a smart TF card based way which combines modern cryptology technique, smart card technique and TF technique. The mobile equipment with embedded smart TF card is joined by the external bus. The two-way authentication, user identity authentication and the signature on the transaction information in the process of payment are implemented by calling the smart card's firmware to ensure the confidentiality of the client info, the legal identity of the exchange, and the none-repudiation of the transaction info.

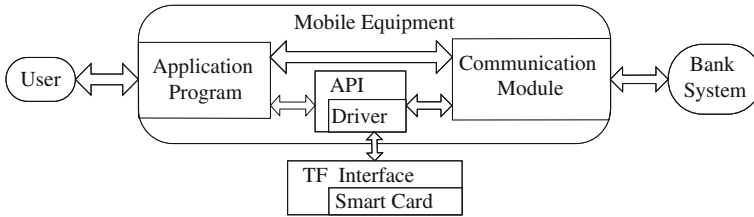


Fig. 104.1 General frame of the mobile phone client system

104.3 Smart TF Card Based Security Improvement Method

104.3.1 Mobile Phone Client System

The general frame of the mobile phone client system is as follow, including the smart card, the firmware of the smart card and the software on the mobile phone (Fig. 104.1).

104.3.1.1 Smart Card

The smart card is designed based on the smart IC card chip. The structure of chip includes the CPU unit, FLASH unit, symmetry cipher operation unit, asymmetry cipher operation, true random generator unit, etc. It favors secure and effective process of the cipher operation availably. In the outside, it supplies TF interface, which joins the mobile’s external interface bus, such as TF bus.

104.3.1.2 Firmware of Smart Card

The firmware is designed as a limited status machine, by calling the peripheral hardware to complete the cryptology operation (such as data’s encryption, decryption, signature, verification and random generation etc.) and to realize the PIN code authentication, the two-way identity authentication with the external terminal or system, and some cryptology application protocol related to some specific business.

104.3.1.3 Mobile Phone Software

The software runs on the mobile equipment, which includes three parts: drive procedure software (Driver), application program interface software (API) and the management software (Manager).

Driver realizes the original interactive mechanism with the smart card, and reveals the function of smart card. It provides service to software on API layer. It has various versions to adapt to distinct OS platform, such as Android, Window Mobile and so on. API layer packages the driver function further, so it can be called by the application procedure and management procedure. It can provide C/C++ interface or JAVA interface. Management software is mainly used to complete the initiation of smart card, PIN code modification and other parameter setting.

According to above three parts, it realizes strong identity authentication protocol in the mobile environment to promote the security of the mobile commerce transaction.

104.3.2 Double Factor Authentication of User Identity

In many commerce applications, one of the critical tasks is to authenticate who the authorized user is [5]. On the assumption of the usability and security, double factor authentication which refers to knowing the PIN code and owning the smart card, is regarded as the most safe and effective method [6]. The user can entry the payment process only when he gains the two factors simultaneously.

104.3.3 Two-Way Identity Authentication Protocol

Two-way identity authentication refers to the authentication between the payment terminal and the smart card. The principle of the two-way identity authentication is that the legal payment terminal and the mobile equipment should share the internal and external key in common, the imitated terminal or equipment is unable to gain the corresponding key. So this chapter adopts a protocol, which uses the shared key based authentication protocol [7] for reference and draws lessons from Ref. [8]. It is combined with smart card technology to improve the proposed protocol [9], which ensures the confidentiality of the shared key and makes the best use of the smart card's operational capability to simplify the traditional shared key protocol.

The symbols in the protocol

Suppose A is the mobile equipment, B is the payment terminal (Table 104.1):
The content of the protocol

- (1) $A \rightarrow B: ID_A \parallel R_A \parallel T_A$
- (2) $B \rightarrow A: R_B \parallel E_{K_{AB}}[f(R_A) \parallel T_B \parallel T_A]$
- (3) $A \rightarrow B: E_{K_{AB}}[f^{-1}(R_B) \parallel T_B \parallel T_A']$
- (4) B permits or denies A to carry on a session

A sends its identification ID_A , random number R_A and the current time T_A to B;

Table 104.1 Meanings of the symbols in the protocol

Symbols	Meaning
ID_A	A's identification
R_A/R_B	Random number generated by A/B
T_A/T_B	The current time of A/B
K_{AB}	The shared key of A and B
$f(M)$	An operation to M using the function f (function f is a simple reversible operation which is agreed by both part)
$f^{-1}(M)$	An inverse operation to M using the inverse function f^{-1}
$E_m[M]$	Encrypt M with DES algorithm by the key m
\parallel	The operator to connect the string
$A \rightarrow B:M$	Send message M from A to B

When B receives the message, it will take f operation to R_A . Then encrypt the result $f(R_A)$, current time T_B at B and received time T_A , and send the ciphertext along with the random number R_B to A;

When A receives the message, it will decrypt the ciphertext $E_{K_{AB}}[f(R_A) \parallel T_B \parallel T_A]$ with the shared key K_{AB} to gain the plaintext. Compare the deviation value between current time T_N and previous time T_A with the double deviation value between T_B and T_A to judge the message's degree of novelty. If the novelty condition meets, then take the inverse f operation to $f(R_A)$. if the result equals to R_A , then judge the legality of B's identity, otherwise rule it to be illegal. If B is legal, then take the inverse f operation to R_B . Encrypt the received time T_B at B, the time received the message T_A' with the result. Send the ciphertext to B;

When B receives the message, it will decrypt the ciphertext $E_{K_{AB}} [f^{-1}(R_B) \parallel T_B \parallel T_A']$ with the shared key K_{AB} . Compare the deviation value between current time $T_{N'}$ and previous time T_B with the double deviation value between T_A' and T_B to judge the message's degree of novelty. If the novelty condition meets, then take the f operation to $f^{-1}(R_B)$. If the result equals to R_B , then judge the legal identity of A, otherwise rule it to be illegal and reject to open out a session.

If the two-way authentication passed, A can choose an session key K_s , encrypt it with the shared key K_{AB} , and then send it to B to carry out a session with K_s .

The analysis of security

(1) Prevent the replay attacks effectively

To avoid the replay attack, the protocol not only uses the request–response handshake method, but also adds the time stamp and random number in the sending message [10]. What's more, it skillfully make good use of the method that distributional opposite time to avoid the problem of complicated time synchronization, which assures the degree of novelty while preventing the replay attack availablely.

(2) Stop the attack-in-middle efficaciously

This protocol is a two-way authentication one, so it carries out a source data authentication on each information exchange side. The attacker cannot

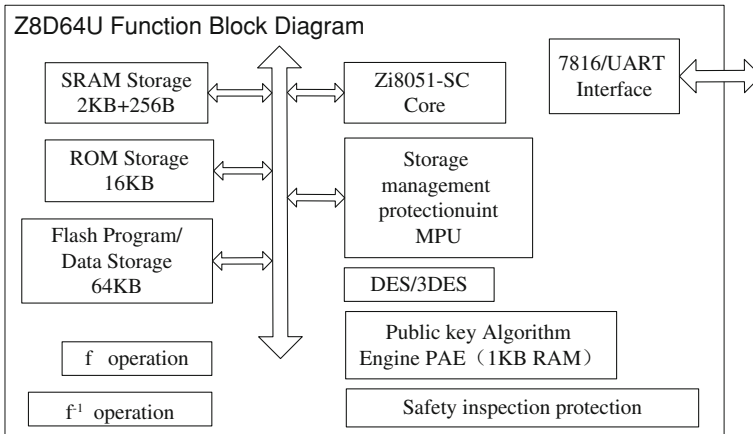


Fig. 104.2 Internal structure diagram

pretend or steer a medium modification and then forward the authentication info.

(3) Avoid the reflection attack effectually

This protocol takes f and inverse f operation to the random numbers with the help of the smart card’s mighty operation capability. So the third party cannot gain some valuable info before he proves himself, and he cannot initiate reflection attack.

104.4 Implementation of the Method

104.4.1 Internal Structure of Z8D64U Chip

We chose the Z8D64U chip as the smart card in our design. The internal structure diagram is shown as in the Fig. 104.2. It commendably fulfils the request we proposed.

104.4.2 State Transition of Firmware

The smart card firmware is designed as a limited status machine, which has eleven states in our design. The state-transition diagram is shown as in the Fig. 104.3

Initial state: wait state, when it receives a authentication request, it changes into the PIN code verification state;

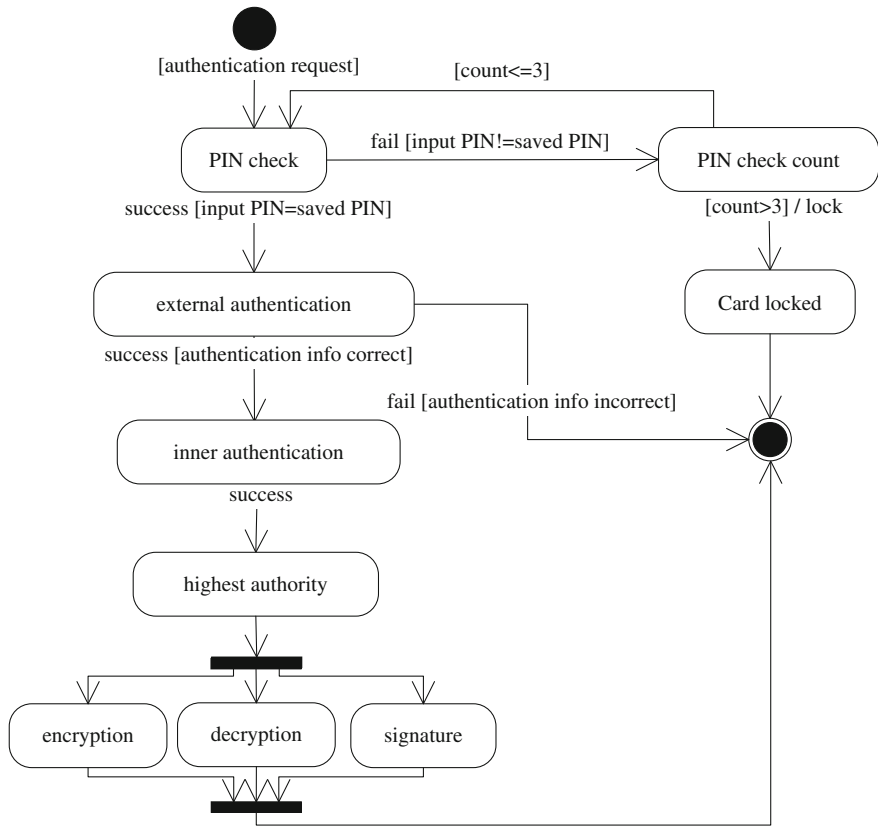


Fig. 104.3 State-transition diagram

PIN code verification state: waiting for the user to input the PIN code, and comparing the input PIN code with the PIN code stored in the smart card. If it is correct then it changes to the external authentication state, otherwise it switches to the PIN code verification count state;

PIN code verification count state: it takes count of the PIN code verification failures. If the count is less than or equal to three, then it changes into the PIN code verification state, otherwise it switches to the smart card locked state;

External authentication state: it is an authentication to the payment terminal. It comes to a conclusion of the authentication result by sending, receiving and verifying certain challenge data. If it returns true, it will switch to the internal authentication state; if it return false, it switches to the cease state;

Internal authentication state: waiting for authentication launched by the payment terminal. Return the related data by receiving and processing;

Smart card locked state: if the PIN code verification failure count outweighs three, then the card switches to this state;

The highest authority state: if the PIN code verification, internal and external authentication are all passed, then it changes into this state, which can take certain operations like encryption, decryption and signature and so on;

Encryption state: under the highest authority, it can perform encryption according to the users' request. When the ciphertext is sent successfully, it returns to the cease state;

Decryption state: under the highest authority, it can perform decryption according to the users' request. When the plaintext processing ends, it returns to the cease state;

Signature status: under the highest authority, it can perform signature according to the users' request to ensure the non-repudiation. When the ciphertext is sent successfully, it returns to the cease state;

Cease state: when all the operation completes, the card changes into this state, after a moment, it switches into the initial state.

104.5 Conclusion

This chapter proposed a smart TF card based method to enhance the security of mobile payment through comparing the advantages and disadvantages of current authentication methods. In the chapter, it introduces the general frame of mobile phone client system in detail, which includes the smart card, the firmware of the smart card and the software on the mobile phone. The two-way authentication between the mobile equipment and the payment terminal, user identity authentication and the signature on the transaction information in the process of payment are implemented by calling the firmware of the smart card to ensure the security of identity authentication and transaction information in the payment process. This system has been on trial in Huaxia and Jiaotong bank. These two banks are satisfactory to the usage of the system.

References

1. Security for Mobility (2004) The Institution of Electrical Engineers, London
2. Zhe C (2010) Research and implementation of authentication mechanism based on USB key. Computer software and theory college in Dongnan University, Jiangsu
3. Jia-neng Y (2009) The security analysis and security system of the building of the mobile payment system. Information and Engineering college of Nanchang University, Jiangxi
4. Xi Y, Yi-ru Y, Qin C (2010) The implementation of handset token based dynamic password authentication system. *Comput Syst Appl* 19(10):32–36
5. Iuon-Chang L, Chin-Chen C (2009) A countable and time-bound password-based user authentication scheme for the applications of electronic commerce. *Inf Sci* 179:1269–1277
6. Yi-zhen T, Zhi-jie W, Ding-yong T, Shan G (2009) USB Key based J2EE authentication system. *War Ind Autom* 28:87–91

7. Gui-xi, X., Xiao-hu, W. (1998) *Computer Networks*. 3rd edn., Prentice Hall, Beijing, pp 467–468
8. Shi-ming D, Lian-zhong L, Zhen L (2005) A USB-KEY based identity authentication protocol. *Microcomput Dev* 15:1–3
9. Xin W, Chao-wei Y (2010) An semi-anonymity offline mobile payment protocol based on smart card. *J China Univ Posts Telecommun* 17:63–66
10. Jie W (2006) *Computer network security theory and practice*, edn, Higher Education Press, Beijing, p 7

Chapter 105

Localization on Discrete Grid Graphs

Anna Gorbenko, Vladimir Popov and Andrey Sheka

Abstract Grid graphs are popular testbeds for planning with incomplete information. In particular, it is studied a fundamental planning problem, localization, to investigate whether gridworlds make good testbeds for planning with incomplete information. It is found empirically that greedy planning methods that interleave planning and plan execution can localize robots very quickly on random gridworlds or mazes. Thus, they may not provide adequately challenging testbeds. On the other hand, it is showed that finding localization plans that are within a log factor of optimal is NP-hard. Thus there are instances of gridworlds on which all greedy planning methods perform very poorly. These theoretical results help empirical researchers to select appropriate planning methods for planning with incomplete information as well as testbeds to demonstrate them. However, for practical application of difficult instances we need a method for their fast decision. In this paper we describe an approach to solve localization problem. This approach is based on constructing a logical model for the problem.

Keywords Localization · Grid graph · Genetic algorithm

A. Gorbenko (✉) · V. Popov · A. Sheka
Ural State University, Ekaterinburg, 620083, Russia
e-mail: gorbenko.aa@gmail.com

V. Popov
e-mail: Vladimir.Popov@usu.ru

A. Sheka
e-mail: andrey.sheka@gmail.com

105.1 Introduction

A testbed is a platform for experimentation of development projects. Testbeds allow for rigorous, transparent, and replicable testing of scientific theories, computational tools, and new technologies. A typical testbed could include software, hardware, and networking components. Testbeds are widely used for planning. In this context, testbeds are planning domains that allow researchers to evaluate their planning methods, communicate performance results of their methods to others, interpret published performance results of others more easily, and compare their methods against these performance results [1]. Planning researchers have studied in detail the properties of their testbeds for planning with complete information, such as blocksworlds and sliding tile puzzles. Examples of such experimental and theoretical studies include [2–5]. In recent years, planning researchers have become interested in planning with incomplete information (see [6–10]). This is an important research direction because, in the real world, complete information is often not available.

Testbeds should be easy to describe, but they should also provide a wide enough variety to mimic real domains. In particular, testbeds must include cases that are not too easy to solve because otherwise planning methods would appear to be more efficient than they actually are in some of the domains of interest.

Gridworlds are popular testbeds for planning with incomplete information. In [6] studied a fundamental planning problem, localization, to investigate whether gridworlds make good testbeds for planning with incomplete information. In [6] found empirically that greedy planning methods that interleave planning and plan execution can localize robots very quickly on random gridworlds or mazes. Thus, they may not provide adequately challenging testbeds. On the other hand, in [6] showed that finding localization plans that are within a log factor of optimal is NP-hard. Thus there are instances of gridworlds on which all greedy planning methods perform very poorly. In [6] showed how to construct them. These theoretical results help empirical researchers to select appropriate planning methods for planning with incomplete information as well as testbeds to demonstrate them. However, for practical application of difficult instances we need a method for their fast decision. In this paper we describe an approach to solve localization problem. This approach is based on constructing a logical model for the problem.

105.2 Grid Graphs Planning Tasks

We study localization tasks in grid graphs. Localization is a prototypical planning task with incomplete information. Before performing this task the robot knows a map of the gridworld but does not know its start cell. Evidently, the robot may need to localize prior to performing many other tasks.

The sensors onboard the robot tell it in every cell whether the cells immediately adjacent to it in the four compass directions (north, east, south, west) are traversable. The border of the grid graph is untraversable and observed as such. The robot can then move one cell to the north, east, south, or west, unless that cell is outside of the grid graph or untraversable. In the latter case the robot remains in its current cell. We assume a point robot with accurate sensing, perfect actuation, and knowledge of its orientation from an onboard compass.

The robot is localized if it knows its current cell. A deterministic localization plan specifies the movement to execute based on all previous movements and observations. A localization plan is valid if and only if there is no matter which cell the robot is started in; it eventually prints out its current cell or correctly determines that localization is impossible. The objective of planning then is to determine a valid deterministic localization plan that minimizes the number of movements for the worst possible start cell. We first calculate the number of movements for each possible start cell. The worst-case performance is then the maximum of these values.

In the decision version the valid deterministic localization plan problem can be formulated as following.

We can suppose that a grid graph is given by a matrix $(g[i, j])$ where m and n are dimensions of G , $g[i, j] = 1$ or $g[i, j] = 0$, and $g[i, j] = 1$ if and only if the cell with coordinates i and j belongs to G .

The Valid Deterministic Localization Plan Problem (VDLPP):

Instance: A grid graph G , a natural number K .

Question: Is there a valid deterministic localization plan such that the worst-case performance of this plan does not exceed K ?

In [6] showed that VDLPP is NP-complete.

105.3 Logical Model

The propositional satisfiability problem (SAT) is a core problem in mathematical logic and computing theory. Propositional satisfiability is the problem of determining if the variables of a given Boolean function can be assigned in such a way as to make the formula evaluate to true. SAT was the first known NP-complete problem, as proved by Stephen Cook in 1971. Until that time, the concept of an NP-complete problem did not even exist. Considered also different variants of the satisfiability problem.

Encoding problems as Boolean satisfiability and solving them with very efficient satisfiability algorithms has recently caused considerable interest. In particular, local search algorithms have given impressive results on many problems. For example, there are several ways of SAT-encoding constraint satisfaction, clique, planning, maximum cut, Hamiltonian cycle, vertex cover, maximum independent set, and colouring problems. There is a well known site on which posted solvers for SAT [11]. These solvers are divided into two main classes: stochastic local

search algorithms and algorithms improved exhaustive search. All solvers allow the conventional format for recording DIMACS Boolean function in conjunctive normal form and solve the corresponding problem [12]. In addition to the solvers the site also represented a large set of test problems in the format of DIMACS. This set includes a randomly generated problem of SAT. Of course, these algorithms require exponential time at worst. But they can relatively quick receive solutions for many Boolean functions. Therefore, it is natural to use a reduction to different variants of the satisfiability problem to solve computational hard problems.

Note that if we have a valid deterministic localization plan such that the worst-case performance of this plan does not exceed K then we have some sequence of instructions. We can assume that these instructions are defined as follows. “If x [north] = a , x [east] = b , x [south] = c , and x [west] = d , then M ”, where x [north], x [east], x [south], and x [west] are Boolean variables whose truth means that corresponding cells are vertices of grid graph, a , b , c , d are Boolean constants, and M is the direction. Not very difficult to construct a Boolean function which is true if and only if there is a valid deterministic localization plan consisting of K actions. This function can be constructed so that, using a set of values on which the function is true, we automatically obtain a sequence of instructions. This function gives us not only a SAT-encoding of VDLPP but also a way to obtain a sequence of instructions.

105.4 Robot Experimental Setup

Assumptions under which we consider VDLPP are simplifying but sufficiently close to reality to enable one to use the resulting planning methods on real robots. Greedy localization, for example, has been used on Nomad 150 mobile robots. The success rate of moving was at least 99.57%, and the success rate of sensing was at least 99.38% (see [10, 13]). These large success rates enable one to ignore actuator and sensor noise, especially since the rare failures are usually quickly noticed when the number of possible locations drops to zero, in which case the robot simply reinitializes its belief state to all possible locations and then continues to use the localization algorithm unchanged.

Among practical applications of VDLPP, we can note the air landing of a robot and a variety of tasks for indoor autonomous service robots. Note that the need of localization for indoor service robots occurs either due to temporary equipment failure or because of deliberate refusal to memorize the route. In the first case conditions of VDLPP are not quite consistent with the real situation. In particular, in many cases, we can assume that the robot is close to some known point. In some other cases, we can assume that the robot motion through a fixed direction. These assumptions greatly simplify the solution of the problem of localization.

For our experiments, we use two mobile robots (see Fig. 105.1 and 105.2). Typical failures for our robots are following.

Fig. 105.1 Design of this robot based on the well-known RC cars. From RC-CAR AT-10ES Thunder Tiger [14] we use only the four wheel chassis, the high torque DC-MOTOR and a steering servo. The DC-MOTOR drives the chassis and a steering servo controls the direction. The electronic system based on SSC-32 microcontroller. Onboard computer based on a motherboard with x86 compatible processor AMD Geode LX600 for embedded systems

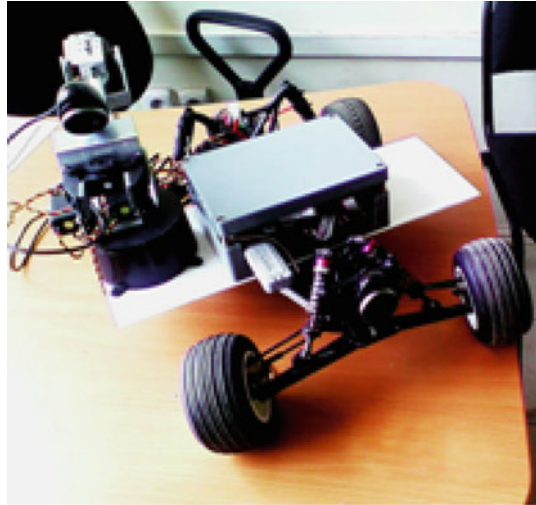
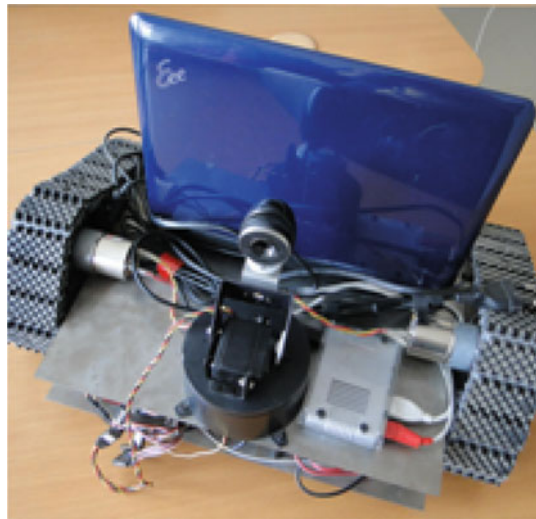


Fig. 105.2 Design of this robot based on the well-known Johnny 5 Robot [15]. By utilizing heavy duty polypropylene and rubber tracks with durable ABS molded sprockets the robot has excellent traction. It includes two 12vdc 50:1 gear head motors and the Sabertooth 2 × 5 R/C motor controller. Onboard computer of this robot is Asus Eee PC 1000HE



A collision with an undetectable object. When there is a collision with an undetectable object robot either stops or changes direction.

A tire started to deflate. In this case we get a predictable distortion of the trajectory of motion.

Rotation of a track is stopped. In this case we get the rotation along a known trajectory.

When the power of a servo motor is turned off, it begins to self-motion. It can give only a single change of the angle.

If you lose the connection with external navigation module motors continue to do the previous commands until the connection is restored. In this case, we know the trajectory of motion. Only the time of movement is unknown.

Loss of the next image in the case of using visual navigation. In this case, it is known that the robot is in a small neighborhood of the known point.

In each of these cases, we can consider instead VDLPP some more simple problem. On the other hand VDLPP of considerable interest in the case when the robot need not memorize the path.

Both of our robots have processors with relatively low performance. The requirement of localization while moving substantially reduces the speed of moving robots. Therefore, in solving many problems, we use visual topological navigation (see [16–18]).

Vision-based navigation systems may use either topological or metrical maps. In topological maps, only places such as rooms and their relations are learned and recognized, whereas in metrical maps, the precise positions of environment features and of the robot are estimated. In realistic scenarios for entertainment robotics, the robot can fall or be blocked in places where sensors will have difficulty to find useful information. In these situations, a metrical approach, that usually requires a continuous tracking of features, will probably fail, whereas a topological approach, able to recognize the rooms and guide the robot between them is more adapted. Moreover, topological approaches may be purely appearance based, thus avoiding the need for camera calibration.

In practice often used simple topological algorithms that allow only follow the target and avoid obstacles. Such algorithms provide very high performance. Note that usually robots do not have enough memory to store full visual series. Therefore, usually after the task the robot needs to solve the problem of localization. For example, the robot is moved directly by the user from one place to another, the robot is moved directly by another robot, robot uses skittles as landmarks and other robot rearranges skittles, etc.

Note that for a relatively small testbeds, we can apply brute force. However, this method is not suitable even for indoor laboratory testing. We create a generator of special hard (see [6]) and natural instances for VDLPP. We use algorithms from [11]. Also we design our own genetic algorithm for SAT which based on algorithms from [11]. We use heterogeneous cluster based on three clusters (Cluster USU, Linux, 8 calculation nodes, Intel Pentium IV 2.40 GHz processors; umt, Linux, 256 calculation nodes, Xeon 3.00 GHz processors; um64, Linux, 124 calculation nodes, AMD Opteron 2.6 GHz bi-processors) [19]. Each test was run on a cluster of at least 100 nodes. The maximum solution time was 14 h. The average time to find a solution was 12.3 min. The best time was 52 s. Note that the calculation exceeds 1 h is quite rare.

It is easy to see that solver for VDLPP give us theoretically a good solver for the valid deterministic localization plan problem whose use gives only a linear slowdown. However, in practice it may be slowing down a hundred times or more. Thus sometimes we can use a supercomputer to find an exact solution. In this case we can apply this solution on the robot. However, in many cases we are forced to

rely on heuristic methods of solution. Nevertheless, using a logical model provides us a good tool for verifying heuristic solutions. In addition, we can expect a significant acceleration of the process by using a more powerful supercomputer. Also we can use a logical model for VDLPP to create a training set for supervised learning of some intelligent algorithm for the valid deterministic localization plan problem.

105.5 Summary

In this paper we have presented an approach to solve localization problem. This approach is based on constructing a logical model for the problem. We also examined practical aspects of using a logical model for the solution of the valid deterministic localization plan problem and for the generation of testbeds for this problem.

References

1. Hanks S, Pollack M, Cohen P (1993) *AI Mag* 14:17
2. Gupta N, Nau D (1992) *Artif Intell* 56:223
3. Koenig S, Simmons R (1996) In: *Proceedings of the national conference on artificial intelligence*, p 279
4. Reinefeld A (1993) In: *Proceedings of the international joint conference on artificial intelligence*, p 248
5. Slaney J, Thiebaux S (1996) In: *Proceedings of the national conference on artificial intelligence planning*
6. Tovey C, Koenig S (2000) In: *Proceedings of the AAAI conference on artificial intelligence*, p 819
7. Koenig S, Likhachev M (2005) Fast replanning for navigation in unknown terrain. *IEEE Trans Robot* 21:354
8. Koenig S, Smirnov Y, Tovey C (2003) Performance bounds for planning in unknown terrain. *J Artif Intell* 147:253
9. Mudgal A, Tovey C, Koenig S (2004) In: *Proceedings of the international symposium on artificial intelligence and mathematics*
10. Tovey C, Koenig S (2010) *IEEE Trans Robot* 26:320
11. Information on <http://people.cs.ubc.ca/~hoos/SATLIB/index-ubc.html>
12. Information on <http://www.cs.ubc.ca/~hoos/SATLIB/Benchmarks/SAT/satformat.ps>
13. Nourbakhsh I (1996) In: *Proceedings of the AAAI-96 spring symposium on planning with incomplete information for robot problems*, p 86
14. Information on <http://www.tiger.com.tw/>
15. Information on <http://www.lynxmotion.com/c-103-johnny-5.aspx>
16. Filliat D (2008) In: *IEEE/RSJ international conference on intelligent robots and systems*. IEEE computer society press, New York, p 248
17. Park IP, Kender JR (1995) Topological direction-giving and visual navigation in large environments. *Artif Intell* 78:355

18. Santos-Victor J, Vassallo R, Schneebeil H (1999) In: Christensen HI (ed) Proceedings of the first international conference on computer vision systems. Springer, London, UK, p 21
19. Information on http://parallel.uran.ru/mvc_now/hardware/supercomp.htm

Part X
Electronic Applications

Chapter 106

The Research of Chinese Q&A System Based on Similarity Algorithm

Liu Zhen, Xiao Wenxian, Wan Wenlong and Yulan Li

Abstract On the basis of the research on the theories of the Chinese question answering system, in-depth analysis on the calculation methods of the current statement similarity has been done; the advantages and disadvantages of the calculation methods of the various statements similarity and its application have been studied; a new method of statements similarity calculation is put forward. On this basis, a Chinese question answering system based on a set of frequent questions was achieved. The results show that: the system is better than the current more popular VSM method in the match of the statement similarity.

Keywords The question answering system · Models · Questions recognition · Questions matching · Threshold · VSM

106.1 Introduction

Chinese Q&A system is an important branch of the information retrieval technique, which adopts new retrieval theory and has an efficient complement based on keywords to existing technical information retrieval deficiency, it is a research direction of new generation of information retrieval. The question answering system is more close to natural language by means of specific questions, the semantic aspect is understood by computers, and it has an important practical value.

L. Zhen (✉) · X. Wenxian · W. Wenlong · Y. Li
Henan Institute of Science and Technology, Henan Xinxiang 453003, China
e-mail: liuzhen@hist.edu.cn

The question answering system of the study and application develops more quickly in abroad. Because the syntactic analysis and lexical analysis based on English are more mature and corpus and test corpus based on English is various, and most of them were open source, the development of overseas question answering system develops first relatively. In contrast, the domestic research on question-answering system both in scale and in level, still exist great gap, the main reasons are: on the one hand, Chinese information processing characteristics and difficulties determine the relevant foreign mature technology and research results not to be directly used in Chinese Q&A system; on the other hand, because of the shortage of corresponding base corpus and corresponding evaluation standard.

106.2 The Theoretical Basis of Chinese Question Answering System

In Chinese Q&A system, user put forward problems; we need to search for the most similar questions of the user questions from the database. The conventional practice calculates similarity between user's problems and candidate problem repository of similar questions. When similarity exceeds a certain specified threshold, the similarity is the biggest value that users are finding sentences; if the user problems and all the candidate problems in the library of similarity between similar questions are less than threshold; it needs retrieval and gather the corresponding answers from corpus or network. Therefore, statements similarity definition and calculation is the key to the Chinese question answering system [1].

106.2.1 HowNet Structure

There are two main concepts in HowNet: "concept" and "righteousness ". We describe them as follows:

- (1) Concept: what is used to describe the lexical semantics is the concept identifier words expressed by words [2].
- (2) Righteousness used in the original: it is the basic and smallest meaningful unit division in describing concepts.

106.2.2 The Classification of Chinese Question Answering System

106.2.2.1 Field Classification

Chinese Q&A system, according to the subject content related to the field can be divided into specialized question answering system for an open field and the question answering system.

106.2.2.2 Characteristics Classification

Chinese Q&A system can be divided into: a. factual question Q&A system; B. listed sexual issues question answering system; C. definitional Q&A system based on the characteristics of question type and the TREC reference standard.

106.2.3 Characteristics of Chinese Information Processing

Chinese Q&A system has the following features compared with western automatic question answering system [3]:

- (1) Chinese words have a continuous writing and the Chinese word segmentation is Chinese information processing foundation. The Chinese Q&A system analysis question and answers and cope with sentences participle processing first.
- (2) The Chinese do not have the same shape change as English.
- (3) The grammars of Chinese sentences are very flexible and sentence types have no way to mention, the relationship between the components of a sentence is so complicated that it is difficult to find sentence rule.
- (4) It lacks of necessary corpus support.

106.2.4 Questions Eigenvector Extraction

Questions characteristic vectors, it points the extraction of Chinese word segmentation and the basis of part of speech tagging, and removes the conjunction, prepositions, onomatopoeic words, etc., and little sense to distinguish between the high-frequency words (for example: “this”) and low-frequency word formation of keywords sequence.

When we have extraction of characteristic vector, it is necessary to avoid removing those itself is low-frequency word, but it is very important terms. This requires a combination of specific context and special terms for special treatment.

106.3 Chinese Q&A System Model

Based on a deep research of plenty domestic and foreign question answering system, combines Chinese information processing features of statements, and puts forward a similarity calculation based on the Chinese Q&A system model in some basis, and constructs a common problem sets Chinese Q&A system.

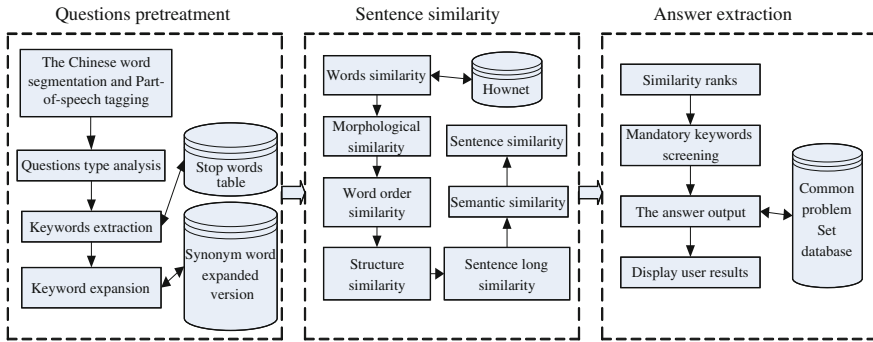


Fig. 1 Chinese Q&A system overall frame-chart

106.3.1 Framework of Chinese Q&A System

Chinese Q&A system is an advanced form of information retrieval, its purpose is to hope user use more appropriate and natural language manner of questions, and get an answer closer to natural language. Chinese Q&A system uses natural language expression for certain analysis question in understanding its semantic, based on the content, according to certain algorithm or take some strategies on the answer, finally present to the user from search web or document extracted Fig. (106.1) [4].

Based on statements from above knowable, similarity Chinese Q&A system consists of three parts, it is respectively: questions of pretreatment, sentence similarity calculation and answer extraction.

- (1) Questions pretreatment module mainly include: the Chinese word segmentation and part-of-speech tagging, questions type analysis, keywords extraction and keywords expansion.
- (2) Sentence similarity calculation module mainly include: words form, similarity calculation, word similarity calculation similarity calculation, structural similarity calculation, sentence long similarity calculation, semantic similarity calculation and sentence similarity calculation [5].
- (3) Answer extraction module mainly include: similarity ranks, mandatory keywords screening, the answer output and user results are shown.

106.3.2 Questions Pretreatment

106.3.2.1 Chinese Word Segmentation and Part-of-Speech Tagging

The Chinese word segmentation is Chinese information processing foundation, because Chinese is participle of ligatures, if we want to understand questions, we need to have the words after word segmentation speech tagging.

106.3.2.2 Determination of Questions Categories

The aim is to determine the question categories determined the semantic category and answer questions with what kind of strategy. After determining the type and search for question answer, we can use category special strategies and technical analysis of questions and questions similarity calculation and ultimately make the answers to the questions.

106.3.2.3 The Extraction of Keywords

After participle and part-of-speech tagging information retrieval, the key words is the input portion question, which directly affect the retrieval results behind.

106.3.2.4 Keywords Expansion

Chinese Q&A system evaluation standard is to use the system information retrieval precision and recall rate.

106.3.3 Weighted Feature Statements Similarity Calculation

106.3.3.1 Words Similarity

In HowNet as system of semantic knowledge base, mainly includes: words similarity calculation, righteousness original similarity calculation and the function concept similarity calculation and notional concept similarity calculation [6].

1. Words similarity calculation

Assume two words W_1 and W_2 , W_1 have n a meanings (concept): $C_{11}, C_{12}, \dots, C_{1n}$; W_2 have m meanings (concept) : $C_{21}, C_{22}, \dots, C_{2m}$; the maximum of each concept similarity is W_1 and W_2 similarity.

$$Sim(W_1, W_2) = \max_{i=1,2,\dots,n; j=1,2,\dots,m} Sim(C_{1i}, C_{2j}) \quad (106.1)$$

2. Righteousness original similarity calculation

We have already known concepts are expressed with righteousness; therefore, original of similarity degree between concepts can be converted to the similarity between righteousness original calculations

$$Sim(P_1, P_2) = \frac{\alpha}{d(P_1, P_2) + \alpha} \quad (106.2)$$

Among them: $d(P_1, P_2)$ is two righteousness in the original P_1 and P_2 righteousness in the hierarchical tree shortest path length (positive integer); α for adjustable parameters.

3. The similarity calculation of function concept

The rules here:

1. Alternation similarity between concepts is 0;
2. Virtual concept of similarity between the syntactic righteousness is their corresponding relationship between the original or the similarity of righteousness.
4. The similarity calculation of notional concept Notional semantic similarity calculation between concepts is divided into four parts:
 - (1) The first independent righteousness original similarity between righteousness is two of the similarities, which notes as $Sim_1(S_1, S_2)$;
 - (2) Other independent righteousness original similarity: because their descriptions type may be various, the specific calculation method as follows: The first, according to the biggest similarity to group; Secondly, the repeated cycle does not stop until all the original group complete independence of righteousness. Finally, calculated respectively the similarity between groups, and then weighted average overall similarity calculation, notes for $Sim_2(S_1, S_2)$.
 - (3) The relationship between the similarity: the righteousness of the righteous described the same relationship as a group, and calculated the similarity between them, noted as $Sim_3(S_1, S_2)$.
 - (4) Symbol righteousness the original similarity: description of the same type of relationship symbols as a group, and calculated the similarity between them, notes as $Sim_4(S_1, S_2)$. We can get two notional concept of semantic expression for the similarity through the above rules:

$$Sim(C_1, C_2) = \sum_{i=1}^4 \beta_i \prod_{j=1}^i Sim_j(C_1, C_2) \tag{106.3}$$

Among them: $\beta_i (1 \leq i \leq 4)$ is adjustable parameter, $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$, $\beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4$, $\beta_1 \geq 0.5$.

106.3.3.2 Morphological Similarity

Reflecting the morphological similarity degree of words in two sentences can be measured by the number of the same words contained in two sentences. Suppose S_1, S_2 are two sentences, the morphological similarity of S_1 and S_1 is:

$$Sim_1(S_1 S_2) = 2 * (SameWord(S_1 S_2) / (Len(S_1) + Len(S_2))) \tag{106.4}$$

Among them: SameWord (S_1, S_2) for S_1, S_2 contains the same number of words, Len (S) is the number of sentences S seen.

106.3.3.3 The Similarity of Trace Word Order

The word similarity of Sentence A and B $Sim_2(A, B)$ is decided by the following formula:

$$Sim_2(A, B) = \begin{cases} 1 - \frac{RevOrd(A,B)}{|OnceWS(A,B)|-1} & \text{when } |OnceWS(A, B)| > 1 \\ 1 & \text{when } |OnceWS(A, B)| = 1 \\ 0 & \text{when } |OnceWS(A, B)| = 0 \end{cases} \quad (106.5)$$

Easy to prove: $0 \leq Sim_2(A, B) \leq 1$.

106.3.3.4 Structural Similarity

Chinese sentence structure similarity mainly reflect the two sentences on the structure, the method of similarity main idea is: according to two Chinese sentence segmentation and part-of-speech tagging, it have been lexicon after and combined with different parts of speech, the sequence of weight, and the optimal sequence matching the match results. Its structure similarity calculation formula is as follows:

$$Sim_3(S_1, S_2) = \frac{2 \sum_{i=1}^C \frac{1}{1 + \sum_{j=1}^{D_i} W_j} W_i}{\sum_{k=1}^E W_k} \quad (106.6)$$

Among them: C is the same sequence number of two parts of speeches, D_i is the number of node merge a ring of node, E is the number of total knot points two parts of speech sequence (repeated computation). W_i is the first I a part of the same node (weights), W_j is ring first j a node, the weights of W_k is all nodes (parts of speech) first k a node metric.

106.3.3.5 The Similarity of Sentence Length

Reflecting two sentences in length form similar degree. The two statements long sentence in similarity calculation formula is as follows:

$$Sim_4(S_1, S_2) = 1 - abs \left| \frac{Len(S_1) - Len(S_2)}{Len(S_1) + Len(S_2)} \right| \quad (106.7)$$

Among them: abs shows absolute value, Len(S_i) stands for the number of words in S_i .

Among them: abs stands for absolute value, Len (S_i) represents the number seen S_i .

106.3.3.6 Semantic Similarity

Suppose the words included in the two sentences of S_1 and S_2 , the words included in S_1 are $w_{11}, w_{12} \dots w_{1m}$ while the words included in S_2 are $w_{21}, w_{22} \dots w_{2n}$, then the similarity between $w_{1i}(1 \leq i \leq m)$ and $w_{2j}(1 \leq j \leq n)$ is $Sim_5(w_{1i}, w_{2j})$, so the semantic similarity between sentence S_1 and S_2 is $Sim_5(S_1, S_2)$:

$$Sim_5(S_1, S_2) = \left(\frac{\sum_{i=1}^m \max\{Sim(w_{1i}, w_{2j}) | 1 \leq j \leq n\}}{m} + \frac{\sum_{j=1}^n \max\{Sim(w_{1i}, w_{2j}) | 1 \leq i \leq m\}}{n} \right) / 2 \tag{106.8}$$

106.3.3.7 Sentence Similarity

Reflecting the similarity degree between the two sentences. Usually expressed as a number between 0 and 1, 0 means no similar values, 1 said the exactly alike, the larger the number is, and the more similar the two sentences are:

$$Similarity(S_1, S_2) = \sum_{i=1}^5 \lambda_i \times Sim_i(S_1, S_2) \tag{9}$$

Among them: $\lambda_i (1 \leq i \leq 5)$ is constant, and meet $\sum_{i=1}^5 \lambda_i = 1$.

106.3.4 Answer Extraction

106.3.4.1 Track of Sentence Similarity Sorting

The similarity calculation is the key problem system database standard question sentences. When we display user results based on the similarity, we needs sort according to the sentence similarity. In addition, the system also need to set up a threshold, it can output retrieval results, only when statements similarity more than threshold.

106.3.4.2 The Mandatory Keywords Filtering Rainfall Distribution

We list filter through mandatory for retrieval results keywords and remove irrelevant results of user.

106.3.4.3 The Rule Final Answer Extraction Technology

It outputs users' retrieval results that the users are familiar with BBS style. In order to meet different user browsing habits, the user can search results custom-made

display styles, including: each page shows the number of records, page CSS style, etc.

106.4 Experimental Results and Analysis

This experiment used traditional vector space model (VSM) and the similarity calculation based on the statements of the test question matching method.

1. There is a comparison between two sets of data. The randomly S @ 1 value in the first set of data (from FAQ library question), after careful analysis, found the value is low, the wrong reasons are: some problems are more casual and contain little useful information. Some questions also have no answer. The second set of data repository (from the candidate selection problem question) S @ 1 value, but it is still a relatively ideal nearly 30% of the answer is wrong. After careful analysis found that the wrong reason was partly due to incorrect question category recognition, on the other hand is due to question or answer contains negative words, leading to questions matching is not correct.

2. The comparison between two kinds of similarity methods. Two groups of testing set is of big difference in similarity calculation method. The VSM method, mainly analysis from statements, statistical frequency surface such information; And matching method based on statements similarity questions, using the semantic knowledge base, is the deep analysis of statements, which raises questions about the accuracy of traditional, matching the VSM method still have certain enhancement and improvement.

106.5 Conclusion

Chinese Q&A system is the core content of natural language processing as well as one of the important research directions of intelligent information retrieval. The most direct embodiment whether the computer can realize intelligentization is whether it can answer the question raised by human with natural language. By means of in-depth research on the calculation methods of Chinese statement similarity, this article discussed and compared several similarity calculation methods; this paper puts forward a much weighted feature statements similarity algorithm, applies this algorithm to Chinese Q&A system, puts forward a Chinese Q&A system model. The experiment illustrates its questions matching accuracy increased significantly. If we can further improve optimization calculation efficiency and take statements of emotional color into consideration, then the answers achieved will be more optimal, which needs to be studied further.

References

1. Sujian L (2002) Research of relevancy between sentences based on semantic computation [J]. *Comput Eng Appl* 38(7):75–76
2. Li J, Cao X ,Yu F (2008) A word semantic automatic classification system based on word similarity computation [J]. *Comput Simul* 25(08):295–299
3. Wu Q, Xiong H (2010) Method for sentence similarity computation by integrating multi-features [J]. *Comput Syst Appl* (11):110 –114
4. Liu J, Zou P, Zhang P, Qi F (2010) Research on an improved algorithm of concept semantic similarity based on ontology [J]. *J Wuhan Univ Technol* 20:112–117
5. Liu Q, Gu X (2010) Study on HowNet-based word similarity algorithm [J]. *J Chin Inf Process* 06:31–36
6. Li S (2002) Research of relevancy between sentences based on semantic computation [J]. *Comput Eng Appl* 38(7):75–76

Chapter 107

A Model-Driven Method for Service-Oriented Modeling and Design Based on Domain Ontology

Ying Zhang, Xiaoming Liu, Zhixue Wang and Li Chen

Abstract In recent years, the continually changing requirements accompanied with the rapid growth of complexity makes it challenging to develop systems by traditional good design principles. Service-oriented Architecture offers a new paradigm to integrate heterogeneous systems flexibly and effectively. This chapter proposed an approach for developing service-oriented system embraced MDA philosophy. The method firstly builds domain model by domain specific modeling language. Then the domain model is mapped to the platform independent model, which is modeled by UML. Finally, the platform independent model is transformed to the platform specific model described by OWL-S. There is a domain ontology is defined to provide domain knowledge for modeling and transformation.

Keywords SOA · MDA · Domain ontology · UML profiles

107.1 Introduction

The continually changing requirements accompanied with the rapid growth of complexity makes it challenging to develop systems by traditional good design principles. Therefore, one of the hottest topics both in industry and research areas is how to integrate the heterogeneous systems fast and flexibly. Service-Oriented Architecture (SOA) [1, 2] offers a new paradigm to solve this problem because it

Y. Zhang (✉) · X. Liu · Z. Wang · L. Chen
Institute of Command Automation, PLA University of Science and Technology,
210007 Nanjing, People's Republic of China
e-mail: zhywl66@163.com

features loose coupling, dynamic binding and independent of development technologies, platforms and organizations.

On the other hand, the Object Management Group (OMG) initiates the Model Driven Architecture (MDA) [3, 4] to implement software development that emphasized model-based and automated code generation. MDA includes a set of standards such as MOF, CWM, QVT, etc. [5]. Generally, MDA has three layers of model, i.e. computation independent models (CIM), platform independent models (PIM) and platform specific models (PSM). And it focuses on the approach to develop applications from business requirement to implement code by creating and automatic transforming among these models.

There are plenty researches about SOA and MDA, follows are the most useful ones for this chapter: Claus [5, 6] presents a Web service composed architecture based on WSPO and WSMO. Karsai et al. [7] gives a brief synopsis on how MDA ought to be extended to handle embedded software development, and illustrates the concepts in practice using a prototype modeling language and tool chain designed for developing mission computing software. References [8–10] provide transformation methods between UML control constructs and OWL-S including for simple process and complex process. Kim and Lee [8] proposes method uses class diagrams to represent a domain ontology and sequence or activity diagrams to represent the behavior of a business process in order to facilitate construction of OWL-S ontology. Grønmo et al. [9] researches the correspondence between WSML, OWL-S and UML, and construct ontology for discovering, invoking, and composing semantics Web services.

The combination of SOA and MDA may be the good solution for system analysis and design. Therefore, this chapter proposes an agile and lightweight model-driven method for service-oriented modeling and design based on domain ontology. Accompany with an example in online entertainment domain, the rest of the chapter is organized as follows: in Sect. 107.2 we presents related works and our approach briefly. We modeling CIM model by domain-specific modeling language in Sect. 107.3, and the mapping from CIM to PSM is addressed in Sects. 107.4 and 107.5. At last, we draw some conclusions and future work in Sect. 107.6.

107.2 Our Approach

Ontology is the formulation of a conceptualization of a domain—usually hierarchically structured based on consumption (classification) relationships, but also other semantic relationship types such as composition [5]. We define domain ontology as follows:

Definition 1: the domain ontology is defined as;

$$\text{Domain ontology} = \langle \text{Concept, Relation, Constraint} \rangle$$

in which Concept is a finite set of domain-specific concepts, such as user, system, member, account, etc. Relation is a finite set of domain-specific relations

associated with a pair of the concepts. Constraint is a finite set of domain constraints of concept and relation such as each member must have a unique account. It can be expressed as logical formulas for model verification.

The core technology of MDA is model transformation. Compare with of typical MDA, there are three differences in our approach. First of all, the typical MDA researches focus on modeling and mapping of PIM and PSM rather than CIM. Our study, however, regard CIM as an important part, as well as the transformation between CIM and PIM. Secondly, we define a domain ontology to provide semantic support for modeling, model transforming and logic reasoning while this is not be emphasized in traditional MDA. In addition, our central aim is developing service-oriented systems, thus we choice Web service technology for platform specific modeling, which is not the same as the MDA that aims at generating executable code.

107.3 Domain Modeling

The purpose of the computation independent model (CIM) is to capture a domain with its concepts and properties [5].

UML [11] is a wide-spectrum language used for almost all applications across different domain. It has a broad range of concerns and consists of several types of diagram. Thus, it may be so complex and unsuitable for domain non-UML experts. On the other hand, Domain-Specific Language (DSL) is software language tailored to address problems in some application domain [12–14]. A well-known technique to create a DSL is extending an existing general-purpose language, which called a host language, with domain-specific constructs. The embedded DSL inherits all features of the host language [7].

Although defining DSL will cost some time, we still believe DSL is more suitable for domain modeling than UML because of it is closer to the specific domain and more easily to understand and use for domain experts. It will increase the efficiency of domain specific modeling by DSL after definition. There are two possible approaches for creating DSML when using MDA standards [12]. The one is defining a new language, which is an alternative to UML actually, based on MOF. The other is specializing pure UML suitable for domain modeling by using UML profiles. Obviously, the former is a heavyweight method, while the latter is a lightweight one. We prefer the latter one and call it Domain-Specific Modeling Languages (DSML) because of the host language is UML rather than general programming languages.

Syntax and Semantics are two essential components [7, 13] to define DSML. We define DSML semantics by the natural language in order to be consistent with its host language, UML. In fact, the domain ontology has given sufficient ontological notation to define the DSML. We give an instance of online entertainment domain, which perhaps is not comprehensive but enough, to illustrate how to

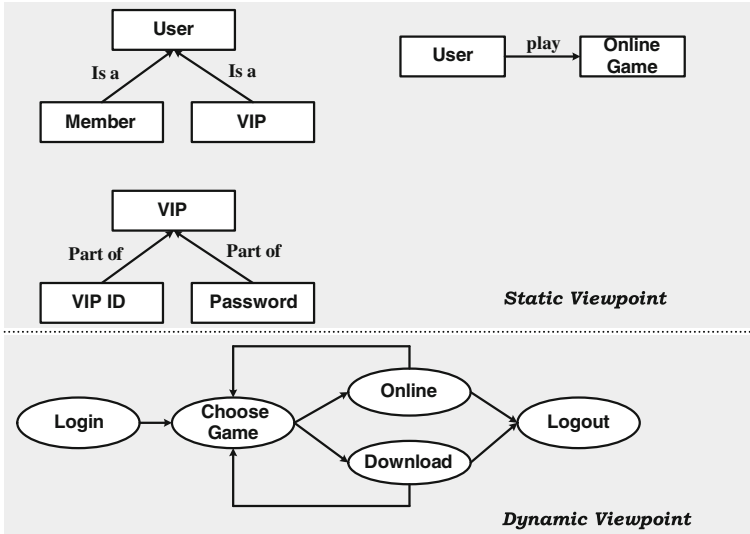


Fig. 107.1 Domain model

construct domain model by DSML. As shown in Fig. 107.1, DSML has three model blocks: concept, relation and viewpoint, as follows:

Two types of concept: static entity and dynamic entity. Static entity is the static objects or elements in domain. In our instance, it divided into three types: person, system and information, represented by rectangular. Dynamic entity usually describes activities or processes of domain which represented by ellipse.

Four types of relation: Is a relation describes inherited relationship between a static entity and its sub-concept. Part of relation defines the composition relationship between concepts. Association relation defines the structural relationship which could have its own name. These three relations represented by arrow with its relation name. Connection relation describes the transactions of dynamic entity which represented by arrow without relation name.

Two types of viewpoint: static viewpoint and dynamic viewpoint. Static viewpoint describes the hierarchical and structural properties of concepts while dynamic viewpoint describes the process of concepts, especially for dynamic entity.

107.4 Mapping and Modeling in PIM Layer

The platform-independent model (PIM) changes the focus from the computation-independent model to architecture modeling.

UML profiles provide a convenient way to transformation from DSML into UML, while all required information of transforming has been given by DSML

Table 107.1 Transformation rules for the CIM to PIM mapping

Rule	Domain constructs	Description
TR_1.1	Static entity	For each static entity in the CIM, create a class in the PIM
TR_1.2	Dynamic entity	For each dynamic entity the CIM, create an activity in the PIM
TR_1.3	Is a	Transform to generalization in the PIM
TR_1.4	Part of	Transform to aggregation in the PIM
TR_1.5	Association	The association relation is the same as in the PIM
TR_1.6	Connection	Transform to control flow in the PIM
TR_1.7	Constraints	Still constraints in the PIM

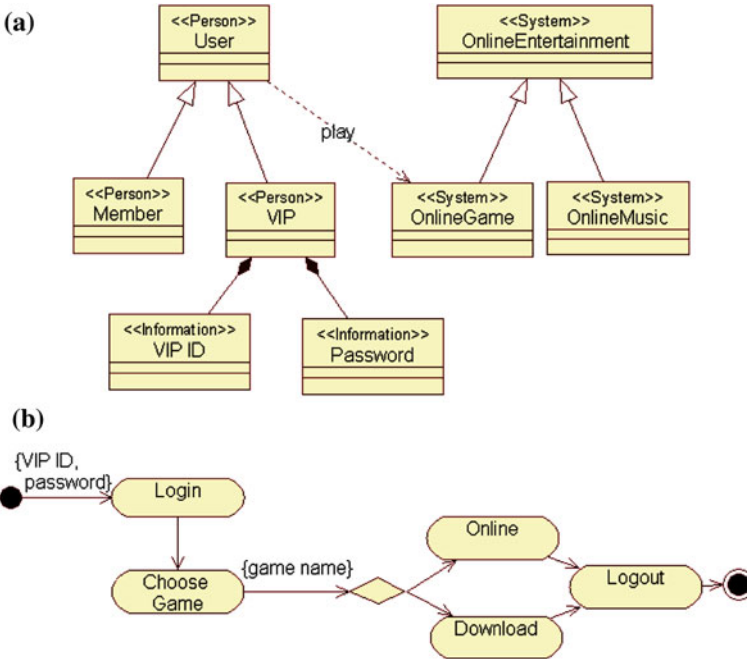


Fig. 107.2 The platform independent model

and domain ontology. As mentioned in Sect. 107.3, we design DSML as an embedded DSL based on UML. That means the transformation from CIM to PIM is almost straightforward. We have defined rules for the CIM-to-PIM transformation in Table 107.1.

The transformation result of Fig. 107.1 is shown in Fig. 107.2. This transformation approach bridges the gap between domain experts who know little about UML and system architects who have little domain knowledge. It will increase not only the modeling efficiency, but also the validity of domain model.

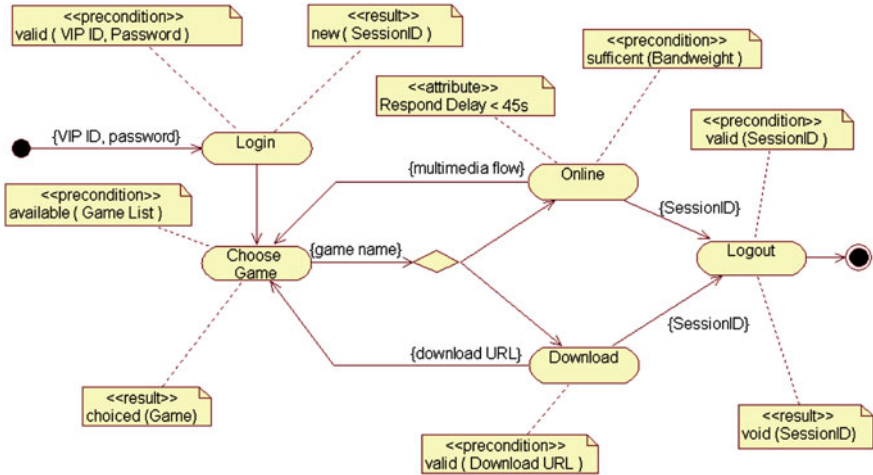


Fig. 107.3 Semantic annotation for UML activity diagram based on domain ontology

Table 107.2 Transformation rules for the PIM to PSM mapping

Rule	UML aspect	Description
TR_2.1	Activity diagram	Transform an activity diagram to a Service Process
TR_2.2	Activity	Transform each activity to one service, while transform an action to an atom service, i.e. atom process.
TR_2.3	ControlFlow	Transform ControlFlow to Control Construct in OWL-S
TR_2.4	Input pin	Transform to Input in OWL-S
TR_2.5	Output pin	Transform to Output in OWL-S,
TR_2.6	Comment	Transform comment to OWL-S respectively according to its content, such as category, IOPR, QoS, etc.

107.5 Mapping and Modeling in PSM Layer

UML is a standard modeling language for object-oriented system development, but not for service-oriented. In order to transform PIM into service-oriented PSM, we must resolve the key problem that how to import adequate semantic to UML models, which required for abstract description of services, including functional and non-functional properties. Considering the agility, we add the semantic annotations by UML profiles with domain knowledge provided by domain ontology, and the result of annotating for Fig. 107.2 is shown in Fig. 107.3. We use the UML activity diagram for example because it represents a system’s business and operation workflows, which is similar to business process in SOA.

There are several optional service-oriented modeling languages for modeling in PSM layer, such as WSMO, WSDL, etc., OWL-S [15] is our preference because it is supporting semantic modeling and reasoning.

```

(...)
<owl:Ontology rdf:about="">
(...)
<owl:Class rdf:ID="onlinegame">
<rdfs:subClassOf>
  <owl:Class rdf:ID="onlineentertainment"/>
</rdfs:subClassOf>
</owl:Class>
(...)
<process:CompositeProcess rdf:ID="OnlineGameProcess">
<process:composedOf>
(...)
  <process:Choice rdf:ID="Choice_03">
    <process:components>
      (...)
        <process:Perform rdf:ID="Perform_04">
          <process:process>
            <process:AtomicProcess rdf:ID="Online">
              <process:hasInput>
                <process:Input rdf:ID="MultimediaFlow">
                  </process:Input>
                </process:hasInput>
              </process:AtomicProcess>
            </process:process>
          </process:Perform>
        </process:components>
      </process:Choice>
    </process:composedOf>
  </process: CompositeProcess>
(...)
  <service:Service rdf:ID="LoginService">
    <service:supports>
      <grounding:WsdlGrounding rdf:ID="LoginGrounding">
        <service:supportedBy rdf:resource="#LoginService"/>
      </grounding:WsdlGrounding>
    </service:supports>
    <service:presents>
      <profile:Profile rdf:ID="LoginProfile">
        <profile:hasInput rdf:resource="#VIP_ID"/>
        <profile:hasInput rdf:resource="#Password"/>
        <profile:hasPrecondition rdf:resource="#Valid_VIP_ID"/>
        <profile:hasPrecondition
          rdf:resource="#Valid_Password"/>
      </profile:Profile>
    </service:presents>
    <service:describedBy rdf:resource="#Login"/>
  </service:Service>
  (...)

```

Fig. 107.4 Transformation results of PIM to PSM mapping (OWL-S ontology segment)

The transformation rules for mapping UML to OWL-S as shown in Table 107.2. Based on the semantic annotation of PIM model and the transformation rules for mapping PIM model to PSM model, the result of PSM modeling shows as Fig. 107.4, which is in OWL-S ontology. Because the limited of paper length, it only shows a segment.

107.6 Conclusion and Future Work

In this chapter, we propose an approach for architecting service-based software systems that embraces the MDA philosophy based on UML and UML profiles. The approach includes modeling and transforming models from CIM to PSM. We define domain ontology to support the whole process. A domain specific modeling language DSML is constructed to build domain models in computation independent layer. Then we modeling platform independent model by UML and transform UML model to OWL-S. Our approach offers an effective, reusable, flexible way to develop applications in heterogeneous and dynamic environments.

In the future, the main tasks include: how to transform the layered models in an automatic way, which is the original objective of MDA, how to generate executable code in platform dependent model. And the method of models checking and reuse, etc.

References

1. Arsanjani A (2004) Service-oriented modeling and architecture[R]. IBM technical report
2. Erl T (2005) Service-oriented architecture (SOA): concepts, technology, and design [M]. Pearson Education, Upper Saddle River
3. Object Management Group (2003) Model driven architecture guide. The Netherlands
4. Object Management Group (2009) Overview of the proposed model driven architecture to augment the object management architecture [EB/OL]. <http://www.omg.org/cgi-bin/doc?omg/00-11-05>
5. Pahl C (2007) Semantic model-driven architecting of service-based software systems[J]. *Inf Softw Technol* 49:838–850
6. Pahl C (2006) Ontology-based composition and transformation for model-driven service architecture[J]. *LNCS* 4066:198–212
7. Karsai G, Neema S, Sharp D (2008) Model-driven architecture for embedded software: a synopsis and an example [J]. *Sci Comput Program* 73:26–38
8. Kim IW, Lee KH (2007) Describing semantic web services: from UML to OWL-S[C]. In: *Proceedings of 2007 IEEE international conference on web services (ICWS 2007)*
9. Grønmo R, Jaeger MC, Hoff H (2005) Transformations between UML and OWL-S[C]. In: *Proceedings of first European conference on foundations and applications*, pp 269–283
10. Timm JTE, Gannod GC (2007) Specifying semantic web service compositions using UML and OCL[C]. In: *Proceedings of WWW2007*
11. Booch G, Rumbaugh J, Jacobson I (2005) *The unified modeling language user guide*, 2nd edn.[M]. Pearson Education, Upper Saddle River
12. Brahe S, Østerbye K (2006) Business process modeling: defining domain specific modeling languages by use of UML profiles[J]. *LNCS* 4066:241–255
13. Cuadrado JS, Molina JG (2009) A model-based approach to families of embedded domain-specific languages[J]. *IEEE Trans Softw Eng* 35(6):825–840
14. Mernik M, Heering J, Sloane AM (2005) When and how to develop domain-specific languages[J]. *ACM Comput Surv* 37(4):316–344
15. DAML (2006) OWL-S specifications. <http://www.daml.org/services/owl-s>

Chapter 108

Design of Three-Dimensional Garage Monitoring System Based on WinCC

Wenzhe Qi, Yanmei Yang, Lingping Shi, Yafei Liu and Jianjun Meng

Abstract This chapter puts forward the automatic three-dimensional garage monitoring system based on WinCC by analyzing the various problems in the automatic three-dimensional garage monitoring system and combining with actual project requirements. Applying to WinCC configuration software monitoring, data acquisition, configuration, development and opening characteristics, this system develops the automatic stacker parking system, and realizes the automatic monitoring system, fault alarming, data processing, user management and project safety, communication settings etc. This chapter describes the seamless connection process between WinCC monitoring program and Master of Siemens PLC programmable controller S7-300 and the constitution of the system's communication.

Keywords WinCC · Automatic three-dimensional garage · Monitoring system · PLC

W. Qi (✉) · Y. Yang · L. Shi · Y. Liu · J. Meng
School of Mechatronic Engineering, Lanzhou Jiaotong University, Lanzhou,
730070 GanSu, China
e-mail: qiwz@mail.lzjtu.cn

Y. Yang
e-mail: yym4459795@163.com

L. Shi
e-mail: shilingpingcl@163.com

Y. Liu
e-mail: 649776899@qq.com

J. Meng
e-mail: mengjj@mail.lzjtu.cn

108.1 Introduction

The automatic three-dimensional garage monitoring system is used to realize the real-time monitor of the running status of the warehousing and the outbound, the fault alarm and diagnose of the garage organization. This system gathers PLC, computer, communication technologies etc. and mainly includes data communication, condition monitoring, operation statistics, system safety management and PLC control and so on. The system monitors the scene parameter and the breakdown information of the garage organizations conveniently, promptly and intuitively. Windows Control Center (WinCC) is a kind of powerful industrial control software of the Siemens Corporation, and is “the true opening” HMI/SCADA software [1].

Combining with the actual project requires, and basing on WinCC configuration software, this chapter has developed the roadway stacking type automatic three-dimensional garage monitoring system. This monitoring system is advantageous to the enhancement of the automation degree of the garage. It has realized the unmanned management and “centralized management, distributed control [2]” process.

108.2 Automatic Three-Dimensional Garage

108.2.1 Working Principles

The main process of the garage’s work was divided into warehousing and the outbound. After the vehicle arrives at the entrance, the vehicle owner takes out the parking card which is registered beforehand, and scans it on the cardpunch, after it receives the parking card information which is sent by the cardpunch, the host computer inquires the database to check user record. If the information is right, the host computer calculates the parking stall where the vehicle will be stored at, and waits for the vehicle owner stopping vehicle to lifting platform. After the vehicle is driven into the lifting platform safely, the vehicle owner leaves the vehicle. The security detection system of the lifting platform checks whether the specific location of the vehicle is right or not, if the location is right, PLC carries out the parking stall instruction which is transmitted by the host computer, to control the stacker and the lifting platform to arrive in the parking stall. After sending the vehicle to the specific location, under the return instruction, the stacker and the lifting platform move to the exit, return to the waiting position, and wait the next instruction.

When vehicle owner wants to take out his vehicle, he can scan the parking card on the cardpunch. Then the cardpunch sends the information of the vehicle owner to the host computer. The host computer inquires the database to check out the vehicle’s location, and sends instruction to PLC. Controlled by PLC, the stacker and the lifting platform moved to the specific location to take the vehicle [3]. After

the vehicle location is checked by the detection system of the lifting platform is right, the stacker and the lifting platform which are controlled by PLC move to the exit, then send the vehicle out. And the vehicle owner drives the vehicle to leave the garage.

108.2.2 Monitoring System Constitutes

The system is suitable for the management of the automation three-dimensional garage, the real-time monitoring of the stereo memory and the transportation equipment, and the monitoring management of the warehousing and the outbound. The system includes the host computer monitoring system, the garage management system, database system, communication with PLC system, communication with MATLAB system and so on [4].

108.3 Design and Implementation on Monitoring System

108.3.1 The Design of Host Computer Software

The system's main functions: switching each monitoring screen; Concrete and comprehensive dynamic demonstration; Communication with the lower computer PLC; Security; Information hinting; Equipment alarm.

The system's host computer uses Siemens Corporation's Wincc6.0 configuration software to accomplish the design of control subsystems, such as the monitoring of storage operations, the communications with the lower computer, equipment alarm, report printing and so on.

108.3.1.1 The Abstraction of System Model

Automated parking system models abstract entirely from the parking example. The formulation of a regional sense based on a rectangular Cartesian coordinate system, each abscissa is allocated from 0, and the vertical axis is allocated from top to bottom. Correspondence between the address and parking spaces are shown in Table 108.1.

Distributing the address of automated parking garage is actually a result of model abstract. The structure of the garage totally abstracted into a four-dimensional array that stands for garage building structure and operations. Automated parking control system operating on the four-dimensional array reflects the true stereo distribution of garage parking space. The whole process then displays on the screen, combined with the corresponding display module.

Table 108.1 Correspondence between the address and parking spaces

Parking	Specific parking
1-2-0-2	the 2nd parking of area 1 row 2 floor 0
2-1-3-3	the 3rd parking of area 2 row 1 floor 3
4-1-2-2	the 2nd parking of area 4 row 1 floor 2

108.3.1.2 The Program of Model Transformation

Throughout the whole project establishment, the core part is to program and achieve the required function of the monitoring system, namely the model transformation program. The following is the monitoring program of passing instructions.

```
#include "apdefap.h"
void OnLButtonDown(char* lpszPictureName, char* lpsz-
ObjectName, char* lpszPropertyName, UINT nFlags, int x,
int y)
{GetTagByte("state3"); //Returns the current state
GetTagByte("z3"); //Back to the current direction of z
GetTagByte("y3"); //Back to the current direction of y
GetTagByte("x3"); //Back to the current direction of x
if(GetTagByte("state3") ==1&&GetTag-
Byte("z3") ==1&&GetTagByte("y3") ==0&&GetTag-
Byte("x3") ==16)
SetBackColor(lpszPictureName, "rectangle
93", CO_YELLOW);
else
if(GetTagByte("state3") ==0&&
GetTagByte("z3") ==1&&GetTagByte("y3") ==0&&GetTag-
Byte("x3") ==16)

SetBackColor(lpszPictureName, "rectangle
93", CO_CYAN);
```

Each rectangular in the graph is a specific number block, so the program is mapped to the corresponding number of rectangular blocks. As shown in Fig. 108.1.

108.3.2 The Implementation of Monitoring System

The warehousing and the outbound process shows in Fig. 108.2.

Monitoring system model uses a four-dimensional array to express the physical structure of the garage. Operations of the array reflect the true distribution of garage parking spaces, which will be displayed on the screen, combined with

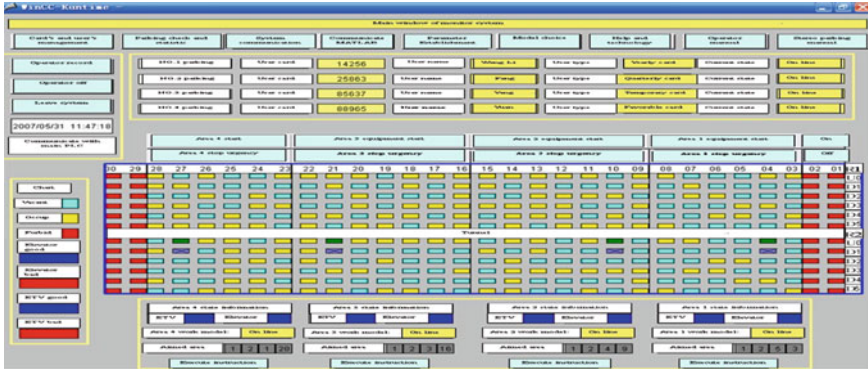


Fig. 108.1 The main interface of monitoring system

corresponding display module. As shown in Fig. 108.1, in this system, the garage is divided into two rows, each row has four work regions and two spare regions, each region is divided into one ground floor, five underground floors, and each region is also divided into six columns.

108.4 Communication Realization

108.4.1 Control Panel Setting

The communication types of WinCC and SIMATIC S7 are MPI [5], industrial Ethernet, PROFIBUS, TCP/IP. Combine the actual needs of project, WinCC and PLC communicate by MPI in the automatic three-dimensional garage monitoring system. The main steps of setting:① Double-click the “Set PG/PC Interface\” in Control Panel and select the MPI (WinCC)—PC Adapter (MPI). ② Setting properties of the PC Adapter.

108.4.2 Setting of WinCC

Main steps: (1) Add the driver: In this system, add SIMATIC S7 Protocol suite to this item. (2) The chosen protocol of configuration: set up the address of “System Parameter” and S7 station. (3) The joint parameter of the configuration: Establish the logic combined parameter of WinCC and PLC port. (4) Add variable tag group and variable tag in the connection: Produce variable tag group and variable tag. (5) Inspect the state of the connection while running [6]: To confirm if the connection between WinCC and PLC is effective, the user can press the “RUNTIME” bond in the tool bar of the Control Center, and inspect if the connection is successful by Status of Driver Connections, OK means the connection is good, and

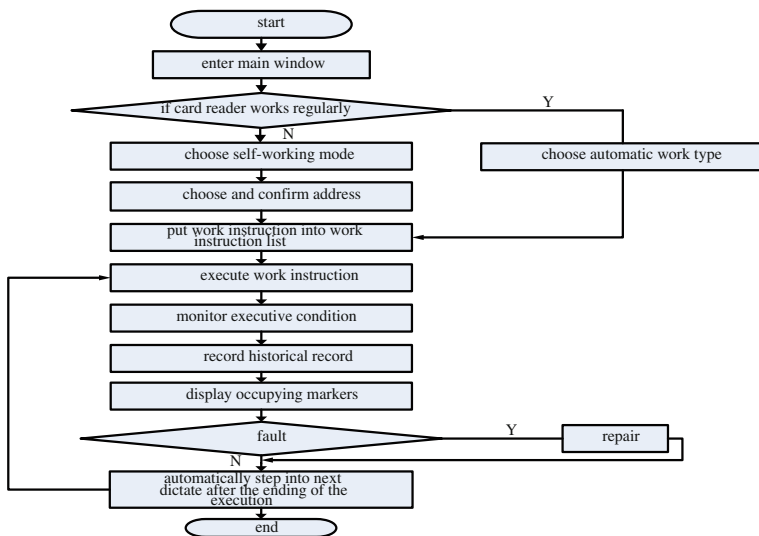


Fig. 108.2 The flow of in and out garage

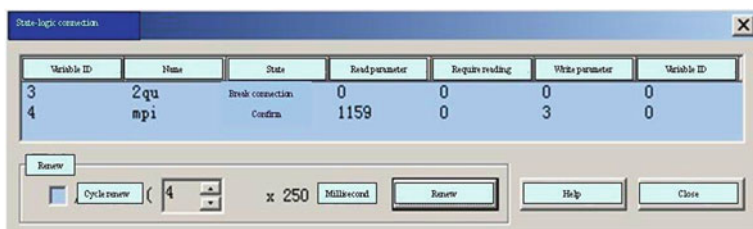


Fig. 108.3 The driver connection

Disconnection means not successful. Click the checking box of “Cyclic Update” to ensure the renovation of the data periodicity. The driver connection state is showed in Fig. 108.3.

The logic connection named mpi and PLC had been under communication, so we can read and write the variable in the monitoring program and PLC.

108.4.3 The Connection and Realization of the System Communication

To realize the communication between WinCC and PLC, we used communicative driver program to connect data manager and PLC. The communication process structure between WinCC and PLC is showed in Fig. 108.4.

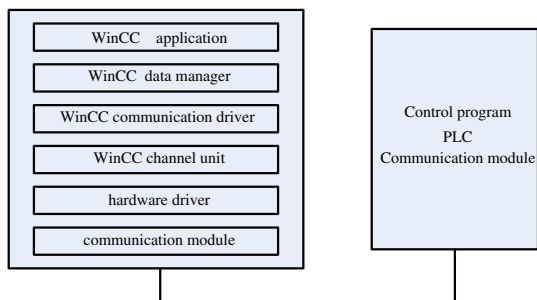


Fig. 108.4 The communication process structure

The application program of WinCC asks data from data manager. The data manager settles the data which is produced by WinCC and the data that is stored in the SQL anywhere system database. According to the connective PLC, the communication driver program is a or many channel units, communication driver program forms the interface between WinCC and processing manager by its channel units, and it transmits the requesting information by the communication processor, communication processor returns the requesting process value in the corresponding answer to WinCC. The channel unit supports hardware component which is needed in the communication, by the operating system interface and hardware driver program. We can access the process data in the PLC by WinCC.

108.5 Conclusion

The automatic three-dimensional garage monitoring system based on WinCC supports the industrial Ethernet and PROFIBUS and so on, which has great Real-time and openness and provides the necessary field information of garage for the operators. This system can simplify the communication between the host computer and PLC and read data from PLC and write data into PLC, and deploys vivid monitoring picture, in addition, it is beneficial for the safety of the program running, so this system can make the roadway stacking type automatic three-dimensional garage monitoring system run more stably, and reduce the project time, manpower and material resource consumption, so this system is highly applied and popularly valuable.

Acknowledgments This work has been supported by the Research Project of Postgraduate Tutor of Gansu province college (No.1004-12) and the ‘Qing Lan’ Talent Engineering Funds of Lanzhou Jiaotong University (No.QL-05-14A). We appreciate all people who provide us physical and spiritual support.

References

1. The automation and actuation group of Siemens limited company (2003). Explain the profound in simple language Siemens WinCC V6 [M] Beijing astronautics aviation university publishing company, Beijing.
2. Chunyuan B, Shuangyan R, Yongkui M (2007) et al S7-300/400 PLC directory of practical exploitation [M]. Mechanism Industry Publishing Company, Beijing
3. Shouming Z, Xunzhou H et al (2003) Multi-Storey lifting and shifting type of three-dimensional garage and control. Mechanical or parking equipment
4. Manli C, Mingxing L (2006) Monitoring system of the three-dimensional garage. J. Instrum 27: 559–560
5. Xuesong Z, Mi T, Youjie M et al (2009) The multi-layered cyclic type system of three-dimensional garage control system of J hoisting and conveying machinery. 5: 58–60
6. Jianjun M, Zeqing Y, Chenghui Y (2006) Simulation system of automated three-dimensional garage based on configuration software. In: The 6th world congress on intelligent control and automation, DaLian, 6198–6202

Chapter 109

A Hybrid Modeling Method for Service-Oriented C4ISR Requirements Analysis

Ying ZHANG, Xiaoming Liu, Zhixue Wang and Li Chen

Abstract With the rapid growth of complexity as well as the continuously innovational concept of network-centric warfare, it is difficult to analyze and design C4ISR systems based on the existing component technology or the object-oriented technology. However, both the workflow method and the AI planning method for service modeling are not very suitable for analysis and design service-oriented C4ISR requirements. This paper imports Service-Oriented Computing (SOC) design paradigm to C4ISR requirements analysis, and proposes a hybrid modeling method based on multi-ontologies. Accompanied with an instance, the method provides a new, flexible, reusable solution for C4ISR requirements evolution and system reconstruction.

Keywords SOC · C4ISR · Service modeling · Ontology

109.1 Introduction

Service modeling technology is one of the hot topics and has received significant attentions in SOC [1–4]. There are many researches proposed various methods, such as Refs. [5–13]. These methods could be divided into two kinds [14]. The one kind is based on the workflow technology. This method decompose the business process into activities, roles, rules and processes in order to integrate the enterprise comprehensive resources automatically or semi-automatically. The notable

Y. ZHANG (✉) · X. Liu · Z. Wang · L. Chen
Institute of Command Automation, PLA University of Science and Technology,
Nanjing, 210007, China
e-mail: zhywl66@163.com

advantage of this method is modeling and compositing services with high efficiency. Another kind is AI planning method. Dynamic and flexibility are the notable advantages of this method.

However, those two kinds of service modeling are all not suitable for analysis and design service-oriented C4ISR (Command, Control, Communications, Computers, Intelligence, Surveillance and Reconnaissance) [15, 16]. Because the workflow method cannot satisfy the dynamic reconstruction and reconfiguration of C4ISR systems, while the AI planning method cannot deal with such huge scale and complex systems and application. In order to resolve this problem, we proposed a hybrid method combining the two methods mentioned above for C4ISR requirements analysis. The method could support the high-layer analysis by process reuse and low-layer analysis by dynamic service integration. It also support the reuse of models and services by domain knowledge based on multiple ontologies. This method provides a more efficient and flexible method for service-oriented C4ISR systems requirements analysis and modeling.

The rest of this article is organized as follows: Sect. 109.2 introduces the ontology definitions for service-oriented C4ISR requirements analysis. Section 109.3 presents the modeling process in detail with an instance. Section 109.4 draws some conclusions and future work.

109.2 Ontology Definitions

There are three reasons of defining multi-ontologies in this paper. Firstly, it is used to give a common semantic among domain experts. Secondly, it is needed to increase modeling efficiency by domain knowledge reuse. At last it could support model checking by logical reasoning.

Domain ontology is used to capture domain specific concepts and relations, which is defined as follows.

Definition 1 Domain Ontology Domain Ontology = $\langle \text{DomConcept}, \text{DomRelation}, \text{DomRule} \rangle$

in which DomConcept is a finite set of domain specific concepts, while DomRelation is a finite set of domain specific relations associated with a pair of the concepts. DomRule is a finite set of domain model constraints specified in Description Logics [17]

Domain ontology forms domain knowledge which is related to common knowledge in a specific domain. It is the guide for constructing application ontology. We define application ontology as follows.

Definition 2 Application Ontology Application Ontology = $\langle \text{AppConcept}, \text{AppRelation}, \text{AppRef}, \text{AppRule} \rangle$

in which AppConcept is a finite set of application concepts while AppRelation is a finite set of application defined relations. AppRef is a mapping function, specifying the mappings from AppConcept to Domconcept or from AppRelation

to DomRelation. AppRule is a finite set of application model constraints specified in Description Logics according to the rules of DomRule.

Service-oriented C4ISR requirements analysis and modeling regard service as the basic element to support various application of high system. Services could represent a variable logical scope and scale because its concerns could be large or small [18]. In order to support different granularity of service discovery and composition, this paper use activity ontology for coarse-grained modeling while service ontology for fine-grained modeling. We define activity ontology as follows.

Definition 3 Activity Ontology Activity Ontology = $\langle \text{ActConcept}, \text{ActRelation}, \text{ActFunction}, \text{ActAttribute}, \text{ActRef}, \text{ActRule} \rangle$

in which AppConcept is a finite set of activity-related concepts, such as the name of an activity, sub-activities of an activity, etc. AppRelation is a finite set of activity relations, used to represent the relationship between activities, it also include hasSubAct and realized by relation. The former is used to represent the relationship between an activity and its sub-activity, while the latter represents the relationship between an activity and services contribute to realize it. ActFunction = $\text{Input} \cup \text{Output} \cup \text{Precondition} \cup \text{Effect} \cup \text{Attribute}$ is a finite set of activity attributes. Input is the input parameter set of activity, and Output is the output parameter set of activity. Precondition is the prerequisite set of activity, while Effect is the result set of activity. Attribute is a finite set of activity attributes. ActRef is a mapping function, specifying the mappings from activity ontology to activity concept in domain ontology or application ontology. ActRule is a finite set of activity constraints specified in Description Logics according to the rules of AppRule or DomRule.

Service is the instance of service ontology which is a reusable assets published in service repository. A service usually has at least one function and can be reused in a variety of domain. We define service ontology as follows.

Definition 4 Service Ontology Service Ontology = $\langle \text{ServiceName}, \text{ServiceFunction}, \text{ServiceRef} \rangle$

in which ServiceName is the name of a service. ServiceFunction = $\text{Input} \cup \text{OutPut} \cup \text{Precondition} \cup \text{Effect} \cup \text{Attribute}$ is functional description of the service. Input represents the input parameter set of the service, Output is the output parameter set of the service. Preconditions represent prerequisite set for service implementation. Effect is the effect set of the service. Attribute is a finite set of service attributes, generally used to describe the QoS properties of service. For C4ISR systems, the main considerations of QoS are reliability, availability, safety, timeliness, accuracy and so on. ServiceRef is a mapping function that maps service ontology to a service concept in activity ontology, or a service (or activity) concept in domain ontology.

109.3 Modeling Process

In this paper, we introduce a hypothetical instance of C4ISR systems to illustrate the service-oriented C4ISR requirements modeling. The instance is belonged to military logistics domain, and the specific application is maintaining a broken-down tank. Regarding domain ontology as a shared ontology, accompanied with application ontology, activity ontology and service ontology construct the domain knowledge repository for modeling. The modeling process consists of four phases, as follows.

Phase 1, capturing the domain knowledge that means domain conceptual modeling. Domain ontology defines all concepts and relations within a specific domain. Therefore, domain experts firstly need to define domain ontology when modeling domain model. In this paper, we use UML as the modeling language and example of the logistics domain model (fragment) is shown in Fig. 109.1. In the whole modeling process, all used concepts and relations should be within domain knowledge. If there are necessary concepts and relations are not defined in the domain knowledge repository, domain experts should add them to the repository.

Phase 2, creating the initial application model. Define the application ontology under the guide of domain ontology for the specific application. Application ontology describes concepts and relationships for specific application in domain. It is the static reusable asset of domain knowledge. Therefore, the application model also should be stored in the repository after built. In the future modeling process, modeler should firstly find whether there is available application model in domain knowledge repository or not, in order to improve the modeling efficiency by reuse. The initial application model for maintaining a broken-down tank is shown in Fig. 109.2. We use UML stereotype represent the mapping relationship between application concepts and domain concepts. For instance, the application concept “ThreeHours” is mapping from the domain concept “DesiredFinishedTime”, and so on.

Phase 3, for each activity established in the initial application model, analyze and modeling activity process. Activities are divided into two kinds: simple activities and complex activities. The former is performed by an atomic activity while the latter is performed by several atomic activities. Based on activity ontology, the atomic activity is an activity that has no sub-activity. The detail steps are as follows: use Algorithm 1 (as shown in Fig. 109.3) in the activity repository to find whether there is existing activity process that could be reused this time or not. If there is, modeler could reuse the activity process directly. Otherwise, the modeler could modify existing activity process according to specific application requirements then reuse, or model a new activity process to supplement and improve the repository.

Phase 4, creating the final application model by matching the atomic activity and services. In this paper, we use activity represents coarse-grained modeling while service represents fine-grained modeling. Activity is implemented by activity process which consists of atomic activities. And atomic activity is realized

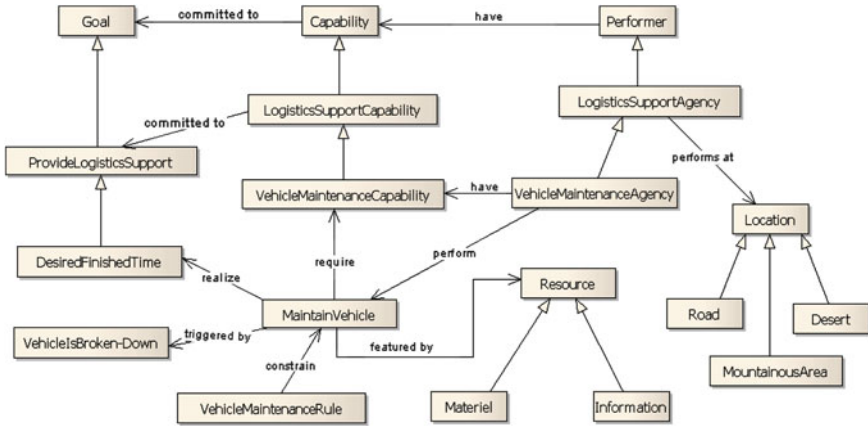


Fig. 109.1 The logistics domain model (fragment)

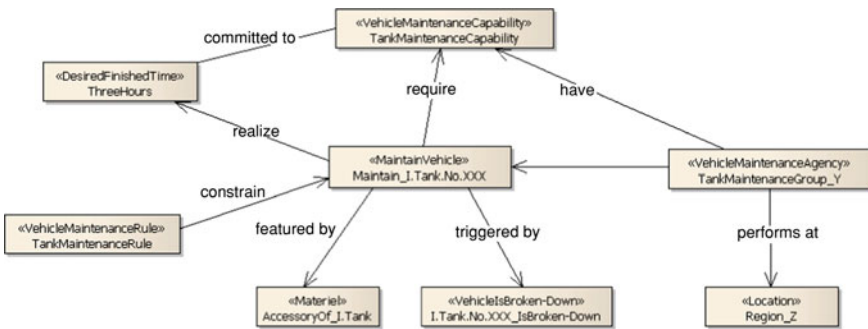


Fig. 109.2 The initial application model of maintaining a broken-down tank

by service which could be single service or services composition. Therefore, this phase is including service discovery and composition.

Because the concepts and relations of activity ontology and service ontology are all included in the same domain knowledge repository, we could regard atomic activity as service requester. The activity function describes inputs, outputs, pre-conditions and effects of an activity could be regarded as service request description. Then according to the service request, discovery in the service repository if there are services could satisfy the request or not, i.e. service discovery.

Semantic similarity is an important facet of service discovery. We divide service matching into three aspects: name matching, inputs/outputs matching and attributes matching. Inputs and outputs matching is the keyword item among all aspects. Based on the shared ontology, i.e. domain ontology, locate concepts of

ALGORITHM 1 Activity Process Construction Algorithm

INPUTS: activity set $AppActSet$ which is modeled in initial application model

OUTPUTS: the atomic activities set $APSet$ which implement the activities in $AppActSet$

INITIALIZATION: load the activity repository of domain knowledge repository, $APSet = NULL$, let the current activity matched atomic activities set $AP = NULL$, i.e. $AP.ActSet = NULL$, $AP.RelSet = NULL$.

ALGORITHM STEPS:

Step1: Select the first activity in $AppActSet$ as the current activity $Activity$

Step2: Retrieve the $Activity$ by its name in activity repository to find it is included or not. If not, do Step3. Otherwise, do Step4.

Step3: There is no reusable activity process for $Activity$ in current activity repository. Therefore, modeler need to build new concepts, relations, attributes, rules of $Activity$ based on activity ontology definition, then do Step7.

Step4: There is reusable activity process in current activity repository for $Activity$. Check whether $Activity.hasSubAct$ is empty or not. If it is empty, do Step5. Otherwise, do Step6.

Step5: The current activity $Activity$ is a simple activity, add $Activity$ to $AP.ActSet$, then do Step7.

Step6: The current activity $Activity$ is a complex activity, and it has reusable activity process, add all sub-activities of the process to $AP.ActSet$, while add the corresponding relations between sub-activities to $AP.RelSet$, then do Step7.

Step7: Add AP to $APSet$, complete the processes of current activity. Then select the next activity in $AppActSet$ do Step2 again, until process all activities in $AppActSet$

Step8: Return $APSet$

Fig. 109.3 The activity process construction algorithm

activity ontology and service ontology in domain ontology, then compute the semantic similarity between the concepts in domain ontology. Currently, there are many domestic and foreign scholars have researched various methods for computing semantic similarity, which is out of this paper's scope. We use the computing algorithm proposed in Ref. [19], the semantic similarity between concept A and concept B is calculated as follows, shown in Eq. 109.1. Max represents the max distance from the root node to any leaf node. Len(A,B) represents the shortest distance between concept A and concept B.

$$sim(A, B) = \frac{2Max - len(A, B)}{2Max} \quad (109.1)$$

The detail matching method is as follows. For each atomic activity, use the service composition algorithm based on the canonical structure, which proposed

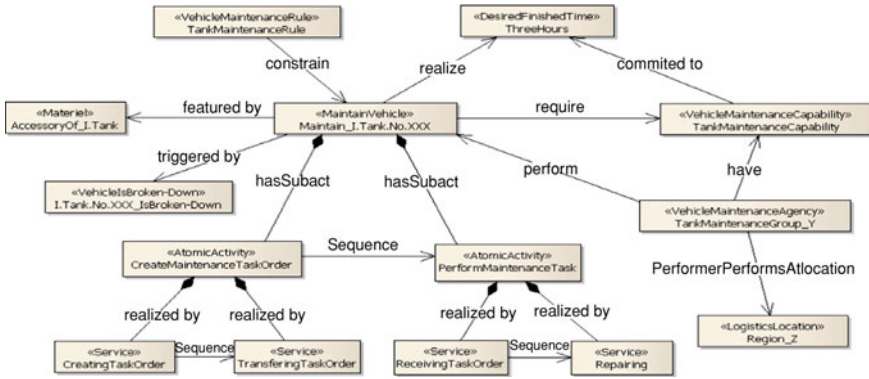


Fig. 109.4 The final application model of maintaining a broken-down tank

by our team members in previous work [20], to find the matched service set which could implement the atomic activity. If the matching is fail, then clue the modeler either modify the inputs and outputs then matching one more time or define a new service. Otherwise, if the matching successfully with only one service set, it is just the matched service we need. Or, if there are more than one service sets when matching is successful, which means there are several optional solution for performing the activity, thus the models could choice the one which as the biggest value computed by Eq 109.2, or choice by his opinion.

$$\begin{aligned}
 \text{Similarity}_{aux}(\text{Act}, \text{Ser}) = & w_1 * \text{Sim}_{Name}(\text{Act.Name}, \text{Ser.Name}) \\
 & + w_2 * \text{Sim}_{Fun}(\text{Act.Precondition}, \text{Ser.Precondition}) \\
 & + w_3 * \text{Sim}_{Fun}(\text{Act.Effect}, \text{Ser.Effect}) \\
 & + w_4 * \text{Sim}_{Fun}(\text{Act.Attribute}, \text{Ser.Attribute}) \quad (109.2)
 \end{aligned}$$

in which $w_i, 1 \leq i \leq 4$ is the weight. $\text{Sim}_{Name}(\text{Act.Name}, \text{Ser.Name})$ represents the semantic similarity between service names, $\text{Sim}_{Fun}(\text{Act}, \text{Ser})$ represents the semantic similarity between service function that including precondition similarity, effect similarity and attribute similarity, which are described respectively as $\text{Sim}_{Fun}(\text{Act.Precondition}, \text{Ser.Precondition})$, $\text{Sim}_{Fun}(\text{Act.Effect}, \text{Ser.Effect})$, and $\text{Sim}_{Fun}(\text{Act.Attribute}, \text{Ser.Attribute})$.

At last, composite all the services into the initial application model, then obtain the final application model with service as the basic granularity, as shown in Fig. 109.4. It is noteworthy that the application model must consistent with the domain model. We ensure the correctness and consistency by ontology reasoning technology and model checking with the Description Logic [21].

109.4 Conclusions and Future Works

This paper studies how to apply SOC with its reusable, flexible, loose couple features to C4ISR capability analysis such as C4ISR systems. We propose a hybrid method for service-oriented C4ISR modeling based on multiple ontologies. Service-oriented C4ISR modeling process focuses on the logical modeling rather than the technical details of application implementations. Therefore, if the higher requirement has changed, the modeler can reconstruct the application model flexibly and fast through invoking and integrating services. In addition, this method supports updating services that the new and better services could replace the old ones without influencing the high-level model.

The future work will include supplement the domain knowledge repository in order to improve the efficiency of reuse, the validation methods to ensure the consistency of the modeling better, as well as study how to map the model to code generation in the system design phase under the guide of domain knowledge.

References

1. Huhns MN, Singh MP (2005) Service-oriented computing: key concepts and principles [J]. *IEEE Internet Comput* (244):021–028
2. Benatallah B, Motahari Nezhad HR (2005) Service oriented computing: opportunities and challenges, LNCS, vol 3372, pp 1–8
3. Papazoglou MP, Traverso P, Dustdar S et al (2008) Service-oriented computing: a research roadmap [J]. *Int J Co-op Inf Syst* 17(2):223–255
4. Papazoglou MP, Schmidt JW, Mylopoulos J (2009) *Service-Oriented Computing* [M]. MIT Press Cambridge
5. Wu B, Jin Z, Zhao B (2008) A modeling approach for service-oriented application [C]. In: *Proceedings of the IEEE international conference on web services*
6. Lee SY, Lee JY, Lee BI (2006) Service composition techniques using data mining for ubiquitous computing environments. *Int J Comput Sci Netw Secur* [J] 6(9):110–117
7. Ma Y, Jin B, Feng Y (2005) Dynamic discovery for semantic Web services based on evolving distributed ontologies [J]. *Chin J Comput* 28(4):603–615 (in Chinese with English abstract)
8. Qiu L, Shi Z, Lin F (2006) Context optimization of AI planning for services composition [C]. In: *Proceedings of the IEEE international conference on e-business engineering*, pp 610–617
9. Bottaro J, Bourcier C, Escoer et al (2007) Autonomic context-aware service composition [C]. In: *Proceedings of 2nd IEEE international conference on pervasive services*
10. Mingkhwan P, Fergus O, Abuelma'Atti et al (2006) Dynamic service composition in home appliance networks. *Multimedia Tools Appl* [J] 29(3):257–284
11. Viroli M, Denti E, Ricci A (2007) Engineering a BPEL orchestration engine as a multi-agent system [J]. *Sci Comput Program* 66:226–245
12. Chi YL, Lee HM (2008) A formal modeling platform for composing web services [J]. *Expert Syst Appl* 34:1500–1507
13. Valero V, Emilia Cambronero M, Díaz G et al (2009) A petri net approach for the design and analysis of Web Services Choreographies [J]. *J Log Algebraic Progr* 78:359–380
14. Deng S (2007) Research on automatic service composition and formal verification [D] (in Chinese). Ph.D thesis of Zhejiang University

15. Liao S, Sun B, Wang R (2003) A knowledge-based architecture for planning military intelligence, surveillance, and reconnaissance [J]. *Space Policy* 19(3):191–202
16. US Department of Defense. DoD Architecture Framework Version 2.0 [R]. 2009
17. Baader F, Calvanese D, McGuinness DL et al (2003) *The Description Handbook* [M]. Cambridge University Press, Cambridge
18. Erl Thomas (2005) *Service-oriented architecture (SOA): concepts, technology, and design* [M], Pearson Education
19. Resnik P (1999) Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language [J]. *J Artif Intell Res* 11:121–126
20. Chen L, Song Z, Zhang Y et al (2010) A method of semantic web service composition based on the canonical structure[J]. *J Inf Comput Sci* 7(2):463–469
21. Dong Q, Wang Ze, Chen J et al (2010) Method of checking capability model based on description logic [J]. *Syst Eng Electron* 32(3):533–539 (in Chinese with English abstract)

Chapter 110

Study on Normalized Operation of Internet Drugs Market

Feng Chen, Ying Lu, Bin Dun and Ming Lu

Abstract Currently, the problem of some websites provide medicines information service and drug transaction services illegally is increasingly prominent. This chapter makes domestic normalized operation conditions of the Internet drug market as research object, reviews the current laws and regulations of domestic Internet drug market and regulation status in quo, analyzes the problems in normalized operation of Internet drug market, and discusses the countermeasures from five aspects below: perfect the law and regulations, adopt high-tech means, take appropriate measures to protect consumers' privacy, strengthen the management of medicines from international online pharmacies and do publicity work well, which is expected to have certain theoretic and practical significance to normalized operation of our country's Internet drugs market.

Keywords Internet drug market · Normalized operation · Regulate

This work is supported by the Key Projects in the National Science and Technology Pillar Program during the Eleventh Five-year Plan Period under Grant No.2009BAH48B01.

F. Chen (✉) · Y. Lu · B. Dun
Information Center of State Food and Medicine Administration,
Beijing, Peoples Republic of China

M. Lu
Beijing ITOWNET Cyber Technology Ltd, Beijing, Peoples Republic of China

110.1 Introduction

As the adoption of Internet in drugs operation has the advantages of information disseminating fast, trade convenient, operation cost low, etc. now the online business of drugs has become an important channel of medicines business. But drugs are different from other commodities; drugs are a kind of special commodity used to treat the illness to save patients. However, the problem that some websites provide drugs information services and drug transaction services illegally has become increasingly prominent. Some websites sell adulterant openly online, which is serious danger to public safety and has attracted great attention in China. Therefore, it is essential to take certain measures to regulate Internet drug market, normalize its business to ensuring people's safety. At present the quantity literature about the Internet drugs market regulation and operation normalization is obviously low; the research is relatively backward; the regulation of related administration departments in our country to the Internet drugs market is also at the exploration stage; the related supervisory system is still not incomplete. However the Internet drugs market supervision level and the perfection degree of related laws and regulations is the key to normalization of Internet drugs market. This chapter is just based on the present status that academic circles rarely study on normalized operation of Internet drugs market, and carries out deeper research on Internet drug market regulation status in quo, the existing problems in normalized operation, and proposes the countermeasures which is expected to have certain theoretic and practical significance to normalized operation of our country's Internet drugs market.

110.2 Regulation Status in Quo of Internet Drugs Market

110.2.1 Current Laws and Regulations

Our country has strict supervision system for Internet drugs information services and transaction behaviors. Current laws and regulations of our country are shown in Table 110.1. Internet information service management method Stipulates explicitly: the country implements the licensing system to the managerial Internet information service; implements on file system to the non-managerial Internet information service. Enterprises that have not obtained the permission or fulfill registration procedure are not allowed to be engaged in the Internet information service. Similarly, regarding the Internet drugs information service and the drugs transaction service behaviors, State Food and Drug Administrator (SFDA), as corresponding staff function of the State Council, also according to relevant laws and regulations such as Drug Administration Law of the people's Republic of China, Internet information service management method and so on, enacted Internet drugs information services management method, temporary provisions of Internet drugs transaction service examination and approve etc. for strict regulation [1].

Table 110.1 Current laws and regulations

Laws and regulations	Date and promulgation department
Internet information service management method	Decree of the State Coucile (No. 292)
Internet drugs information services management method	Decree of SFDA (No. 9)
Notice about related problems of implementation of Internet drugs information services management method	SFDA No.340
Temporary provisions of Internet drugs transaction service examination and approve	SFDA No. 480
Notice about related problems of implementation of temporary provisions of Internet drugs transaction service examination and approve	SFDA No.515
Supplemental notice about related problems of implementation of temporary provisions of Internet drugs transaction service examination and approve	SFDA No. 82
Guidance of Department of Commerce on the Internet trading	Department of Commerce announcement No.19

The regulations strictly define the access provision of Internet drug information service and drug transaction services, and also stipulate drug information provides by Internet drug information service websites must be scientific and accurate, and must conform to national laws, regulations and relevant rules about State drug management, and clearly defined the cases in which drug information is not allowed to publish. Temporary provisions also limit drugs transaction behaviors of enterprises that provide drug transaction services: enterprises can only operate drugs produced by themselves; for individual consumers, an enterprise can only sell non-prescription drugs operate by itself; medical institutions can only purchase drugs, but cannot sell drugs on the Internet.

Approval enterprises will obtain the State food and Drug Administration the Internet drugs transaction service qualification certificate issued by SFDA, valid for 5 years, which include: company name, legal representative, the address and zip code, IP address, network name, certificate number, scope of services, certification authority, and certificate validity. Qualified enterprises need mark qualification certificate number of Internet drugs transaction service organization in a prominent position of its website homepage.

110.2.2 Regulation Status

In order to carry out the analeptic regulation work deeply, SFDA also carries out supervision and inspection on illegal publication behaviors of analeptic information, mainly strengthens supervision and inspection on the behaviors of publishing the protein anabolic preparation and peptide hormones information and providing online ordering services through the websites without approval, treats this as a gradual special task, which has achieved initial results [2]. In order to unify the

Internet drug transaction service acceptance standards, normalize inspection program on the spot, guarantee the quality of acceptance work, SFDA made Internet drug transaction services acceptance and evaluation criteria on the spot and its implementing rules according to temporary provisions, evaluating the enterprise from software and hardware respectively, and providing credentials to the enterprises that meet the criteria.

Besides, in order to provide objective evidence for judging whether an enterprise software system meets relevant indices prescribed in Internet drug transaction service agency acceptance criteria, and to provide reference for evaluating enterprises and drug regulatory agency, SFDA also made Internet drug transaction service system software evaluation framework. This evaluation framework is made by SFDA in conjunction with China Software Evaluation Center, the quality supervision and inspection center of national application software product, mainly making assessment on the basic functions of Internet drug transaction service system software and the quality of relevant software [3].

110.3 Problems in Normalized Operation of Internet Drugs

110.3.1 Related Laws and Regulations of Internet Drugs Transaction are Imperfect

Because the Internet is virtual and opening, in the regulation of Internet drug transaction has emerged some new problems below, for which current laws and regulations do not make strict stipulation. The entity of the object investigated cannot be identified. As a result of network identity virtualization; it is very difficult to ascertain the real name, address, etc. of the transaction entity. Competence is hard to confirm. The competence range of Internet drugs transaction includes the location that IP address belongs to, illegal operation location, drug delivery location, and the location that damage results are made [4]. The gathering and effectiveness of electronic evidence are not enough. Tortious amount of the vast majority of Internet drug transaction is less, not yet reach criminal standard. It is very difficult for executive authorities in collecting and finding the transaction records. They have no way to give the illegal sale of medicines on the Internet the corresponding sanctions.

110.3.2 Monitoring Methods of Internet Drugs Information

Currently, along with the fast development and application of Internet information technology, in the drugs circulation field, providing drugs information

service and drugs transaction service through the Internet increases day by day. However, the problem that some websites provide drugs information services and drug transaction services illegally has become increasingly prominent, but the monitor method used at present mainly depends upon the man-power, retrieving according to key words by search engine, then inspecting search results one by one. This method cannot monitor drugs information of websites, the network advertisements delivery channels, and advertisement operation business of search engine dynamically. Not only the efficiency of this monitor method is low, but also has omission, so it cannot achieve the effective monitor goal.

110.3.3 Lacking of Consumer Privacy Protection Measures

Before the consumers enter a website to purchase products, the website usually require consumers to register and fill in some personal information, which may cause the following problems. On-line pharmacy is negligent in the management regarding consumers' individual material. Because on-line pharmacy does not have perfect network management mechanism or does not pay more attention on registered users' individual privacy security, consumers' individual material is often betrayed. And what is more, some network pharmacy personnel carries on the information transaction with the illegal information dissemination enterprises, which inevitably create the hidden danger to the consumers' individual information security. How to guarantee consumers' individual privacy security is an important problem that the development of Internet drugs profession is confronted with in our country.

110.4 Countermeasures

110.4.1 Legal System of Internet Drug Transactions Regulation

Amend relevant laws, regulations, and effectively solve the laws and regulations imperfect problem. Top priority, the competence range of Internet drugs transaction, the effectiveness of electronic evidence, information share of regulation agency and collaborative disposition, conditions and procedures of close a website, a site with a certain amount of complaints accumulation and the same type of fraud should be transferred to public security organ and provided with detailed prescriptions [5].

110.4.2 High-Tech Means for Internet Drug Transaction Regulation

Regulating Internet drug transactions with high-tech means is more in line with the digitization and systematization development tendency of regulation. For legitimate online pharmacies, mark legitimate approval certificate number of SFDA in the form of a link in a prominent position on its website, so that consumers can click on the logo to determine its legality [6]. Accelerate construction of regulation platform for Internet drug transaction, which should meet SFDA's requirements of day-to-day regulating all domestic enterprises that work on Internet drug transaction. All levels of drug regulatory agencies should be able to enter the online pharmacy transaction directly through that platform, implement real-time monitoring on the whole online drug transaction process, and develop a valid network precaution mechanism. Once you find some problem in the drugs transaction process, you can alert trading parties through this platform in time, and interrupt the transaction, extract relevant evidence to provide a basis for future processing. In the process of implementing the regulation, the platform should take effective measures to guarantee its security, so as not to betray the trade secrets of both sides. In addition, the platform should be able to accept various applications of pharmaceutical enterprises carrying out online drug transactions.

110.4.3 Appropriate Measures to Protect Consumers' Privacy

Online pharmacies conduct transactions primarily in the form of an electronic platform, and electronic data saved can be modified or removed without leaving any traces, so once the quality of drugs exist problems, it is hard for consumers or regulatory agency to extracts evidence. To avoid such things the following means may be adopted. Only consumers' privacy is guaranteed can online pharmacies save information in the form of voice recording. Online pharmacies should provide effective sales notes for consumers, for example at the end of the sales provide for consumers with sales receipt which shows names of names and approval number, saledate, amount, etc. Once problems arise, consumers can report to regulatory agencies by effective notes, and regulatory agencies also can inspect the relevant enterprises according to the specific contents of sales notes.

110.4.4 Regulation on Medicines from International Online Pharmacies

Strengthen regulation on illegal medicines and strictly controlled medicines from international online pharmacies, make laws and regulations about drug import and

customs processing, simultaneously strengthen international cooperation with related institutions of other countries and regions, define regulatory forms, research and discuss countermeasures for regulation on transnational drug sales of online pharmacy [7, 8].

110.4.5 Publicity Work

Select typical illegal Internet drug sales case to give corresponding sanction, invite media to participate in the whole process, make an example and highlight the deterrent effect. In Internet drug cases, encourage informed staff to provide real-time information and reliable first-hand material for cases inspection. Enhance public self protection sense. Increase the public's regulatory system, legitimate online drug distributors of publicity to raise public recognition illegal online drug transaction capabilities, enhanced public awareness of self protection.

110.5 Conclusions

As the Internet is becoming mature day by day and the competition in the pharmaceutical market is more intense, more and more pharmaceutical retail chain enterprises begin to provide Internet trading services. It brings certain convenience for people who buy medicine, but it also will change business pattern of traditional medicine economic, forcing pharmaceutical enterprises of each link to reconstruct enterprise value chain, reposition and think their organizational structure, business process and business channels, which will bring some problems about security, law, economy, etc. Currently, the Internet drug market regulation is a new thing in drug regulation. In order to ensure the normalized operation of Internet drug market, corporate, government, society and the public need work together to continually enhance the theoretical research and exploration of this field, with a view to exploring a win-win way which can not only promote the fast development of the Internet pharmaceutical market, but also ensure drug safety.

References

1. State Food and Drug Administrator. <http://www.sda.gov.cn/WS01/CL0006/>
2. Xintao Z, Zhilu H, Lei C, Qineng P (2008) Current status and analysis of drug information administration over internet [J]. Chin JMAP 25(8):776-778
3. Li S (2009) Study on countermeasures for strengthening regulation on online pharmacies in our country [D]. Shenyang Pharmaceutical University, Shenyang
4. Wu B (2008) Government regulation and standardized operation of medical E-commerce [D]. Henan University, Henan

5. Yongfa C (2005) Supervision of internet drugstore in USA and its enlightenment [J]. *China Pharm* 16(20):1597–1599
6. Dajun M (2010) Comparative study of regulatory mechanisms for internet drug management [J]. *Chin Pharm Aff* 24(3):250–254
7. Zhang J (2007) Government should strengthen supervision of the sale of drugs online [J]. *China Pharm* 16(2):2–3
8. Zhang Z, Wang G, Shen A (2004) How the U.S. government manage online pharmacies [J]. *Her Med* 5(23):353–354

Chapter 111

Study on Accounting Shenanigan in Universities Using Improved Game Theory

Jinfang Zhao

Abstract The accounting shenanigan is one of the most terrible causes responded to the graft and corruption phenomena. The graft and corruption definitely do a deadly hazard to the social trust and integrity. It is therefore critical for the governments and organizations to punish severely the accounting shenanigan and prevent the impending graft and corruption phenomena to build honest social and working environments. However, most economic corruptions occurred in universities is caused by accounting shenanigan. The necessary of accounting shenanigan module analysis and detection is urgent. To investigate the accounting shenanigan in universities can not only eliminate economic corruption but also ensure the healthy of the finance management. For this reason, a new method to model and analyze the accounting shenanigan in universities based on the improved game theory is proposed in this paper. The evolutionary game between universities and Supervisor was investigated. The analysis results show that the financial management system using in the universities is not effective to avoid accounting shenanigans. However, by reasonable strategies, the supervise efficiency is improved, and the universities will abide by the accounting standards. As a result, the university financial management system can be consummated to eliminate accounting shenanigans.

Keywords Accounting shenanigan · University · Game theory

J. Zhao (✉)

Tourism College of Zhejiang, Hangzhou 311231, China

e-mail: jfang@tczj.net

111.1 Introduction

Education is an important part of Chinese well-to-do society. This is special true for the national colleges and universities. In past two decades, the State Council has been dedicated to implement the strategic objectives of developing the country through science and education. Great achievements have been made in the construction of colleges and universities. There are nearly 2000 colleges and universities in China now. It can be seen that Chinese government almost have attached great importance to the development of high education plan. During 2003–2010, State Statistics Bureau clearly pointed out that over 100 billion RMB have used as education investment. This rapid development have changed universities a lot, including school scale, social influence, and school subject, etc. [1]. While in the developing process, the universities have confirmed their legal entity status in the society. Since economic man assumes that people are self-interested, some phenomena of the construction raise in universities, such as accounting shenanigans. But the current financial accounting system in universities lags to the economic progress. In 1997, China has promulgated national financial standards, including financial criteria for institutions and higher school, etc. [1]. Although these standards issued accelerated the reform of financial accounting system in universities, the existing financial accounting system still cannot cope with the sophisticated market economy system. The cause is determined by the special Chinese political and economic environment. This discordance may hasten the college financial corruption cases [2].

Accounting shenanigans have become the hot topic in the society. Many academic and political organizations have paid great attention to the research and management of accounting shenanigan. The representative achievements are the analysis of the causes and countermeasures for accounting information distortion [3]. The shenanigan phenomenon are analyzed and discussed from different perspectives by case studies. Several techniques for accounting shenanigan analysis and detection have been proposed [4]. These include iceberg theory [5], theory of shenanigan factors [6], game theory [7], and risk theory [8] and so on. Literature review shows that profit organizations and enterprises are researched very much, but do few institutions, especially for university accounting shenanigan. Hence, it is critical to investigate deeply on financial accounting system in universities. Feasible and available reform ideas and patterns should be proposed for the health development of universities.

In this work, university financial management system has been investigated from the prospect of economics, and the analysis and solution of the accounting shenanigan occurred in universities has been implemented based on the improved game theory. The analysis results show that the existing university financial management system is not efficient and hence causes false accounting information. It is imperative to call for effective accounting supervision function to take effective rewards and punishments measures to depress accounting shenanigans.

Thus, several strategies have been proposed to hand the university financial accounting management, and thus to improve finance management system.

This work is organized as follows. The accounting shenanigan in universities is described in [Sect. 111.2](#). In [Sect. 111.3](#), the proposed method using improved game theory is discussed to analyze accounting shenanigans, and the stability analysis of evolutionary stable strategy between supervisor and universities is investigated in [Sect. 111.4](#). Conclusions are drawn in [Sect. 111.5](#).

111.2 Accounting Shenanigans in Universities

The reason for the economic man appeared in accounting shenanigan is usually to pursue for economic or political benefits. As an economic man, the economic interests are the inevitable problems even for universities. Therefore, this paper argues that university accounting shenanigan is similar to the accounting shenanigans happened in profit organizations but different in specific aspects.

111.2.1 Jobbery

Jobbery means corruption among public officials. Presidents of a few universities are on fund management, the tendency of jobbery is common occurrence. Presidents use the special accounting work duties to instigate for profitability. Some phenomenon existed is serious such as tampering with accounting data, false invoices, shenanigan accounts, false number, false table, peripherals accounts, transfer state-owned assets and so on. Through these means, including not only political power, but also economical motivation, the persons occupied a key position could benefit from its political and social status. In this situation, the accountants somehow are forced to do the crimes.

111.2.2 Organized Crime

When facing huge economic or political benefits, organized crime always happens. The chairmen in different departments collude with each other and change accounting information to obtain benefit. This crime way is very serious, and very complex to discover. However, the organized crime usually involves a large amount of critical persons, and once the crime is accused, a lot of people related will be punished. Hence, the impact to the society is severe.

111.2.3 Small Private Treasury

After providing the public services, the economic benefits do not enter the school financial management system. For this reason, a portion of income has flowed into the packets of some individuals and some small teams. As a result, the illegal financial benefits have avoided the supervision, i.e. gray income. These are several means to grab gray income.

- (1) Making false accounts. Some universities do not implement accounting system and confusion category artificially when involving some spending. These actions may result in financial statement distortion.
- (2) Imaginary budget projects. Expense budget are not real used in career development and business work through exchanging project events or false project budget report which lead to fund waste or erosion.

From the perspective of economics and accounting shenanigan environment, these unhealthy phenomena reflect the following two points:

- 1) The economic man in economic activities selects the chances to obtain greater profit and pursue max profits. Information asymmetry is the direct cause of university accounting shenanigan [2]. As the accounting information holders, college and university administrators have natural information superiority and have the relationship with government as client and agency. Because of the asymmetry of information, it is very easy to make violating operations, which makes the agency have the chance to have illegal profits, i.e. accounting shenanigan. Currently from the domestic accounting shenanigan cases in universities, the duty crime proportion is greater. College administrators in economic activities adopt deceit, hide, and fake to seize the interests.
- 2) Universities accounting shenanigan highlights the defects of current management system. Two conditions are necessary to figure out. The first one is to avoid appointing people by favoritism. The second one is to punish the shenanigan severely according to the laws [9]. Obviously, Universities are difficult to supervise themselves, and the best supervisor should be the government.

However, the internal feature of government department activity [10] may make it ineffective in the supervision management. The definition of internal feature is that the costs and benefits are experienced by traders but not explained in trade terms. The existence internal feature of the government departments enables them to have their own objectives.

The government activity aims not only to public interests. Therefore supervisors may not fulfill their duty well. They always are requested to pay certain cost for social benefits. On the other hand, supervisors have to pursue personal interests. This creates a possibility that supervisors may make decision based on max personal utility in a similar way to economic man activities. Thus this will aggravate accounting shenanigans.

111.3 Accounting Shenanigans Analysis via Improved Game Theory

In traditional game theory, the participants are often assumed to be completely rational, and act in complete information condition. However, it is not exactly true in reality, and the complete information condition is impossible to get. Taking the varieties of the participatos into account, the complexity of the incomplete information caused by economic environment and the limited rationality of the participants are obvious. To solve these problems, the improved game theory, i.e. evolutionary game theory [11], is employed to analyze the relationship of the government and the universities in the accounting shenanigans.

111.3.1 The Evolutionary Game Characteristics of Government and Universities

There are two strategies for government in supervising the accounting shenanigans in universities, that is, supervision and non-supervision. Given: A as the basic earnings of government before supervising the accounting shenanigans, B as the supervision investment, u as the probability that universities decrease the rate of accounting shenanigan under supervising, b as increase rate of cost for accounting shenanigans under supervising, V as the benefits of government from the decrease of accounting shenanigans, a as the probability that no accounting shenanigan occurs under supervising. So when the government choose supervision, the expected return is $A + (1-a)uV - B - bG$; Otherwise, the expected return is A .

In the same way, the universities face the accounting shenanigan risk, that is, shenanigan and non-shenanigan. Given: G as the basic cost of shenanigan without supervision, z as the shenanigan rate for the government supervision, C as the cost of shenanigan under supervision. Thus, the expected return of accounting shenanigan is $u[(1-a)(1+b)G] + (1-u)G - (1-z)C$; Otherwise, the return is $G + zC$.

111.3.2 Dynamic Replication Equation between Government and Universities

Suppose x is the probability of government to supervise the accounting shenanigans in universities, and y is the probability of universities to make accounting shenanigans. The probability could be interpreted as the proportion of the participants who select the strategy in the game. The payoff matrix of government and universities is shown in Table 111.1, where $H = (1-a)uV - B - bG$, $I = u[(1-a)(1+b)G] + (1-u)G - (1-z)C$, $J = B$, $K = zC$, and $I > L$, $|J| > L$.

Table 111.1 The profit matrix between government and universities

		Universities	
		Shenanigan	Non-shenanigan
Supervisor	Supervision	(H, I)	(J, L)
	Non-supervision	$(0, 0)$	$(0, L)$

Assuming $G(1, x)$ and $G(0, x)$ respectively stand for expected profit in choosing supervision and not supervision, at the same time use $E(y, 0)$ and $E(y, 1)$ respectively stand for university expected profit in choosing shenanigan strategy and not shenanigans. So both for the profit function can be expressed:

$$G(1, x) = y[(1 - a)uV - B - bG] + (1 - y)B = yH - (1 - y)J, \tag{111.1}$$

$$G(0, x) = 0, \tag{111.2}$$

$$E(y, 1) = x\{u[(1 - a)(1 + b)G] + (1 - u)G - (1 - z)C\} + (1 - x)^*0 = xI, \tag{111.3}$$

$$E(y, 0) = x(zC) + (1 - x)(zC) = L. \tag{111.4}$$

111.4 Stability Analysis of Evolutionary Stable Strategy between Government and Universities

The evolution of government and universities can be described with the dynamic system as Eqs. (111.5) and (111.6).

$$\frac{dx}{dt} = x(1 - x)[yH - (1 - y)J], \tag{111.5}$$

$$\frac{dy}{dt} = y(1 - y)[xI - L]. \tag{111.6}$$

By analyzing the dynamic system, we can learn that there exist five equilibrium points. Then we can calculate the Jacobian matrix of the system to investigate the stability of equilibrium point. The Jacobian matrix is

$$J_b = \begin{bmatrix} x(1 - 2x)[yH - (1 - y)J] \\ x(1 - x)[H - J] \\ y(1 - y)I \\ (1 - 2y)[xI - L] \end{bmatrix}. \tag{111.7}$$

Table 111.2 lists the local stability analysis of the five equilibrium points.

Table 111.2 The local stability analysis

Equilibrium points	Determinant of J_b	Trace of J_b	Stability analysis
(0, 0)	$J^*(-L)$ (+)	$J - L$ (-)	ESS
(0, 1)	H^*L (+)	$H + L$ (+)	Instability
(1, 0)	$(-J)^*(I - L)$ (+)	$-J + I - L$ (+)	Instability
(1, 1)	$(-H)^*(-I + L)$ (+)	$-H - I + L$ (-)	ESS
$(\frac{I}{J}, \frac{J}{J-H})$	0	$\frac{I^3[(1-2L)(J-H)^2] + HJ + L(H-J)^3}{I^2(H-J)^2}$	Saddle point

We can learn some problems from the analysis of the equilibrium point. Firstly the supervision department of punishment is greater, and the probability of universities shenanigan is lower. Secondly supervision, the supervisory cost of supervision department is greater, and the probability of universities shenanigan is greater. Thirdly universities shenanigan consume more social resource which will lead to supervisor to strengthen supervision and punishment. Finally universities may gain huge income through accounting shenanigan, and then the probability of supervision choosing to supervise is more.

Literature [10] pointed out, by introducing cold strategy, the ultimate game equilibrium point of universities and supervision department is not supervision and not shenanigan. The reason is that choosing implement shenanigan will only make the universities loss less than the gains.

111.5 Conclusions

After game model analysis we can draw the conclusion that whether universities accounting shenanigan or strict supervision, universities and supervisor both make the choice after tradeoff the benefits and costs. In view of our country college development rapid with accounting system innovation slow, establish incentive-restricted mechanism for supervision and the punishment of universities shenanigan and take effective measures of rewards and punishments, these will play the role of improve the regulation efficiency. Accordingly, some proposal can be referenced. Firstly we should establish and perfect the incentive and restraint mechanisms. Supervision department should supervise effectively and make clear of the supervision responsibility, and to strengthen supervision responsibility punishment of accident and increase punishment strength of dereliction of supervisor. Meanwhile, to increase effective supervision actions and excellent supervisor award, establish long-term material and spiritual incentive mechanism. From the long term, this is the effective method to promote supervision positively. Secondly, increasing shenanigan punishment strength is the effective way of inhibiting accounting shenanigan in short-term. Intensify sanctions would raise the cost of accounting shenanigan which curb to commit frauds to a certain extent. This will produce a certain degree of warning effect for cheating and also can

reduce the probability of accounting shenanigan. For supervisor this will not affect their costs, thus increasing penalties for fraudulent is practical and feasible measures. Thirdly, we can improve the supervision quality. On one hand this requires improve the quality of supervisor. On the other hand is to mobilize social forces for supervision. Finally, we can strengthen and improve the supervision construction of state-owned capital accounting system. Owning a perfect accounting supervision system can largely inhibit shenanigan and improve the effectiveness of the accounting supervision.

References

1. Yang Y (2008) Research on the reform of financial accounting system in high education. Kunming University of Technology Press, Kunming
2. Fan Z (2010) College's accounting fraud causes and countermeasures. *Netw Wealth* 15:39–40
3. Jiang Y (2002) Accounting information distortion status, causes and countermeasures study: the listed company profit operation empirical research. Chinese Financial and Economic Press, Beijing
4. Zhang Z (2010) Accounting fraud and stock market regulation of the game theory. *Account Commun* 9:125–127
5. Hines T (2002) Developing an iceberg theory of cost comparisons in relation to sourcing decisions by UK fashion retailers. *J Text Inst Part 1 Fibre Sci Text Technol* 93:3–14
6. Liu M, Zhang F (2007) What have caused consumer fraud in auction website? In: Proceeding of 2007 international conference on wireless communications, networking and mobile computing, pp 3392–3395
7. Geng X, Whinston AB, Zhang H (2005) Health of electronic communities: an evolutionary game approach. *J Manag Inf Syst* 21:83–110
8. Budovskii VP (2009) Evaluation of dispatcher decisions by risk theory methods. *Power Technol Eng* 43:333–335
9. Le Y, Zhang Q (2011) Financial report frauds game analysis and innovative financial report means using XBRL. *Account Inf* 11:62–63
10. Yu H (1999) Regulation and market. People Press, Shanghai
11. Vincent TL, Brown JL (2005) Evolutionary game theory. Natural selection and Darwinian dynamics. Cambridge University Press, UK

Chapter 112

Improved Iterative Decoding Algorithm for Turbo Codes

Zhenchuan Zhang and Jiang Nana

Abstract This chapter primarily studies the MAP and the SOVA, and discusses the suitability of combining both algorithms. An iterative decoding algorithm is improved. The frames are first applied with SOVA iterations with error detection for iteration termination. After a fixed maximum number of SOVA iterations, the erroneous frames are continued with Log-MAP iterations. The extrinsic soft output from the last SOVA iteration are selectively fed to the first Log-MAP iteration as initial priori input. The result of the simulation shows that the improved algorithm could reduce the complexity and keep high performance.

Keywords Turbo decoding · SOVA · Log-MAP · Complexity reduction · Decoding performance

112.1 Introduction

Since their invention in 1993, Turbo codes have attracted great interests because of their novel coding/decoding techniques, MAP and iterative decoding. At a SNR of 0.7 dB, its BER is about 10^{-5} . It is a large work of decoding, because the numerical computation decoding algorithm is very complicated. The MAP is used firstly, other algorithms are improved in order to reduce the complexity [1]. It consists of two types: SOVA and MAP. SOVA has a moderate complexity but

Z. Zhang (✉) · J. Nana
Department of Information Science and Engineering, Northeastern University,
Shenyang 110004, China
e-mail: zhangzhenchuan@ise.neu.edu.cn

J. Nana
e-mail: jnn362@163.com

incurs some performance loss. MAP has lower BER, it could decode while the SNR lower than 1 dB, but it is complicated. In the basic of studying the arithmetic in decoding, this thesis improves a new algorithm.

112.2 Log-MAP and SOVA

MAP is one of the SISO, it makes the rate of priori maximum. The paper denotes $\Lambda(x_k)$ as LLR. Comparing Λ with threshold value, if Λ is larger than 0, y_k is 1, otherwise y_k is 0.

Denoting:

$$\Lambda(x_k) = L(x_k) + L_c(x_k) + L_e(x_k) \tag{112.1}$$

Where $L_e(x_k)$ is extrinsic value, it is translated to the next decoder as a priori information, such as $L_2(x_k) = L_{1e}(x_k)$.

112.2.1 Log-MAP Algorithm

Denoting $A_k(s)$, $B_k(s)$, $M_k(e)$ with the $\alpha_k(s)$, $\beta_k(s)$, $\gamma_k(e)$ of MAP. $\alpha_k(s)$ and $\beta_k(s)$ have exponential calculation in the MAP, it denotes [2]:

$$\max_e^*(f(e)) = \ln \left(\sum_e e^{f(e)} \right) \tag{112.2}$$

$$A_k(s) = \ln \sum_{e:s_k^E(e)=s} \alpha_{k-1}(s_k^E(e))\gamma_k(e) = \max_{e:s_k^E(e)=s}^* [A_{k-1}(s_k^E(e)) + M_k(e)],$$

$$k = 1, \dots, N - 1 \tag{112.3}$$

$$B_k(s) = \ln \sum_{e:s_k^E(e)=s} \beta_{k+1}(s_{k+1}^E(e))\gamma_{k+1}(e) = \max_{e:s_k^E(e)=s}^* [B_{k+1}(s_{k+1}^E(e)) + M_{k+1}(e)],$$

$$k = N - 1, \dots, 1 \tag{112.4}$$

In order to correct the function, it is $\log(e^{x_1} + e^{x_2}) = \max(x_1, x_2) + \log(1 + e^{-|x_2-x_1|})$,

so the correct function is $f_c(x) = \log(1 + e^{-|x|})$ [3]. The paper gets the output value of the LLR of the information:

$$\Lambda_k(u; O) = \max_{e:M(e)=1}^* [A_{k-1}(s_k^E(e)) + M_k(e) + B_k(s_k^E(e))] - \max_{e:M(e)=0}^* [A_{k-1}(s_k^E(e)) + M_k(e) + B_k(s_k^E(e))] \tag{112.5}$$

112.2.2 SOVA

SOVA can provide soft decision value and extrinsic information. The algorithm not only computes log-likelihood path, but also considers the priori values and transmits them between two decoders when it computes path metric [4].

112.2.2.1 Computing the Accumulative Path Metric

At time k , the accumulative path metric of the path m is:

$$M_k^{(m)} = M_{k-1}^{(m)} + \sum_{n=1}^N x_{k,n}^{(m)} L_c y_{k,n} + x_{k,1}^{(m)} L(u_k), \quad m = 1, 2 \quad (112.6)$$

Where x is output bit of the coding list, y is the acceptant list. $L_c = 4 \cdot E_s/N_0$, u_k is the input bit of time k .

112.2.2.2 Computing a Soft Decision Value

At time k , the probability of path m is $P_k^m = C \bullet e^{M_k^{(m)}/2}$, $m = 1, 2$, C is constant.

The probability of selecting wrong path is

$$\Phi_k^e = \frac{P_k^c}{P_k^s + P_k^c} = \frac{1}{1 + e^{(M_k^{(s)} - M_k^{(c)})/2}},$$

where $M_k^{(s)}$ and P_k^s are the accumulative path metric and the rate of survive path; $M_k^{(c)}$ and P_k^c are the accumulative path metric and the rate of collateral survive path. At time k , the LLR of the decision path (soft decision value) is

$$\Delta_k = \log \frac{1 - \Phi_k^e}{\Phi_k^e} = \frac{M_k^{(s)} - M_k^{(c)}}{2}.$$

112.2.2.3 Computing Extrinsic Information Before Updating Soft Decision Value

The paper can get the value of extrinsic information by reducing intrinsic value from the LLR of hard decision list.

$$L_E\left(\hat{u}_k\right) = L\left(\hat{u}_k\right) - L_I\left(\hat{u}_k\right) = L\left(\hat{u}_k\right) - y_{k,1} - L(u_k) \quad (112.7)$$

where $y_{k,1}$ is the systemic code of the acceptant code, $L(u_k)$ is the priori value of the decoder.

Table 112.1 Complexities of Log-MAP and SOVA

Operation	Log-MAP	SOVA
Max ops	$5 \times 2^M - 2$	$3(M + 1) + 2^M$
Additions	$15 \times 2^M + 9$	$2 \times 2^M + 8$
Mult. by + 1	8	8

112.3 Improved Iterative Decoding Algorithm

The soft input and output values of the SOVA and Log-MAP algorithms are of the same logarithm likelihood ratios. It is therefore of practical interests to consider if the two algorithms can be utilized in a cooperative way so that the advantages of the both algorithms can be obtained together. The feasibility of combining the two algorithms comes from two facts. First, although at each signal to noise ratio (SNR) region SOVA is significantly inferior to Log-MAP. If SOVA is only able to correctly decode half of the frames, further apply Log-MAP for the other erroneous frames. At a moderately SNR, the improve algorithm can keep the average complexity close to that by applying SOVA alone. Second, the extrinsic soft output information can be properly utilized by properly pre-set initial a priori values can improve decoding performance and reduce the number of necessary iterations.

Frames are first applied with SOVA iterations with error detection for iteration termination. After a fixed maximum number of SOVA iterations, the erroneous frames are continued with Log-MAP iterations. In order to reduce the number of necessary Log-MAP iterations, the extrinsic soft output from the last SOVA iteration are selectively fed to the first Log-MAP iteration as initial priori input [5].

112.3.1 Complexity Reduction by Combining SOVA and Log-MAP Iterations

To quantify computational complexities, the paper denotes the complexity required for a SOVA iteration as C_S and that for Log-MAP as C_M . It is showed in Table 122.1.

Denoting:

$$C_m = 15 \times 2^M + 9 \quad (112.8)$$

$$C_s = 2 \times 2^M + 8 \quad (112.9)$$

The relative complexity of the two algorithms depends on encoder memory length. For memory length of 2–4, $C_M \geq 3C_S$. If N_1 SOVA iterations result in a FER and the erroneous frames are further applied with N_2 Log-MAP iterations to achieve a FER, the average complexity is:

$$C_{Avg}(F_e^{Log-MAP}, F_e^{SOVA}) = N_1 C_S + N_2 C_M F_e^{SOVA} \quad (112.10)$$

112.3.2 *Initializing a Priori Input for Log-MAP Iterations by Selecting Extrinsic Output from SOVA Iterations*

Correctly biased a priori input can reduce necessary decoding iterations and even lower the error rate, while erroneous a priori input slows down convergence and degrades performance. It is also shown that erroneous a priori inputs have stronger negative effects than correct inputs have positive effects. Therefore, avoiding erroneous input may be more important than selecting correct input. Also the convergence rates by SOVA iterations are much slower than that by Log-MAP iteration [6]. Thus feeding all extrinsic output from SOVA iterations to Log-MAP could degrade the performance if many erroneous values are used. This is especially the case at a SNR which is relative low for SOVA but moderately high for Log-MAP, since SOVA outputs are still very noisy but Log-MAP expects fast converging extrinsic values after a few iterations.

In order to better initialize Log-MAP iterations, we select extrinsic outputs from SOVA that are more likely to give correct hard decisions. The paper considers two approaches for extrinsic output selection. The first is by examining the absolute value of extrinsic outputs. Since larger magnitudes in extrinsic output indicate stronger confidence in correct decisions, it can apply a threshold value so that data bits with higher output magnitudes are selected and their priori inputs for Log-MAP are pre-set with their extrinsic values, while the rest are pre-set with zero a priori inputs. As the extrinsic values from most erroneous bits fluctuate from -10 to 10 in early iterations, the paper can set the threshold value to be 10 .

Another approach is using the cross-entropy values to preclude most erroneous extrinsic outputs. Denoting $T_n(i)$ for the cross-entropy for bit n after the i th iteration [7], the paper has:

$$T_n(i) \approx \frac{|L_n^a(i+1) - L_n^a(i)|^2}{\exp(|L_n^Q(i)|)} \quad (112.11)$$

where $L_n^a(i)$ is the priori input to bit n for the i th iteration, and $L_n^Q(i)$ is the posteriori output for bit n from the first component decoder after the i th iteration. The paper measures the ratio of $T_n(i)/T_n(1)$ after each iteration.

The concrete steps of improved iteration decoding algorithm (IIDA) are as flows:

Initialize variables, iter = 1;

Compute path measure of each state, look for surviving path measurements and competitive path measurements, compute metric sent;

Update state metric;

Compute LLR, put extrinsic information as a priori information;

If iter < N1, iter = iter + 1, then return to 2; else storage error frame, selectively put external information as the priori information input;

Remove the wrong frames as new input information, compute $\alpha_k(s)$, $\beta_k(s)$;

Compute LLR, make stop decision.

The flow of improved decoding algorithm is showed in Fig. 112.1.

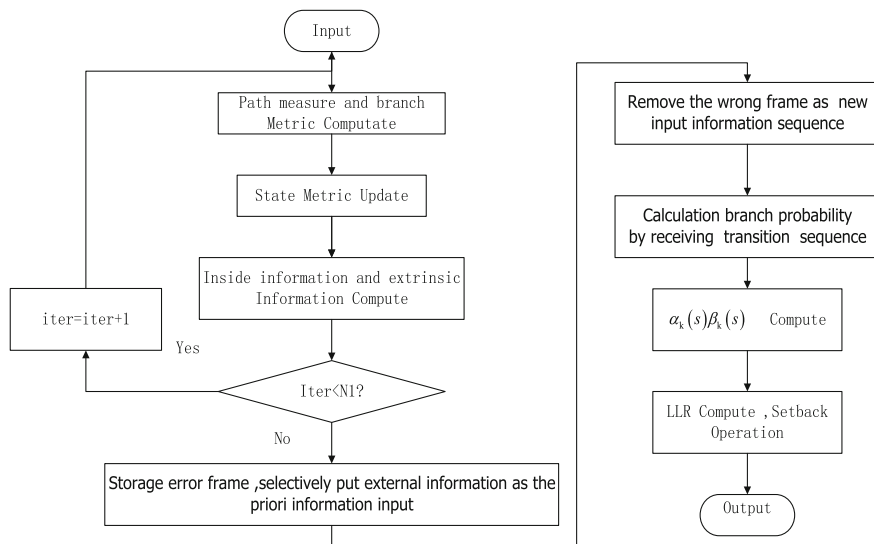


Fig. 112.1 The flow of improved decoding algorithm

112.4 Simulations

In simulations the length of the frame is 400, the Turbo encoder consists of two identical rate 1/2 recursive systematic convolution (RSC) component encoders with generator polynomials (7, 5). After puncturing, the overall code rate 1/2 is maintained. The paper assumes binary phase shift keying (BPSK) modulation on additive white Guassian noise (AWGN) channels. The frames are first applied with 5 SOVA iterations with error detection for iteration termination. After a fixed maximum number of SOVA iterations, the erroneous frames are continued with 5 Log-MAP iterations. The paper compares with SOVA and Log-MAP.

The paper compares the BER and FER, it is shown in the Fig. 112.2 and 112.3. The improved iterative decoding algorithm has similar performance as Log-MAP. So it combines the benefit of SOVA and Log-MAP, then it improves the performance.

The storage length of encoder is 2, the complexities of Log-MAP and SOVA are 690 and 160. The average complexity of improved algorithm could compute by (112.10). The data are shown in Table 112.2.

The complexity of improved algorithm is shown in Fig. 112.4.

Figure 112.4 shows that the average complexity of the improved algorithm degrades with the increasing of SNR and it closes to a stability value. The complexities of Log-MAP and SOVA are 690 and 160. So the improved algorithm has better complexity which is similar to SOVA. The improved algorithm keeps the performance as Log-MAP, but the complexity reduces a lot.

Fig. 112.2 BER of IIDA, SOVA and Log-MAP

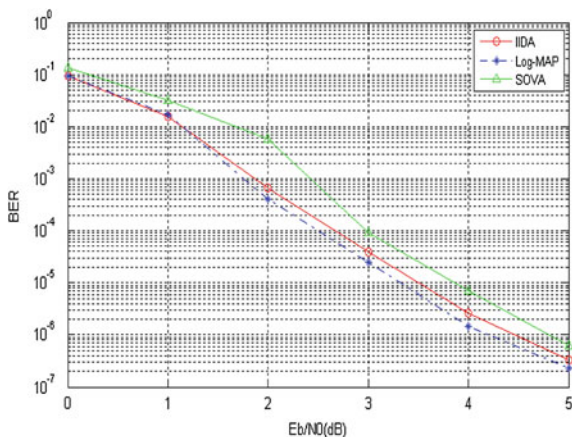


Fig. 112.3 FER of IIDA, SOVA and Log-MAP

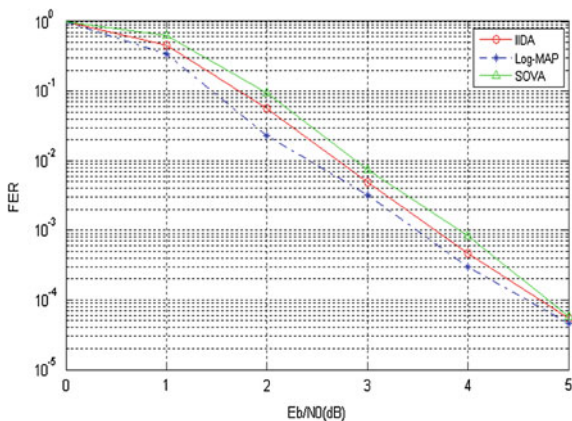


Table 112.2 Complexity comparison

SNR(dB)	0	1	2	3	4	5
Complexity of Log-MAP(times)	10C _M	10C _M	10C _M	10C _M	10C _M	10C _M
Complexity of SOVA(times)	10C _S	10C _S	10C _S	10C _S	10C _S	10C _S
Complexity of IIDA(times)	425	202	92.5	81.3	80.1	80.0

Figure 112.5a shows that the Log-MAP priori input are selected by SOVA extrinsic value magnitudes. The result shows that the scheme slightly degrades the FER. In order to compensate for that performance gap, the paper may apply more iteration. Log-MAP priori input are selected by SOVA cross-entropy ratios in the

Fig. 112.4 Complexity of IIDA

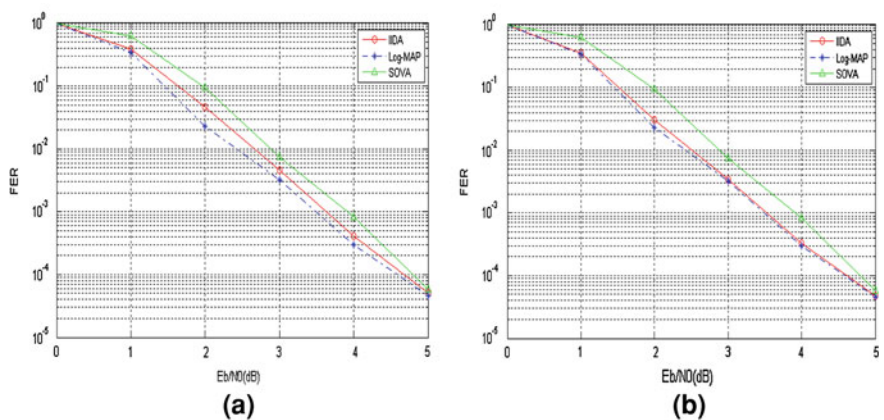
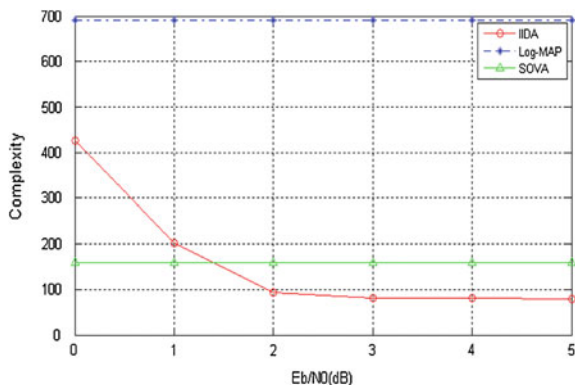


Fig. 112.5 Log-MAP a priori input selection by SOVA

Fig. 112.5b, the scheme is better than the first one about FER, at the same time, it is feasibility.

112.5 Conclusions

The Turbo decoding principle has found widespread applications, not only in theory, but also in practice, distinguishingly in the transmission of the channel coding, so it is used in widespread field such as mobile communication. The usual decoding algorithms are not fit for the translation of the real-time data because of time delay. The paper provides the advantage and disadvantage of some decoding algorithms, and then it improves a new one. In simulation the complexity and convergence are much better. In order to reduce the soft output of SOVA, the

paper improves Log-MAP, it introduces better performance but does not increase the complexity. In recent years, some research works have been carried out to improve the performance of these practical Turbo decoding algorithms.

References

1. Zhang Z, Wang J (2008) Performance and simulations of improved adaptive iterative decoding algorithm of Turbo Codes. *J Northeastern Univ (Nat Sci)* 29(2):205–208
2. Li D, Dong L (2007) Improved Log-MAP algorithm for turbo-codes decoding. *J Civil Aviation Univ Chin* 25(1):49–51
3. Zhang L, Liu X (2006) Approximated Log-MAP decoding algorithms for turbo codes. *J Circuit Syst* 11(3):70–74
4. Jiang X, Jie C, Qiu Y (2005) A simplified SOVA algorithm. *J Chin J Electron Devices* 27(3):467–469
5. Wu J, Vojcic BR (2009) Combining iterative SOVA and Log-MAP algorithms for turbo decoding. In: 43rd annual conference on digital object identifier of IEEE information science and systems. pp 420–423
6. Robertson P, Villebrun E, Hoeher P (1995) A comparison of optimal and sub-optimal map decoding algorithms operating in the log domain. In: Proceedings of IEEE international conference on communications, Seattle, pp 1009–1013
7. Wang Z, Parhi K (2003) High performance, high throughput turbo/sova decoder design. *J IEEE Trans Commun* 51:570–579

Chapter 113

Financial Evaluation of the Listed Companies Based on Statistical Analysis Methods

Guofu zhang

Abstract Use multivariate statistical analysis methods, This chapter took a research on financial evaluation of Beijing's listed companies. Firstly, we used factor analysis to choose eight indexes to analyze and sort them. During cluster analysis, two methods were used, and their effects were compared, then we took a better one to evaluate. Then, with discriminated analysis, those were validated above conclusions. Finally, the existed problems were put forward. The result could offer company some references, and provide the gist for the investor to judge listed company's quality.

Keywords Multivariate statistical analysis • Financial evaluation • Listed company

113.1 Introduction

Financial evaluation experimental work has been started in state-owned company in China in 1998 [1]. For example, “financial evaluation rules”, “financial evaluation operation detailed rules” “financial evaluation index explanation”, “financial evaluation billiard-marker method” had come on by state Ministry of Finance, Ministry of Personnel, Committee for Economical Trade and Committee for Family Planning together [2].

“Fortune” and “Business weekly” make use of gross indexes to evaluate the scale of the company, but wrong individual order was obvious in the result because

G. zhang (✉)

Department of Economics and Management College of Economics and Management
Heilongjiang Bayi, Agricultural University, Daqing, China
e-mail: byndzgf@qq.com

different indexes were chose. So the result always cannot reflect the real situation of the company comprehensively. At the same time, Research center of company in Qing Hua University and Negotiable Security Magazine together came on a new method which needs four seasonal data [3], to choose more than 30 indexes to compute and compare, and to compute the score of each index. This method needs high quality fundamental data, overabundant index, and singular information. But this method has been adopted by many evaluation units, with different indexes and score computation. In this paper, indexes mainly refer to the state-owned financial evaluation system.

113.2 Methods and Sample Analysis

Keeping to the rules of all sidedness, fairness, scientificalness, and maneuverability and so on, the index system reflects the prosecution of the company as comprehensive and genuine as possible. Data are mainly from www.cnlist.com and www.cninfo.com.cn. They threw daylight on annals and correlative finance information. During analysis, eight financial indexes of 73 listed companies' work performance in 2002 are selected. The eight indexes are EPS for earnings per shares ($\times 1$), NAY for net asset yield ($\times 2$), RTA for rate of return on total assets ($\times 3$), TTC for turnover of total capital ($\times 4$), TCA for turnover of current assets ($\times 5$), ADR for assets debt ratio ($\times 6$), CAR for capital accumulation rate ($\times 7$) and POG for primary operation growth rate ($\times 8$). Firstly, we compute eight financial indexes in these companies, and process each factor's score which is analyzed with factor analysis. We also declare advantages and disadvantages of the performance in listed company commonly. Secondly, we classify the listed company with cluster analysis. Thirdly, discriminated analysis is used to establish a set of rules to distinguish the sample belongings. The result could provide company references, and provide the gist for the investor to judge listed company's quality.

Factor analysis belongs to category of reducing dimensionality [4]. Its core point takes the factor model as a start. The method is to extract common factor when data loss is in granted bound to reflect correlation between original variables and factors, viz. what kind of factor has effect and how. In generally, the number of selected common factors is less than that of the original variables, that is to say, using less uncertain but explicable variables to describe divivable original variable.

Factor analysis model for Beijing listed company's financial performance all-around estimation is as follows:

$$X = AF + \varepsilon$$

Where $X = (X_1, \dots, X_8)'$ is the vector matrix, $F = (F_1, \dots, F_m)'$ is common factor vector matrix, $m \leq 8$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_8)'$ is special factor vector matrix, $A = (a_{ij}) 8 \times m$ is component matrix.

Table113.1 Sample total variance explained

Component	Eigen value	Percentage of Variance	Cumulative %
1	2.954	36.923	36.923
2	1.683	21.043	57.966
3	1.161	14.51	72.476
4	0.908	11.345	83.821
5	0.667	8.343	92.163
6	0.256	3.198	95.362
7	0.215	2.685	98.047
8	0.156	1.953	100

Table113.2 Component matrix

	Component				
	1	2	3	4	5
IPS	0.873	-0.234	0.177	-0.071	0.056
NAY	0.83	-0.22	0.183	-0.272	0.144
GAY	0.862	-0.256	0.072	-0.141	-0.022
GAV	0.543	0.721	-0.285	0.133	-0.058
CAV	0.526	0.71	-0.35	-0.056	-0.101
AIR	-0.16	0.587	0.523	-0.044	0.591
CAR	0.367	-0.291	-0.205	0.806	0.288
PWOIR	0.171	0.248	0.756	0.369	-0.445

Take eight indexes data of 73 listed companies in Beijing as a sample, and compute sample covariance's eigenvalue as Table 113.1.

Taking together, we choose frontal five eigenvalues, namely $m = 5$. All frontal five eigenvalues account for more than 85–92.163%. These five factors can describe 92.163% information of sample covariance. At this time, component matrix A as Table 113.2.

In order to explain common factors more accurately and easily, it is necessary to rotate matrix A. The more scattered for the rate of the variance, the better for rotating. Each common factor's rate is polarized to 0 and 1. Here we choose varimax rotation, and rotate component matrix as Table 113.3.

From Table 113.4, final common factor deviation estimate indicates that all variable can be explained effectively by five common factors, because common factor deviation is in the range of 0.833 and 0.997.

After common factors are ascertained, in order to estimate listed company comprehensively, common factor estimation relevant with each sample is needed to be reviewed, namely factor score. Then factor score function is computed with regression method. Component score coefficient matrix is as Table 113.5.

Factor score function is as follows :

Table113.3 Rotated component matrix

	Component				
	1	2	3	4	5
IPS	0.895	0.117	0.163	0.055	-0.107
NAY	0.926	0.084	0.007	0.033	-0.019
GAY	0.878	0.146	0.098	-0.17	0.043
GAV	0.124	0.937	0.103	0.081	0.077
CAV	0.148	0.943	0.063	-0.023	-0.024
AIR	0.114	-0.089	-0.096	0.974	0.125
CAR	0.16	0.027	0.979	-0.095	-0.001
PWOIR	0.079	0.038	0.002	0.12	0.987

Table113.4 Communalities

	Initial	Extraction
IPS		
NAY	1	0.856
GAY	1	0.866
GAV	1	0.833
CAV	1	0.917
AIR	1	0.916
CAR	1	0.994
PWOIR	1	0.994
IPS	1	0.997

$$\begin{cases} F_1 = 0.368X_1 + 0.436X_2 + 0.355X_3 - 0.09X_4 - 0.053X_5 + 0.074X_6 - 0.086X_7 - 0.055X_8 \\ F_2 = -0.057X_1 - 0.09X_2 - 0.02X_3 + 0.538X_4 + 0.551X_5 - 0.061X_6 - 0.015X_7 - 0.061X_8 \\ F_3 = 0.027X_1 - 0.125X_2 - 0.061X_3 + 0.086X_4 - 0.105X_5 + 0.092X_6 + 1.017X_7 - 0.025X_8 \\ F_4 = 0.051X_1 + 0.167X_2 - 0.082X_3 - 0.019X_4 - 0.086X_5 + 1.039X_6 + 0.097X_7 - 0.131X_8 \\ F_5 = 0.033X_1 - 0.124X_2 - 0.001X_3 - 0.034X_4 - 0.055X_5 - 0.124X_6 - 0.026X_7 + 1.019X_8 \end{cases}$$

Then take deviation communalities of each factor as weight timed each score to frame a score function. Mark communalities of I common factors as follows [4]:

$$w_i = \lambda_i / \sum_{i=1}^n \lambda_i \quad i = 1, \dots, 5$$

Where is latent root of sample variance matrix,.

The synthesis score model for financial evaluation in listed company is as follows:

$$F = w1F1 + w2F2 + w3F3 + w4F4 + w5F5$$

Table113.5 Component score coefficient matrix

	Principal component				
	1	2	3	4	5
IPS	0.368	-0.057	0.027	0.051	0.033
NAY	0.436	-0.09	-0.125	0.167	-0.124
GAY	0.355	-0.02	-0.061	-0.082	-0.001
GAV	-0.09	0.538	0.086	-0.019	0.034
CAV	-0.053	0.551	-0.105	-0.086	-0.055
AIR	0.074	-0.061	0.092	1.039	-0.124
CAR	-0.086	-0.015	1.017	0.097	-0.026
PWOIR	-0.051	-0.016	-0.025	-0.131	1.019

113.3 Comprehensive Evaluation

After computing each factor score of each company, we rank listed company according certain factor score. From the rotated component matrix and function of five factors, the quantity relation between each main factor and original variable should be observed. F1 is mainly affected by income per share, net assets yield and general assets yield, namely profit capability factor; F2 is mainly affected by general assets velocity and current assets velocity, called operational capability factor; F3 is mainly affected by capital accumulated ratio, named saving capability factor; F4 is affected by assets indebted ratio, which is solvency factor; F5 is affected by primary working operation increase rate, called pullulating capability factor. Each factor's weight is ascertained according deviation communalities. Then we obtain synthesis factor. Its expression is as follows:

$$F = 36.923 * F1 + 21.043 * F2 + 14.510 * F3 + 11.345 * F4 + 8.343 * F5$$

Table 113.6 is the rank of synthesis score and each common factor. It is a relevant quantity of high or low score, just for ranking. It indicates no more or less profit itself.

113.3.1 Factor Analysis Evaluation

The result with method is close to many ministries and institutions. Top ten stock companies are: WYFZ, LTKJ, JRJ, ZRGF, FTQC, ZHGJ, GYGF, THBY, TDKJ and TRT. All these companies are frontal in different places, only some order has few differences. The last five stock companies are: ZGC, JXLY, ZSYY, YYGG, STLK, and BTKJ.

Factor analysis is convenient to reduce many factors into several classes, compute systematic score, and then rank them in order. It could find the main factor affected financial performance level. Cluster emphasizes particularly on

Table113.6 Rank of synthesis score

ZCGF	-35.16	62	ZGWX	-2.35	39
JRJ	94.35	3	ZQL	1.38	36
SDCM	-14.36	50	ZTCY	-40.8	66
LXQX	-31.85	60	TTSW	2.88	35
YHGX	-38.57	64	FTQC	82.49	5
BTKJ	206.66	73	ZGHJ	18.83	21
JDFA	12.38	24	ZMGF	-14.37	51
BJHG	-29.32	58	DTDx	-15.8	53
YJPJ	-12.9	49	YYGG	-78.4	71
ZSJS	-31.12	59	XFGF	34.87	13
BXJC	-23.1	54	SLGF	14.31	23
ZSYY	-36.78	63	LQGF	-11.49	45
JXLY	-74.14	70	BJCJ	14.91	22
LYSD	-39.3	65	WYFZ	102.12	1
ZXGA	8.26	26	DHKJ	4.72	32
SXNY	-12.09	47	XXCL	4.83	31
HLGF	5.02	30	ZNZY	-55.51	68
ZGFZ	-27.41	56	HLZC	41.93	11
HBGS	-43.5	67	THBY	45.78	8
ZGC	-56.93	69	BJBS	26.66	19
QHZZ	2.96	34	ZHGJ	80.59	6
SGGF	6.29	29	GYGF	58.25	7
ATKJ	-29.13	57	ZRGF	90.31	4
ZKSH	-0.62	38	JTZ	22.61	20
ZGGM	-6.88	42	JNRD	8.18	27
SCGF	-9.63	44	TDKJ	44.66	9
HNGJ	34.08	14	YYRJ	7.16	28
MSYH	27.82	18	QNTQ	-11.65	46
ZGSH	29.15	17	ZWKJ	10.01	25
GHYX	-15.07	52	XDSC	-32.19	61
ZGLT	1.25	37	GDDL	30.52	15
WDYL	-3.38	40	STLK	-95.02	72
ZJMY	29.53	16	HTCF	-12.6	48
LTKJ	102.04	2	WFJ	-8.39	43
SHYY	37.03	12	BRGF	-5.58	41
TRT	43.8	10	BJCX	-23.94	55
QHTF	4.06	33			

classifying, via two different cluster methods: Between-groups linkage and Ward's method. Two opposite discriminated analysis (98.6–95.9% respectively) is given to find five class listed companies. Therefore, two methods combined together, there is not only making for analyzing each list company status from individual angle and collectivity angle, but also being convenient to compare transversely and analyzing in details.

Table113.7 Clusters and discriminated analysis of companies

LC	C1	C2	D1	D2	LC	C1	2	D1	D2
ZCGF	1	1	1	1	WYFZ	1	1	1	1
SDCM	1	1	1	1	DHKJ	1	1	1	1
LXQX	1	1	1	1	ZNZY	1	1	1	1
YHGX	1	1	1	1	HLZC	1	1	1	1
JDFA	1	1	1	1	BJBS	1	1	1	1
BJHE	1	1	1	1	JNRD	1	1	1	1
YJPJ	1	1	1	1	YYRJ	1	1	1	1
ZSJS	1	1	1	1	ZWKJ	1	1	1	1
BXJC	1	1	1	1	XDSC	1	1	1	1
ZSY Y	1	1	1	1	STLK	1	1	1	2
JXLY	1	1	1	1	HTCF	1	1	1	1
ZXGA	1	1	1	1	WFJ	1	1	1	1
SXNY	1	1	1	1	BRGF	1	1	1	1
ZGFZ	1	1	1	1	BJCX	1	1	1	1
HBG S	1	1	1	1	JRJ	1	2	1	2
QH ZG	1	1	1	1	LYSD	1	2	1	2
SGGF	1	1	1	2	ZGC	1	2	1	2
ATKJ	1	1	1	1	MSYH	1	2	1	2
ZKSH	1	1	1	1	ZGLT	1	2	1	2
ZGGM	1	1	1	1	ZGHJ	1	2	1	2
SCGF	1	1	1	1	DTDX	1	2	1	2
HNGJ	1	1	1	1	BJCJ	1	2	1	2
ZGSH	1	1	1	1	XXCL	1	2	1	2
GHYX	1	1	1	1	QNTQ	1	2	1	2
WDYL	1	1	1	1	GDDL	1	2	1	2
ZJMY	1	1	4	4	BTKJ	2	3	2	3
SHYY	1	1	1	1	HLGF	3	3	3	3
TRT	1	1	1	1	XFGF	3	3	3	3
QHTF	1	1	1	1	THBY	3	3	3	3
ZGWX	1	1	1	1	LTKJ	4	4	4	4
ZQL	1	1	1	1	FTQC	4	4	4	4
ZTCY	1	1	1	1	ZHGJ	4	4	4	4
TTSW	1	1	1	1	GYGF	5	5	5	5
ZMGF	1	1	1	1	ZRGF	5	5	5	5
YYGG	1	1	1	1	JZTZ	5	5	5	5
SLGF	1	1	1	1	TDKJ	5	5	5	5
LQGF	1	1	1	1					

113.3.2 Cluster Analysis and Discriminated Results

In cluster analysis, distance is used to scale close or distant. There are many distances. Euclidean distance is employed wildly [5]. Although it has some limitations, compared with the others, it is more suitable to this research. From choosing distance between linkages, after compared, two methods were chosen in

Table113.8 Test for equality of cluster method one

Clu1	0.07914	-0.1643	-0.2222	-0.0223	-0.2151
Clu2	-5.7334	-0.2594	0.19926	-0.167	1.13833
Clu3	-0.0223	-0.3063	-0.235	0.49273	4.03236
Clu4	0.26657	3.44295	-0.2155	0.73106	0.10942
Clu5	0.02329	0.25841	3.73177	-0.53	-0.0572
λ	0.535	0.474	0.179	0.951	0.264
F	14.774	18.843	77.955	0.872	47.396
Sig	0	0	0	0.485	0

this paper: between-groups linkage and ward's method. Simultaneously the cluster results are given. So it is easy to compare for reader.

Based on the cluster analysis, we carry discriminative analysis into execution to the sample, aimed at validating the veracity of clusters. At the same time we point to the miscarriage of justice sample. Table 113.7 shows five classes and their opposite discriminated analysis's results.

Annotated: Class number is only the mark of class, only reacted on discriminating classes, not the rank of class.

113.3.3 Cluster Statistical Feature

For five cluster variables of the two methods (factor), cluster means are showed in Tables 113.8 and 113.9, where, λ value is a Wilks statistical quantity. In Table 113.8, the statistical significance of F4 is in apparent. It is obvious that with the method of between-groups linkage, differences between each category are indistinctive, so the impact of classify is also imperfect. With the method of ward's method, Table 113.9 shows that the test of equation of each variable is apparent. By this token, the impact of this classify method is favorable. There are evident differences between five categories, category can distinguish the sample better.

113.3.4 Cluster Results Evaluation

ZJMY, LTKJ, FTQC and ZHGJ are great company in payoff capability [6], operational capability, financial solvency and growing ability. They are outstanding companies. Listed Companies belonged to cluster 2, cluster 3 and cluster 5 have their own strong point. But other majority companies' performances are general, although these companies compared with others are better. So many companies belong to this kind, which indicates that it is important to increase integrated outstanding achievements, so it is a striving trend for majority Beijing listed company al the same.

Table113.9 Test for equality of cluster method two

Clu1	0.04736	-0.03573	-0.29149	-0.35109	-0.14282
Clu2	0.22652	-0.76017	0.09921	1.50195	-0.55007
Clu3	-1.45005	-0.29458	-0.12641	0.32779	3.30885
Clu4	0.26657	3.44295	-0.21546	0.73106	0.10942
Clu5	0.02329	0.25841	3.73177	-0.52997	-0.05721
λ	0.871	0.408	0.162	0.524	0.33
F	2.523	24.63	88.061	15.431	34.454
Sig	0.049	0	0	0	0

113.4 Conclusions

With advent of an economic crisis, this research has more practical meanings. It introduced multivariate statistical method to classify listed company and chose purity economical indexes. There also exists incomprehensive place, such as add experts' evaluation to listed company as a factor. On the other side, choosing the factor variable needs further discussions. It can add indexes' choice. At the same time, it may be influent the equilibrium of cluster result. For existing problems, deep analysis research is worth while.

References

1. Zang YT, Fang KT (1982) Multivariate statistical analysis introduction. Science Publishing Company, Beijing
2. Qun HX (1998) Contemporaneity statistical analysis method and application. Ren Ming University of China Publishing Company
3. Min ZX (2001) Financial report analysis of listed company. Foreign Economic and Trade Publishing Company
4. Min MJ (2002) Financial evaluation of China company. Financial Economic Publishing Company in China
5. Gu DF (2002) Synthesis evaluation of listed company. Shang Hai Financial University Publishing Company
6. Lin YX, Song RX (2003) Multivariate statistical analysis. China Statistical Publishing Company, Beijing, pp 301-303

Chapter 114

Research of the Timer Granularity Based on Linux

Geng Qingtian, Jiang Nan, Zhao Hongwei and Liu Junling

Abstract Time interrupts play an important role in the system. When time interrupts occur, system will inspect process running state, providing an opportunity to schedule, which is important for improving system real time performance. For the low system real-time ability of Linux, the paper adopt enhancing time granularity to improve system real time ability by means of one-shot triggering mode. The kernel of one-shot mode time interrupt is that: the system need not have to respond to any change every time, but only make response when having to respond. Experimental results show that the system real-time ability of Linux is improved through enhancing time granularity.

Keywords Computer applications · Real-time · One-shot mode · Timer granularity · Linux

G. Qingtian · J. Nan · Z. Hongwei (✉) · L. Junling
Department of Computer Science and Technology, Jilin University,
130012 Changchun, China
e-mail: zhaohw@jlu.edu.cn

G. Qingtian
e-mail: qtgeng@163.com

G. Qingtian
Changchun Normal University,
130032 Changchun, China

Z. Hongwei
Key Laboratory of Symbolic Computation and Knowledge Engineering
for Ministry of Education, Jilin University,
130012 Changchun, China

114.1 Preface

Linux is essentially a time-sharing operating system but it is lack of real-time, for example rough clock granularity, low real-time, non-preemptive kernel, virtual memory management technology and delay and uncertainty due to inhibit interrupt [1]. These problems are unacceptable. To enhance real-time, enhancing clock granularity is necessary.

In Linux, timing measurement is finished with clock interrupt. Clock interrupt is finished with regularly sending interrupt messages from PIT [2], and is generant according to fixed cycle. Process scheduler is regularly called. The function `timer_interrupt()`, which calls `do_timer_interrupt()`, manages the information about time and checks whether the process scheduler is executed. Therefore system clock is very important for real-time of system.

In Linux, system clock is measured though HZ. In the kernel of version 2.6, the default of HZ is 1,000 [3, 4], i.e. one interrupt is made every 1 ms. The clock granularity doesn't satisfy real-time in some circumstances, for example aerospace industry. Recently clock system is modified, and this can improve real-time capability of Linux. Therefore clock granularity is refined, which can improve the real-time of system, in this paper.

114.2 Implementation of Clock Granularity Detailing

114.2.1 Choice of Optimization Scheme

Clock frequency is increased in the kernel of version 2.6. To improve real-time performance, the first method is to increase the value of HZ. Some concepts as follows in this paper:

- (1) Clock tick
Clock tick is the difference between starting and the first interrupt.
- (2) Clock frequency (HZ)
Clock frequency is the number of interrupt that is made by PIT in one second.
- (3) Macro LATCH
LATCH is that the clock cycle frequency divides the clock frequency.

In `arch\i386\kernel\timers\time_pit.c`, the function `setup_pit_timer()` is defined.

```
void setup_pit_timer(void)
{
    unsigned long flags;
    spin_lock_irqsave(&i8253_lock, flags);
    outb_p(0 × 34,PIT_MODE);
    udelay(10);
    outb_p(LATCH & 0xff, PIT_CH0); /* LSB */
}
```

```

udelay(10);
outb(LATCH >> 8, PIT_CH0); /* MSB */
spin_unlock_irqrestore(&i8253_lock, flags);
}

```

From the function, as same as the general steps of operation to a chip, the chip is initialized by PIT_MODE [5], which can make it work according to mode 2. Then the value of LATCH is transferred into the counter of channel 0 at twice. 8,254 can be used after initialization is achieved. And this function is called when the mode of timer and interrupt frequency are set.

The clock cycle of system is shortened by increasing frequency, after the function is called. When the frequency is increased by the small amplitude, the system time slice is increased, the response time of tasks is shortened, and real-time performance is enhanced. But the system is not stable after the frequency is changed, because the parameter of some program is related to the frequency. If clock frequency is sequentially increased, the stability and performance of system is suddenly reduced, even the phenomenon of Suspended animation appears. Therefore the idea that improving the performance only depends on increasing the clock frequency is impossible.

The reason: if clock frequency is increased, the number of system clock interrupt is more, and more tasks can be executed by system. But the tasks of saving scenes are much more, which responds interrupt. More tasks badly affect the performance, so the system performance is instead reduced.

Based on an advanced idea of nonperiodic clock from KURT system original timer is modified in this paper [6].

(1) Periodic trigger mode

This is the default mode of Linux that an IRQ is termly sent to CPU from PIT. Although real-time performance is not very good, the design and implement is comparatively simply and a few materials are occupied. As a whole, system performance is enhanced.

(2) One-shot mode

This is nonperiodic clock interrupt mode based on the time when clock interrupt time happens. When it is time that is set in advance, clock interrupt happens and the related tasks are executed. Because the process status of system is not examined to ensure whether the clock is changed or not all the time when the system is running.

In application, the aim to design the one-shot mode for the system is that the tasks can be responded and executed if the tasks need to be executed.

114.2.2 Implement of Optimization Scheme

If the kernel is modified using one-shot mode, the data structure of kernel timer is used. The data structure is in the file “\include\Linux\timer.h”. To enhance the

timer granularity, the file has to be modified. And new data structure is redefined as follows:

```
struct timer_list
{
    struct list_head entry;
    int oneshot;
    unsigned long expires;
    unsigned long expires_tsc;
    unsigned long magic;
    void (*function)(unsigned long);
    unsigned long data;
    struct timer_base_s *base;
};
```

In original kernel, kernel timer is used in the driver program, and the statue of special time is examined by the equipment. The equipment is controlled in this way. It is excellent that the program is executed or the process is awoke after certain time. The timer granularity is detailed using kernel time in the one-shot mode.

It is impossible that timer granularity is not achieved using only one 8254. If there is only one 8254, clock frequency is not arbitrarily altered when the system is running. Higher precision value is not defined for the system, so it can't be judge whether the timer is used before next clock interrupt.

For solving this question, time stamp counter is used to time for the system in optimization scheme. The signal of processor is used by the counter, so it can be a high frequency and an exact time is received. But the system clock interrupt is greatly increased, and the cycle of clock interrupt is still 1 ms in application. In addition, another variable jiffies_to_tsc is defined, which value is the number of 1 tick time stamp counter.

When the system is initialized, initialization timer, which is called by start_kernel() in main.c, makes the kernel timer work. Then all the oneshots of kernel timer are set 0. This shows that the kernel timer is a cycle kernel timer. Particular arithmetic as follows:

- (1) If one interrupt happens, related services are achieved when the system is running.
- (2) In the process of dealing with interrupt, the minimal expires value in the alignment of kernel timer is compared with current jiffies +1.
- (3) If the value is greater than or equal 0, one-shot mode is not used. After the default clock cycle, the interrupt can happen. If the one-shot value is 1, this shows that the timer is a one-shot mode timer. The value of default clock cycle LATCH is calculated, and then 8254 is set through the value. Finally oneshot is set 0, and then return to (1).

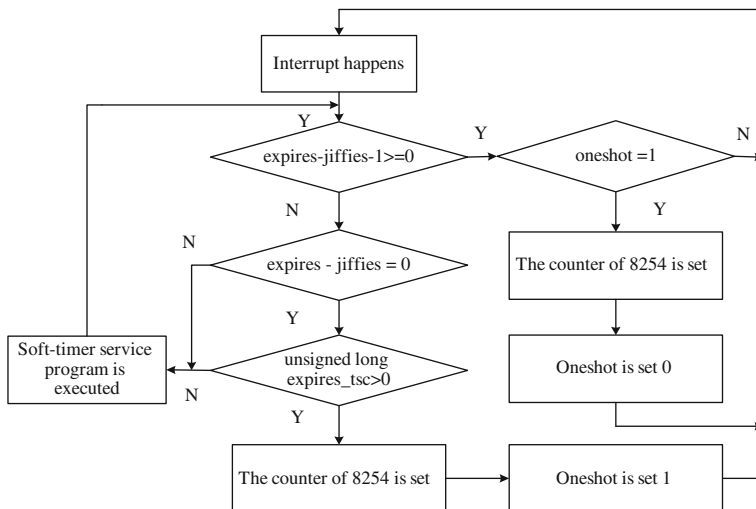


Fig. 114.1 Optimized soft-timer flow chart

(4) If expires is unequal to jiffies, the reason is that interrupt is shut down to protect some important critical regions when last interrupt happened. So the timer is not checked in the interrupt, and the timer is overdue. And return to (2), after the timer service program is executed. If expires is equal to jiffies and the value of unsigned long expires_tsc is greater than 0, this shows that a one-shot mode interrupt will happen in clock cycle. The LATCH value is recalculated based on the time of last interrupt, then 8254 is set through the value. Finally oneshot is set 1, and then return to (1).

The process is shown in Fig. 114.1.

114.3 Experiment

Experiment environment: CPU: P4 2.0, memory: 512 M, kernel: Linux 2.6.14 [7, 8].

For checking the performance after modifying, the system is tested. The kernel needs to run through the steps as follows:

Some files need to be cleared with “#make clean” and “#make mrproper”, otherwise the bad effect might appear when the kernel is compiled;

The kernel is set with “#make menuconfig”, and certain item is chosen according to actual situation;

The dependence and the integrality of the kernel code are checked by calling the related commands after configuration. And the kernel code is compiled with compiler gcc, then the new kernel is transformed into zImage;

The new kernel file is copied to boot folder, and the config file is started.

If the kernel is replaced, the system can run after restarting.

The count of timestamp counter is read by the macro rdtsc (low, high) supplied by Linux kernel, and the macro is defined as follows:

```
#define rdtsc (low, high) _asm_ _volatile_ (“rdtsc”:”=a” (low),”=d” (high))
```

The value of counter in 8254 is read, and the value LATCH is sent to channel 0. After the situation, the value of LATCH is reduced, and the next clock interrupt is happened until the value of LATCH is 0. And the value is set LATCH and is reduced. The time that the interrupt happens is calculated after reading the count of channel 0.

```
rqreturn_t timer_interrupt(int irq, void *dev_id, struct pt_regs *regs)
{
    unsigned long lost,delay;
    int count;

    rdtsc(last_tsc_low, last_tsc_high); //read the value of TSC
    spin_lock(&i8253_lock);
    outb_p(0 × 00, PIT_MODE);
    count = inb_p(PIT_CH0);
    count |= inb(PIT_CH0) << 8;
    spin_unlock(&i8253_lock);
    count = ((LATCH-1)-count)*TICK_SIZE;
    delay_at_last_inerrupt = (count + LATCH/2)/LATCH
    if(only one interrupt)
    {
        printk(delay_at_last_inerrupt);
    }
}
```

In experiment, if the original process loses the system control, spot protection will start. And then the new process gets the system control and its instruction is executed. The time from spot protection to new instruction execution is called context switch timing. This is an important index. If the average time is short, it shows that the new work will be executed quickly. Context switch speed can show the real-time ability of the system.

To test the respond time of system, the experiment is achieved through pipe and a parent and child process. The pipe is built using pipe() function, and the token is transmitted through the pipe. The pipe that is built by the pipe() function is half-duplex, so the two pipe are built for communication.

The data transfer mode in the pipe is first in first out, scilicet the data is read from one side, and the data is left from the other side. The default mode is that input and output are coinstantaneous. For example, for one pipe fd, fd[0] is input terminal, and fd[1] is output terminal. The token is transferred between parent

Table 114.1 Context switch time table

Test objects	The shortest switch time (ms)	The longest switch time (ms)	Average switch time (ms)
Standard Linux kernel	23	52	32
Improved Linux kernel	18	44	25

process and child process, and the process is forcibly switched through uninterupted self-block. Then the system time is read and recorded through the function `rdtsc()`, and the time of switching process is counted. The time that the data is read and written in the pipe is not counted, since this time is very short.

In the Linux platform, and improved platform to experiment and record data were, the results shown in Table 114.1.

114.4 Conclusion

A new schema of enhancing time granularity is proposed in this paper. Through analyzing the experiment data, enhancing time granularity using the method of one-shot mode is equal to increasing the clock frequency. This make the system respond the change quickly, and the real-time performance is enhanced.

Acknowledgements The paper is supported by Specialized Research Fund for the Doctoral Program of Higher Education(20050183032) and Science Foundation of Jilin Educational Committee(2009604)and Changchun Normal University Natural Science Foundation

References

1. Lijun C (2007) Understanding the LINUX kernel [M]. China Electricity Press, Beijing 9:7
2. Oikawa S, Rajkmar R (1998) Linux/RK: A portable resource kernel in Linux[J]. IEEE real-time systems symposium work-in-progress, pp 43–47
3. Xiao-Qun LL, Hui-Bin Z, Yi-Min Y et al (2003) RFRTOS: based on linux real-time operation systems[J]. Software 14(7)
4. Corbet J, Hartman KG, Rubini A (2005) .Linux device drivers, 3rd edn. O’Reilly publishing, 345–347
5. The interrupt mechanism of Linux kernel clock.14-15, <http://www.docin.com/p-1624090.html#docTitle>
6. Leucht A, Lineo (2002) Hard real-time capability under Linux. Electron Eng Design 2:50–54
7. Qin H, Hongtao W (2008) Linux2.6 Kernel standard tutorial, vol 10. Posts and Telecom Press [M], Beijing, pp 323–324
8. Barabanov M, Yodaiken V (1997) Introducing real-time Linux [J]. Linux J 34(2):19–23

Chapter 115

A New Software to Realize the Optimization of City Sound-Planning

Ji Qing

Abstract This new software can be considered as a new method to deal with city sound-planning problems. By the traditional method, specialists arrange sound spaces of a project in zooms with different noise exposure levels. But the new method here considers all the sound spaces of a project as a whole. It arranges them in site towards minimizing the sound conflicts caused by all neighboring sound spaces, and realized this process in software automatically. Three new theories about how neighboring sound spaces interact on each other help us find out and establish the optimization strategy of city sound-planning in software. In this process, a new database that can help us represent different city sound spaces is indispensable. Nowadays, by the help of this new software, normal users can get the sound-planning references of their own project easily and effectively.

Keywords Software · Optimization · City sound-planning

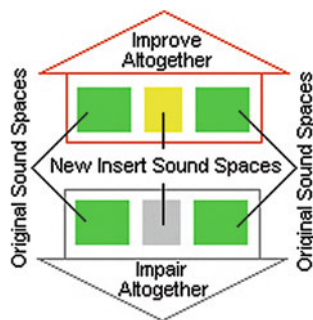
115.1 Initial Clue Come from Questing a City Sound Phenomenon

A city sound phenomenon has often been observed by specialists in previous researches. This phenomenon can be explained by Fig. 115.1 [1], for some organization ways of sound spaces, by the interactions between them, the whole quality of their sound environment will improve. But for others organization ways,

J. Qing (✉)

State Key Laboratory of Subtropical Architecture Science, South China
University of Technology, Guangzhou, 510640 Guangdong, China
e-mail: sboxi@163.com

Fig. 115.1 Different organization ways of sound spaces lead to different results



by the interactions between them, the whole quality of their sound environment will impair. By the knowledge gained from the previous researches, we can find out the emerging conditions and internal social dynamic of the sound phenomena happened in a single sound space easily, but we still know little for the mechanism of interaction between two neighboring sound spaces.

In order to expand our knowledge about how neighboring sound spaces interact on each other. Several researches had been carried out in Canton and in France in the past several years [2, 3]. In Canton, a series of case studies about different organizations ways of sound spaces had been investigated. These cases are typical samples and come from a stable residential district. Then, the following research focused on responding two questions. When two sound spaces are on neighborhood, by which channels can they realize the influences to each other? How to consider two neighboring sound spaces as a whole? The resolution of them can, not only help us comprehend the phenomenon mentioned above, but also give us a clue to optimize the sound-planning of our city automatically.

115.2 Three New Theories to Describe the Interactions Between Neighboring Sound Spaces

By analyzing all critical sound situations being observed in the case studies and comparing the different results brought by changing different neighboring “partner” of certain sound spaces, researchers established three new theories to describe the interactions between neighboring sound spaces.

Interaction Channels. Researchers found that almost all interactions between two neighboring sound spaces can be classified into five channels as following: degree of permeation, degree of public, degree of nature, collective memory and degree of insecurity. But only first three of them exist differences in degree. Collective memory and degree of insecurity seem to be narrative.

Sound Requirement and Sound Contribution. This theory can help us consider two neighboring sound space as a whole. But what’s more important, it tells us that

the status of a sound space sometimes may be unstable. When a sound space neighbors with different “partner”, its status may change in the process of interaction. So a sound space may have two statuses, the inborn status and the acquired status. If we want to determine the sound requirement and the sound contribution of a sound space, we must distinguish what status it belongs to firstly.

Time-gear. This theory is used to describe the process of interaction between neighboring sound spaces on considering the time goes on in a day. It involves three aspects: firstly, the sound requirement and the sound contribution of a sound space may keep on changing in different hours of a day. Secondly, the sound requirement and the sound contribution of a sound space are determined by a sound event which dominate the sound requirement and a sound event which dominate the sound contribution at each hour. These two sound events may be different or may be the same one. Thirdly, according to city sound reality, in each hour, the sound requirement of a sound space can be absent, but the sound contribution of a sound space cannot be absent. If there is not an obvious sound event to dominate the sound contribution, researcher suggests to use a new sound event called “calm” to represent the sound contribution in this hour.

115.3 A New Database to Represent Sound Spaces

These new theories show us the mechanism of the interactions between neighboring sound spaces. So, people get to understand the phenomenon mentioned at the beginning of this article [4]. Conversely, can people utilize these theories as tools to realize some further practical applications, such as to optimize the sound-planning in computer automatically? It seems possible. But the first step is to create a new database, which can help us represent each sound space and simulate the organization ways of sound spaces in computer effectively.

With this purpose, a new database had been created following guides of the three theories above. Firstly, three interaction channels (degree of permeation, degree of public and degree of nature), which can be measured in degree and recognized by computer, had been used to reflect the interactions between neighboring sound spaces. Secondly, to represent a sound space in database must consider its sound requirement and sound contribution at the same time. So it can be linked with the others. Thirdly, since the optimization of a sound-planning always takes place in the design period of a project, so we only consider the “inborn” status of each sound space in the database. Fourthly, in database, the sound requirement and the sound contribution of a sound space ought to evolve according to different hours. Fifthly, the sound requirement and the sound contribution of a sound space in each hour are determined by the dominating sound events. All the sound events that are not in dominating position will not be considered. Finally, in the database, a new sound event “calm” had been used to fill in the hours that the sound contribution is absent [5].

Table 115.1 Sound space form (take “garden with pavilion” as example)

Hour	Sound event	Sound requirement			Sound contribution		
		I	P	N	I	P	N
0–7	Insect voice				43	–2	2
7–8	Insect voice/Morning exercise	55	0.47	1.53	43	–2	2
8–9	Play mah-jone/Pedestrian	55	0.53	0.73	54	1.33	–1.1
9–10	Play mah-jone/Pedestrian	55	0.53	0.73	54	1.33	–1.1
10–11	Play mah-jone	55	0.53	0.73	54	1.33	–1.1
11–12	Pedestrian	55	0.53	–0.7	51	1.07	0.60
12–13	Pedestrian	55	0.53	–0.7	51	1.07	0.60
13–14	Play mah-jone	55	0.53	0.73	54	1.33	–1.1
14–15	Play mah-jone	55	0.53	0.73	54	1.33	–1.1
15–16	Play mah-jone	55	0.53	0.73	54	1.33	–1.1
16–17	Pedestrian	55	0.53	–0.7	51	1.07	0.60
17–18	Pedestrian	55	0.53	–0.7	51	1.07	0.60
18–19	Rest/Insect voice	40	–0.6	1.47	43	–2	2
19–20	Rest/Insect voice	40	–0.6	1.47	43	–2	2
20–21	Rest/Insect voice	40	–0.6	1.47	43	–2	2
21–22	Rest/Insect voice	40	–0.6	1.47	43	–2	2
22–23	Insect voice				43	–2	2
23–24	Insect voice				43	–2	2

Note: “I” refers to degree of permeation (Unit is dBA). “P” refers to degree of public. “N” refers to degree of nature. For degree of public and degree of nature, “–2” to “2” means the level from the lowest to highest

Table 115.1 is the example of sound space form in the database. Each sound space will have its own sound space form and be represented in computer by this means. As we can see in Table 115.1, in this data form, the sound event dominating sound requirement and the sound event dominating sound contribution in each hour have been recorded. The sound requirement and the sound contribution brought by these sound events are described by six reference values, which cover three interaction channels. To degree of permeation, the reference value is reflected by L_{eq} (dBA). The values that belong to the sound contribution are obtained by acoustic measurement in site and the values that belong to sound requirement determined by national norms. To degree of public and degree of nature, all the reference values are obtained by the sound-reaction test technique, since these two interaction channels concern human sense. After listening the sound record of a sound event, testers will judge the degree of public and the degree of nature belong to which level (five reference values –2, –1, 0, 1, 2 had been set to indicate the level from the lowest to the highest). In order to improve the practical ability, these sound space forms are adjustable. Users can regulate the appearance time and appearance type of its sound events following their project’s reality.

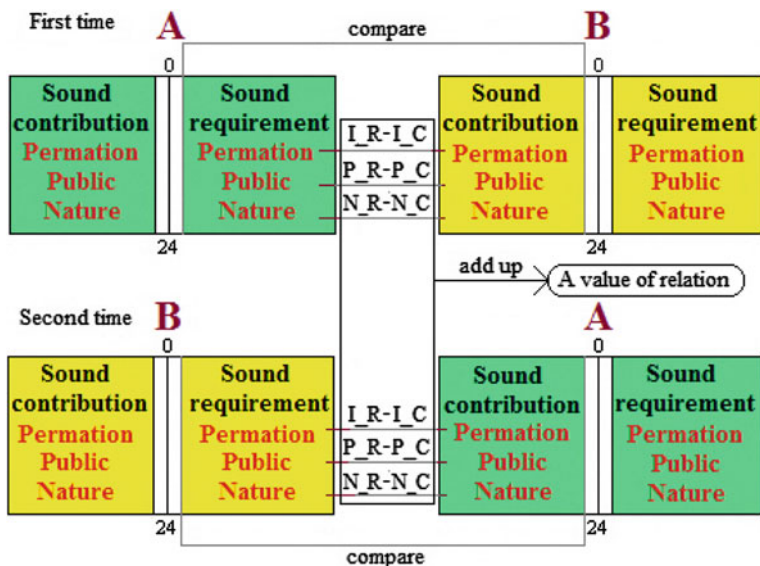


Fig. 115.2 We can use a value of relation to evaluate the matching level of two neighboring sound spaces by comparing their sound space form

115.4 Realization of the Optimization of City Sound-Planning

The optimization of city sound-planning cannot be realized in database directly [6]. we must depend on a computer program to finish a series of calculations. In the program, a data channel must be established to connect the database we mentioned above, since the optimization must use the sound space forms. Specifically speaking, the program use four steps to realize the optimization of city sound-planning as following:

Firstly, the user ought to tell the program what kinds of sound spaces he wants to deal with in his project and each sound space’s quantity.

Secondly, following the information provided by the user, the program would call the sound spaces forms concerned from the database and match each two of them. The matching result of each two sound spaces will be presented by a value of relation. In detail: for example, to match the sound space “A” and the sound space “B” (such as Fig. 115.2). For the first time, the program will match the sound requirement of “A” and the sound contribution of “B”, and obtain three values of difference from three interaction channels. For the second time, the program will match the sound contribution of “A” and the sound requirement of “B”, and obtain the other three values of difference from three interaction channels also. After two times’ matching, we will have six values of difference for each hour. Then, there are 6×24 values of difference for a day (24 h). The sum of

City road	City road	City road	City road	City road
City road	Vacancy	Vacancy	Vacancy	City road
City road	Vacancy	Vacancy	Vacancy	City road
City road	Vacancy	Vacancy	Vacancy	City road
City road	City road	City road	City road	City road

Fig. 115.3 An example of construction site

these 6×24 values of difference is a value of relation. All the values of relation concerned in the project can be presented by a matrix of relation.

Remark 2 Since the values from degree of permeation is presented by dB(A) , the values from degree of public and degree of nature are presented by “ $-2 \sim 2$ ”, we ought to put them in the same scale by adding a coefficient $[2 - (-2)] / (70 - 40)$ to the values from degree of permeation, because the values from degree of permeation in database, the highest one is 70 dB(A) , the lowest one is 40 dB(A) .

After the program has established the matrix of relation of the project, we ought to add a new sound space called “vacancy” in this matrix. This is necessary, since at most of the cases, the vacancies that we can use in a project may be more than necessary. The proprietors always want to remain a few vacancies for the development in future. So what kinds of sound space will be installed in these remaining vacancies is not sure in period of design. In the program now, no matter what kinds of sound space matching with the “vacancy”, the value of relation is fixed to 73. “73” is an average value of all the values of relation on considering all the sound spaces in database now except those sound spaces which are something impossible to be installed in the remaining vacancies such as “city road”, “crossroads” and “city highway”. By the way, the value “73” can be changed by the user, because the user can narrow the possible kinds of sound space following the reality of his project.

Thirdly, The user ought to tell the program the situation of the construction site of the project. It includes telling what appearance the site is, how to divide the vacancies in this site and what sound spaces surround it. In the program now, the situation of the site is presented by matrix. Such as Fig. 115.3, it shows the situation of a site set by user. It is a rectangle construction site, which is surrounded by the sound space “city road”, and divided into 3×3 vacancies. This is a simple example. In fact, the user can set a site bigger (such as 4×4 , 5×5 , 4×5 etc.) and more complicated (the appearance may be irregular) than this one. But the surrounding sound spaces of the site must be sure. It is said that the site must be encircled by knowing sound spaces.

Fig. 115.4 The matrix of relation of the example project

	②	⑧	⑬	⑰	⑳	㉓	㉕
②	54	162	56	64	105	73	73
⑧	162	45	16	60	39	77	73
⑬	56	16	0	11	16	31	73
⑰	64	60	11	35	75	61	73
⑳	105	39	16	75	27	81	73
㉓	73	77	31	61	81	47	73
㉕	73	73	73	73	73	73	73

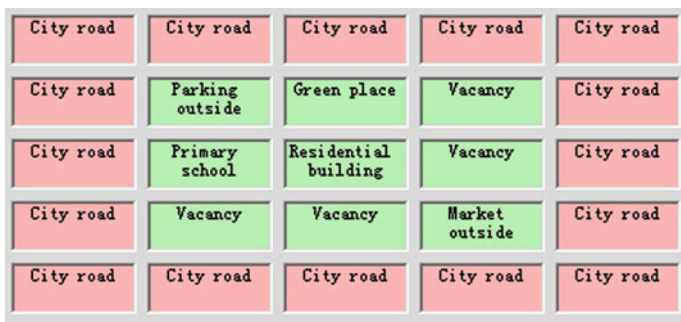


Fig. 115.5 The optimization result of the example project issued by the program

Finally, basing on the matrix of relation of the project, the program can use a linear optimization equation to calculate all possible organizations way of the sound spaces in this site and find out the best ones.

Now, we use an example to show how to realize the optimization of city sound-planning. In this example, we will organize five different sound spaces (one residential building, one parking outside, one green place, one market outside and one primary school) in a site like Fig. 115.3. The matrix of relation of this sound-planning project is shown in Fig. 115.4 and the optimization result issued by the program is shown in Fig. 115.5.

Remark 4 In this matrix of relation, “2” refers to residential building, “16” refers to parking outside, “19” refers to green place, “21” refers to market outside, “23” refers to primary school, “8” refers to city road and “25” refers to vacancy. By this matrix, computer can master the matching level of each two sound spaces engaged in this project.

From the optimization result issued by the software (Fig. 115.5), we can find that the “Residential building” is arranged in central of the site, which belongs to the quietest zone of the site. “Parking outside” is arranged at a corner, since it has not sound requirement. “Market outside” as a noisy sound space is arranged at a corner. The arrangement of “Green place” and “primary school” at two edges not

neighboring with the “market outside” seems to be reasonable also. They like to neighbor with “parking outside” rather than “vacancy”, since the future use of “vacancy” is not sure yet. This result as the best organizations ways of sound spaces in this example project is pertinent.

115.5 Prospects

Benefited from the calculation ability of the computer, we can realize the optimization of sound-planning by thousands matching automatically, which is unimaginable before. Even a very simple project, such as the example above, there exists 15120 ways of organization. With this new method, the sound planning of the big project comes to realize in near future.

By the way, we must pay attention to a new problem following. Since a successful urban planning is combined the intelligences from many special fields (such as light using, thermology, the circulation of pedestrian, the circulation of vehicle and the arrangement of function). The best organization way of sound spaces may be not the best one for the other special fields. So, in practical use, the reasonable way is to issue several good organization ways from the software, for example, to issue the top ten best results. Then, the designer can make an overview decision by considering the requirement from the other special fields.

References

1. Augoyard J F(1999). Du Bruit à L'environnement Sonore Urbain. J. Données urbaines n°3. France: Paris, Ed.Anthropos
2. Augoyard J F (1995) L'environnement sensible et les ambiances architecturales. J. L'espace géographique n°4
3. Thibaud J P (2002) Comment Observer Une Ambiance?. J. France: Ambiances Architecturales ET Urbaines, 23:77–81
4. Amphoux P (1991) L'identité Sonore des Villes Européennes___Guide Méthodologique. Cresson, France
5. Balay O, Arlaud B (1999) La Représentation de L'environnement Sonore Urbain à L'aide d'un Système d'Information Géographique. Cresson and LISI, France
6. Balay O (1999) Les Indicateurs de L'identité Sonore d'un Quartier. Cresson, France

Part XI
Graphics and Visualizing

Chapter 116

Automatic Classification and Recognition of Particles in Urinary Sediment Images

Yan Zhou and Houkui Zhou

Abstract In the past, cast cells in urine sediment were recognized and sorted mainly by human. We proposed an automatic method for classification and recognition of particles mainly white blood cell (WBC) and red blood cell (RBC) in urinary sediment. It is composed of three stages: First, Original urinary sediment microscopic images are transformed into binary image by image pretreatment including median filtering, color image conversion to gray scale image and image segmentation. Second, we select and extract some features as feature vectors for classification and recognition. In the last, eleven texture and shape characteristics of casts are extracted from both gray scale image and binary image. Based on these characteristics, we develop an SVM classifier to distinguish casts from other particles in the image. Experimental results show that our method achieves an easy-implemented classifier and has good recognition performance.

Keywords Feature extraction · Feature selection · SVM · Classification

116.1 Introduction

Urinalysis is crucial for the diagnosis of renal and urinary tract diseases. It aims to obtain quantity information of particles in the urine sediment which is mainly composed by red blood cells (RBC), white blood cells (WBC), epithelial cells (EP), yeast-like cells, hyaline or pathological casts, crystals and mucus [1]. Among these particles, cast cells are extremely important since as long as there exist cast cells,

Y. Zhou · H. Zhou (✉)

College of Information Engineering, Zhejiang Agriculture and Forest University,
Hangzhou, 311300, People's Republic of China
e-mail: zhouhk@zafu.edu.cn

even only one of them in the patient's urine, doctor can assure that there is something wrong [2]. The tradition detection method of the visible urinary sediment components uses manually microscope observation or takes pictures and then identifies artificially the result. This method has many disadvantages such as subjective, different standard, heavy workload and difficult getting the conclusion and so on.

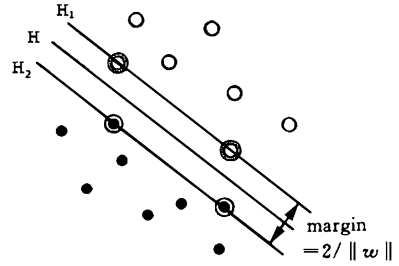
There are some literature concerning segmentation and classification based on images. Li et al. [3] used Gabor-based combining texture segmentation method to segment the particles from background in the image. Li et al. [4] came up with a strategy for segmenting urinary sediment based on wavelet, morphology and combination method. Zhang et al. [5] employed cellular neural network to segment targets from images after being preprocessed. Both Jiang et al. [6] and Luo et al. [7] introduced improved level set in combination with Mumford-Shah model for segmentation. In aspect of classification, some methods were proposed for classifying the substances in the urinary sediment image. Murasaki et al. [8] recognized the substances using the fuzzy-neural networks in which the features membership functions are also gotten from the binarized images. Zeng et al. [9] categorized particles of interests by neural networks and fuzzy reasoning. Dong et al. [10] developed a naïve Bayesian classifier after distinguishing the visible compositions, feature extraction and feature selection for classifying of object entities. Not all particles locate in the same depth. Different cells have varied reflecting property even in the same depth. Moreover, there exists uneven illumination problem. Previous methods gave birth to unsatisfactory segmentation results due to above mentioned facts when it comes to casts. In addition, recognition stage needs efficient classifier for time saving and real time application. Consequently, a sufficient classifier with high efficiency is needed.

We explore a novel scheme for casts classification and recognition in Microscopic urine sediment images. In the pretreatment stage, a combination of segmentation methods is presented, which is include two steps: first acquire the casts location in the image and extract every cell all alone respectively and then adopt adaptive bi-threshold segmentation algorithm. As for the recognition, we bring forward a SVM classifier with respect to texture and shape characteristics of casts extracted from both gray value image and binary image, to distinguish casts from other particles. Experimental results show that our strategy is easy-implemented, time-saving, and displaying pleasant performance. In Sect. 116.2, some theoretical used in the chapter will be bring forward. In Sect. 116.3, we design a classification and recognition algorithm. In Sect. 116.4, experiment results will be got using recognition method introduction in Sect. 116.3. In the end, we give some conclusions and look forward to some feasible future work.

116.2 Theoretical Background

A key problem is how to construct the classify function (classifier) when using SVM classifies automatically the urinary sediment [8]. The un-proposed sediment is classified by the classify function. The classify function can be obtained by some

Fig. 116.1 OSH and the maximum margin



learning algorithm and SVM shows the better performance in the research of urinary sediment classifier learning algorithms. So, this chapter adopts SVM learning algorithm to classify urinary sediment visible components.

Support Vector Machines (SVMs) provides a novel means of classification using the principles of structural risk minimization (SRM; Vapnik 4, 5, 1998) [9, 10]. It is one of the most sophisticated nonparametric supervised classifiers available today, with many different configurations depending on the function used for generating the transform space in which the decision surface is constructed. In a binary classification task in this study, the aim is to find an optimal separating hyper-plane. In Fig. 116.1, an Optimal Separating Hyper-plane (OSH) is shown which generates the maximum margin (dashed line) between the two data sets. SVM firstly transforms input data into a higher dimensional space by means of a kernel function and then constructs a linear OSH between the two classes in the transformed space. Those data vectors nearest to the dash line in the transformed space are called the Support Vectors (SV). SVM is an approximate implementation of the method of “structural risk minimization” aiming to attain low probability of generalization error [11–13].

Consider a training set $(x_i, y_i), i = 1, \dots, n$ with each input $x \in R_d$ and the associated output $y \in \{-1, +1\}$. Each input x is first mapped into a higher dimension feature space H via a nonlinear mapping $\varphi: R_d \rightarrow H$. Considering the case when the data are linearly separable in H ; there exists a vector $w \in H$ and a scalar b that define the separating hyper-plane as $x \cdot w + b = 0$. Such that

$$y_i[(w \cdot x_i) + b] - 1 \geq 0 \tag{116.1}$$

By maximizing the margin of separation between the classes $(2/\|w\|)$, SVM constructs a unique OSH as the one that minimizes $w \cdot w/2$ under the constraints of Eq. 116.1. When the data are linearly non-separable, the above minimization problem is modified to allow classification error by introducing some non-negative variables $\xi_i \geq 0$, often called slack variables, such that

$$y_i(w \cdot z_i + b) \geq 1 - \xi_i, \forall i \tag{116.2}$$

A non-zero ξ_i indicates a misclassified data point and can be regarded as a measure of misclassification. Though new kernels are being proposed by researchers, beginners may find in SVM books the following three basic kernels:



Fig. 116.2 Automatic recognition system

Polynomial: $K(x_i, x_j) = [(x_i \cdot x_j) + 1]^d$, d is the degree of polynomial; Exponential radial basis function (ERBF):

$$K(x_i, x_j) = \exp\left\{-\frac{\|x_i - x_j\|}{2\sigma^2}\right\} \quad (116.3)$$

σ is the width of ERBF function; Multi-layer perception (MLP):

$K(x_i, x_j) = \tanh(b(x_i \cdot x_j) - c)$, b is the slope and c is the bias. In this chapter, the Exponential radial basis function is selected and processes the multi classification question.

116.3 Algorithm Design

The system is composed by the image sampling hardware system and image processing identification software system. In order to inspect the urinary sediment visible components, first, microscope enlarges the visible components in the urinary sediment quantitative analysis board. Then, light signals is transformed to video electrical signals by CCD camera and the image sampling board samples and saves it as files. The software system processes the sampled image in computer and gets the analysis result. The automatic recognition system is depicted by Fig. 116.2.

We design a novel classification and recognition algorithm to sort urinary sediment visible component. This algorithm divided into two processes. One is training process which teach the computer recognizing different cells in urinary sediment images and acquire parameters of the classifier for reorganization process. Another is reorganization process which is responsible for recognizing the different particles in the images.

Training process is shown as follows:

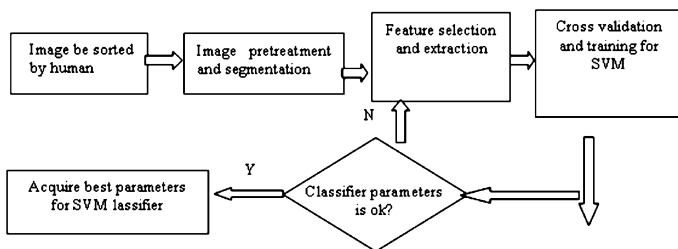


Fig. 116.3 Flowchart of training process for each substance

Figure 116.3 shows the flowchart of training model for each substance. In order to get the classifier parameters, we should establish a samples dataset of the urinary sediment particles. This dataset include substance belonging to each kind of particle and not belonging to them. The reorganization of each particles conduct by human. After we get the sorted dataset including each kind of particles, we can start training process. Then the sorted particles be pretreated and segment, furthermore each cell feature be selected and extracted and feature vectors are got. Further again the feature vectors are inputted into the SVM classifier for training and cross validation, if the parameters satisfy, then we get the right parameters; otherwise go to the feature selection and extraction step.

Testing method for each substance is shown in Fig. 116.4.

Figure 116.4 shows the flowchart of recognition model for each sample. First, a test sample are pretreated and segmented into binary image. Then, feature selected are extracted from both binary image and gray scale image. After that, feature vectors of the test sample are inputted into the SVM classifier for recognition. Last, recognition results are gotten.

116.4 Experiment Results

We took two substances as examples. One is white blood cell and another is red blood cell. Our method is not only to classify between white blood cell and red blood cell [14], but also to separate the white blood cell or red blood cell from the all of the substances. Some typical samples of red blood cell and white blood cell are shown in Fig. 116.5.

From Fig. 116.5 we know that basically red blood cells are smaller than white blood cell in size. But there are also some exceptions, in Fig. 116.4 red blood cell (a) is similar with some white blood cell in size. Except that, red blood cell has clear edge and like a circle plate while white blood cell has vague edge and their gray scale distribution don't balance. So shape features and texture features are selected as the distinguish features for recognition. In this chapter, we totally select 14 features as classifier parameters. These features are list in Table 116.1.

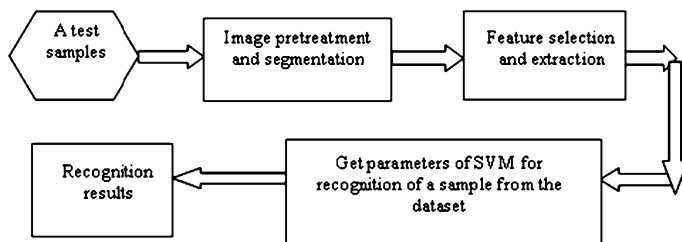


Fig. 116.4 Flowchart of recognition process of a sample

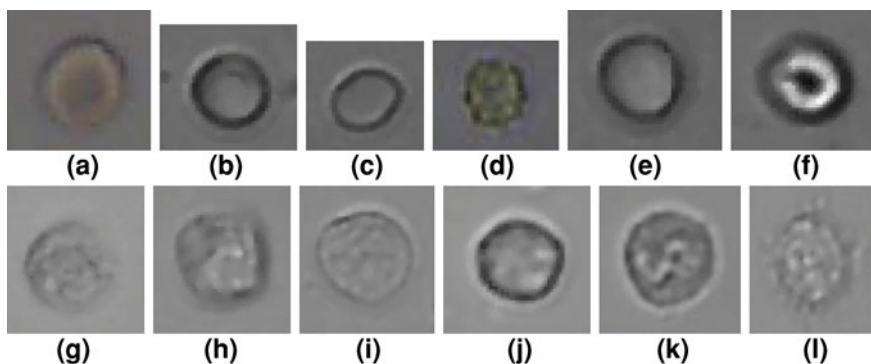


Fig. 116.5 Typical cell samples of red blood cell and white blood cell. a–f are red blood cell. h–m are white blood cell

SVM classifier is used for both training and recognition process in our experiment. SVM has better learning and popularize for little sample classification. It is show that this method has higher recognition rate than the traditional methods. In our experiment, the recognition rate is about 96% but the traditional method, such as nerve network; only arrive at 87% [15]. Although the SVM method has good performance, the kernel function and parameter selecting is still difficult question. Selecting different kernel function and parameter, the experiment has the different results. We select four main type kernel function for SVM, they are linear kernel function, polynomial kernel function, tanh kernel function and Exponential radial basis function (ERBF). We totally use 526 cell samples which have the same number of two type cells and are sorted beforehand, among which half for training and another half for testing. The correct rate of recognition using four type kernel functions respectively by SVM classifiers are shown in Table 116.2. From Table 116.2, we find that the Exponential radial basis function (ERBF) is better than others.

Table116.1 Features for classification

Shape features					Histogram statistic features	
Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7
Area	Girth	Long axis	Short axis	Shape factor	Variance	Secondary moment
Texture features		LBP features				
Feature 8	Feature9	Feature 10	Feature 11	Feature 12	Feature 13	Feature 14
Correlation	Std	Mean	Variance	Smoothness	Entropy	Coherence

Table116.2 Recogniton correct rate of four different type SVM kernel function

Kernel function	Linear	Polynomial	Tanh function	ERBF
Correct rate (%)	85.7	89.8	91.3	94.5

116.5 Conclusion and Discussion

This chapter discusses the algorithm based on computer identification system and introduces briefly work process of the urinary sediment images classification and recognition. First, we pretreat and segment the original urinary sediment image and get the binary image. Then, we decide what features should be selected according to the cell characteristic. After that, we adopt the feature extracting method to acquire the feature vectors. Last, we use SVM process to distinguish the urinary sediment styles. Experiment show that the identify performance is better than tradition methods. This system has two advantages: one is that the system operation cost is low and the speed is higher; the other is that using the SVM classifies sediment and gets the high identification rate according to the characters of urinary sediment components complexity and little samples number. This system improves the traditional diagnostic method with eyes and shows the image in big field, high resolution, and real-time property. It also offers the subject quantify measure, excellent diagnostic project and the doctor classify standardization. Through improving the algorithm this system can implement in the urinary sediment visible component classification system and has excellent application in future.

References

1. Pornvaree L, Krisana P, Pitimon L, Pimpawee S, Pornsri T, Nongnute K, Boonsong P (2007) Urine sediment examination: a comparison between the manual method and the iQ200 automated urine microscopy analyzer. *Clinica Chimica Acta* 384:28–34
2. Chien TI, Kao JT, Liu HL, Lin PC, Hong JS, Hsieh HP, Chien MJ (2007) Urine sediment examination: acomparison of automated urinalysis systems and manual microscopy. *Clinica Chimica Acta* 384:28–34

3. Li YC, Li ZC, Mei YH, Zhang JX (2005) Detecting algorithm based gabor in microscopic image. In: Proceedings of the 4th international conference on machine learning and cybernetics, Guangzhou, pp 18–21 August 2005.
4. Li YM, Zeng XP (2006) A new strategy for urinary sediment segmentation based on wavelet, morphology and combination method. *Comput Method Program Bbiomed* 84:162–173
5. Zhang ZC, Xia SR, DuanHL (2007) Cellular neural network based urinary image segmentation. *ICNC August 2007* 2:285–289
6. Jiang X, Nie SD (2007) Urine sediment image segmentation based on level set and Mumford-Shah model. *Bioinfo Biomed Eng* 1028–1030
7. Luo T, Kramer K, Goldgof DB, Hall LO, Samson S, Remsen A, Hopkins T (2004) Recognizing plankton images from the shadow image particle profiling evaluation recorder. *IEEE Trans Syst Man Cybernet* 34(4):1753–1762
8. Tanimura Y (1996) Automatic classification of white cells. *Med Img Technol* 14(1):14–22
9. Jain LC, Vemuri VR (1999) Industrial applications of neural networks. CRC press, New York, p 65
10. Shen M et al. (2005) Study on urinary sediments classification and identification techniques. 20th congress of the international commission for optics 2005
11. Shen M et al (2006) Application of support vector machine in the classification of the visible urine sediment components. *J Electron Dev* 29:98–101
12. Grace AE, Spann M (1991) A comparison between Fourier–Mellin descriptors and moment based features for invariant object recognition using neural networks. *Pattern Recog Lett* 12:635–643
13. Koenderink JJ, van Doorn J (1987) Representation of local geometry in the visual system. *Biol Cybernet* 55:367–375
14. Langford M, Taylor GE, Flenley JR (1990) Computerised identification of pollen grains by texture. *Rev Paleobot Palynol* 64:197–203
15. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proc IEEE* 86(11):2278–2324

Chapter 117

Ultrasound Strain and Strain Rate Imaging of the Early Stage of Carotid Artery with Type 2 Diabetes Mellitus

Cun Liu, Yanling Zheng, Yuanliu He, Hongxia Xu, Juan Su, Lili Zhang, Xiaohong Zhou and Changchun Liu

Abstract To evaluate the value of velocity vector image in evaluating the motorial characteristics of the early stage of the CCA in patients with type 2 diabetes mellitus (DM2). *Methods.* Fifty patients without vascular complications with type 2 diabetes and fifty healthy volunteers underwent carotid ultrasound examinations, the dynamic image was analysed by the off-line software (syngo Velocity Vector Imaging technology (VVI), Siemens). *Results.* Vmax of anterior wall, anterolateral wall and posterolateral wall were higher than those of posterior wall, posteromedial wall and anteromedial wall ($P < 0.05$). VTTP, Vmax, Smax and SRmax of corresponding segments had significant differences in study group and control group ($P < 0.05$). *Conclusions.* Velocity Vector Imaging can be used to evaluate the change of common carotid elasticity in early stage of AS in patients with type 2 diabetes.

Keywords Ultrasonography · Velocity vector imaging · Speckle tracking · Common carotid · Atherosclerosis · Diabetes mellitus

The number of people with diabetes mellitus (DM) in 2010 is estimated to be 285 million, which was approximately 7% of the adult world population [1]. Macro-

Drs Cun Liu and Mrs Yanling Zheng contributed equally to this article.

C. Liu · Y. He · H. Xu · J. Su · L. Zhang · X. Zhou
Jinan Central Hospital, Shandong University, Jinan, China

C. Liu · C. Liu (✉)
School of Control Science and Engineering, Shandong University, Jinan 250061, China
e-mail: changchunliu@sdu.edu.cn

Y. Zheng
School of Mathematical Sciences, University of Jinan, Jinan, China

vascular disease, characterized by atherosclerotic changes in large blood vessels, is the major cause of morbidity and mortality (80%) in type 2 DM [1]; cardiovascular disease (stroke) is the leading cause of death in diabetes mellitus patients [2]. The overall relative risk of stroke is 1.5–3 times greater in patients with DM [1]. Recurrent stroke is also twice as frequent in patients with DM [3]. More importantly, short- and long-term mortalities after stroke are significantly greater in patients with DM [3]. It has been well established that plaque rupture with subsequent intraluminal thrombosis is the most common cause of acute cardiovascular events [4, 5]. Identifying patients at an early stage before clinical complications such as myocardial infarction or stroke occur and assessing the total atherosclerotic burden are clinically important.

Ultrasound is a non-invasive imaging technique, which has been successfully used for the morphological and functional assessment of plaques. It also offers a reliable platform for their biomechanical assessment. Various ultrasound techniques have been used to detect and track the vessel wall motion. Recently, velocity vector imaging (VVI) has been obtained using ultrasounds [6–8]

This study is aimed at exploring the value of VVI in evaluating the motorial characteristics of the early stage of Atherosclerotic plaque on the CCA in patients with type 2 diabetes mellitus (DM2) and 30 health volunteers underwent ultrasonograph.

117.1 Methods

Fifty patients without vascular complications with type 2 diabetes referred to our department for evaluation of cardiovascular status with carotid artery ultrasound and were included in the study. Additionally, 50 age and gender matched healthy volunteers without type 2 diabetes mellitus (DM2) in the corresponding period were recruited as controls. General conditions of the patients with type 2 diabetes and the healthy subjects including age, sex, BP, HR, BMI were recorded, and blood sampling indicators such as TG, TC, LDL were examined following over night fasting. All patients gave their informed and written consent to study participation and the human part of the study was also approved by the Local Ethics Committee. All patients underwent CCA ultrasound examination. The examined subjects were in supine position, breathe calmly. Neck was fully exposed with face toward opposite side. Ultrasound was performed through transverse and longitudinal directions from up to down according to a standardized protocol over the CCA; images were stored and transferred to a computer for off-line analysis. Intima-media thickness was measured at the point 1 cm proximal to the carotid artery bifurcation from the lumen–initial interface to the medial adventitial border [9]. Atherosclerotic plaques were assessed in long axis view of the carotid bifurcation by manual delineation. Presences of plaques were defined as a focal lesion with an IMT of >1.2 mm [10]. Two-dimensional dynamic image for was collected when patient was told to hold breath for several seconds. The dynamic image was stored to be analyzed later. Concisely, the

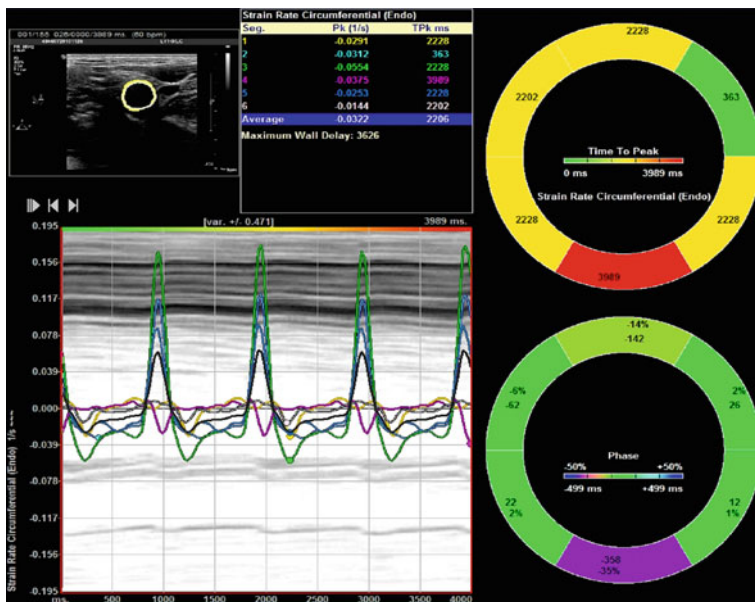


Fig. 117.1 Example of a velocity vector imaging of the CCA. Twenty-four points were marked on the endomembrane of carotid uniformly by hand, three sampling points would be added automatically between every two marked points, and the wall of carotid was divided into ninety-six small segments. The reference point was put in the center of lumina. Time to the peak of velocity (VTTP), maximum of velocity (Vmax), maximum of strain (Smax), and maximum of strain rate (SRmax) of six segment were measured

new off-line software (syngo Velocity Vector Imaging technology (VVI), Siemens) provides angle-independent 2D velocity, strain, and SR was used to derive vessel wall displacement off-line. When the photograph was frozen, twenty-four points were marked on the endomembrane of carotid uniformly by hand. Three sampling points would be added automatically between every two marked points, and the wall of carotid was divided into ninety-six small segments. The reference point was put in the center of lumina. Time to the peak of velocity (VTTP), maximum of velocity (Vmax), maximum of strain (Smax), and maximum of strain rate (SRmax) of six segment (anterolateral wall, anterior wall, anteromedial wall, posteromedial wall, posterior wall, postemlateral wall) were measured, which were supplied by VVI automatic analysis software (Fig. 117.1). EF and SV were acquired using Simpson’s method by 2DE sonography. General average of each index in 3 continual heart beating cycles was recorded.

Reproducibility of the VVI method: for intraobserver variability, the same observer reviewed the ultrasound images and repeated VVI measurements several weeks later, during various hemodynamic states. To obtain interobserver variability, another observer who was blinded to VVI data repeated the VVI measurements.

117.2 Statistical Analysis

Data are expressed as mean \pm standard deviation. Differences in continuous variables between two groups were assessed by unpaired t-test, and comparison among multiple groups was performed by analysis of variance with ANOVA. Categorical variables were analyzed by Fisher's exact test. Intra- and inter-observer variability were reported as the correlation coefficient between measurements as well as the mean difference between respective measurements. All data analysis was performed by SPSS version 13.0 (SPSS Inc., Chicago, USA). A p value <0.05 was considered statistical significance.

117.3 Results

All patients successfully underwent Carotid artery ultrasound examinations and VVI-measurements could be determined in all investigated. The study group median age was 40.81 years, range 39–51 years and the control group median age was 40.17 years, range 38–53 years. The relationship between clinical variables and laboratory analysis in patients and controls are displayed in Table 117.1. Age, BP, HR, BMI and SV had no significant differences in the control group and the study group ($P > 0.05$). IMT of carotid had no significant difference between control group and study group ($P > 0.05$), TG, TC, LDL in the study group were higher than those in control group, HDL in the study group were lower than those in control group, but there were no significant difference between control group and study group ($P > 0.05$). There were significant differences in some segments in control group and study group, V_{max} of anterior wall, anterolateral wall and posterolateral wall were higher than those of posterior wall, posteromedial wall and anteromedial wall ($P < 0.05$). VTTP, V_{max} , S_{max} and SR_{max} of corresponding segments had significant differences in study group and control group ($P < 0.05$). There was no significant difference between any intra- or inter-observer measurements of VTTP, V_{max} , S_{max} and SR_{max} .

117.4 Discussion

The vascular tissue receives blood pressure and shear stress caused by blood flow during cardiac cycles. Especially, the effect of blood pressure in the radial direction is important in determining 2D tissue velocity. The displacement caused by blood pressure would be large in the soft material and the displacement would be small in the hard material. Whereas displacement and velocity characterize wall motion, strain and strain rate describe wall deformation. The term "strain", which in everyday language can mean "stretching", is used in echocardiography to

Table 117.1 Relationship between clinical and CCA variables in patients and controls

	Study group	Control group	P values
Clinical variables			
Age (yrs)	40.81(39–51)	40.17(38–52)	NS
Body mass index (kg/m ²)	21.4(19.45–23.58)	21.7(19.14–24.12)	NS
Blood pressure (mmHg)			
Systolic	111.4(99–125)	110.9(99–127)	NS
Diastolic	77.5(65–85)	76.8(66–84)	NS
HR (bpm)	71.2(55–81)	70.5(54–80)	NS
Triglycerides (mmol/L)	1.14(0.4–2.9)	1.34(0.5–2.3)	NS
Cholesterol (mmol/L)	5.78(4.2–6.6)	5.65(4.7–8.2)	NS
LDL (mmol/L)	3.89(2.8–5.3)	3.68(3.34–6.4)	NS
HDL (mmol/L)	1.34(1.1–1.8)	1.56(0.8–2.6)	NS
EF (%)	61.34(55–68)	63.47(56–70)	NS
SV (ml)	74.87(63.59–84.92)	75.81(64.12–86.17)	NS
Bifurcation IMT (mm)	0.95(0.78–1.07)	0.81(0.79–1.02)	NS
VTTP (x10 ² ms)			
Anterior wall	1.63(1.41–1.82)	2.21(1.95–2.51)	<0.05
Anterolateral wall	1.67(1.42–1.90)	2.19(1.96–2.47)	<0.05
Posterolateral wall	1.63(1.42–1.85)	2.21(1.97–2.55)	<0.05
Posterior wall	1.71(1.49–1.91)	2.17(1.94–2.45)	<0.05
Posteromedial wall	1.65(1.43–1.87)	2.12(1.96–2.41)	<0.05
Anteromedial wall	1.69(1.51–1.92)	2.16(1.93–2.47)	<0.05
Vmax			
Anterior wall	8.39(6.58–9.69)*	16.13(13.17–19.24)*	<0.05
Anterolateral wall	8.14(6.25–9.78)*	15.33(12.35–18.56)*	<0.05
Posterolateral wall	7.92(5.99–8.97)*	13.45(10.08–16.84)*	<0.05
Posterior wall	1.31(0.99–1.63)	2.36(2.11–2.67)	<0.05
Posteromedial wall	1.28(0.98–1.64)	1.98(1.68–2.29)	<0.05
Anteromedial wall	1.36(1.01–1.75)	2.24(1.98–2.64)	<0.05
Smax			
Anterior wall	5.44(3.16–7.75)	8.11(5.97–10.28)	<0.05
Anterolateral wall	6.23(3.74–8.68)	9.17(6.91–11.13)	<0.05
Posterolateral wall	5.98(3.49–8.47)	8.56(6.24–10.33)	<0.05
Posterior wall	5.36(3.01–7.18)	7.99(5.97–9.24)	<0.05
Posteromedial wall	5.47(1.66–7.68)	8.27(5.96–10.57)	<0.05
Anteromedial wall	5.67(3.56–7.79)	8.47(6.18–10.09)	<0.05
SRmax			
Anterior wall	0.31(0.13–0.49)	0.50(0.29–0.73)	<0.05
Anterolateral wall	0.29(0.11–0.47)	0.49(0.28–0.72)	<0.05
Posterolateral wall	0.33(0.17–0.52)	0.53(0.23–0.71)	<0.05
Posterior wall	0.29(0.15–0.38)	0.49(0.29–0.68)	<0.05
Posteromedial wall	0.31(0.14–0.49)	0.48(0.31–0.50)	<0.05
Anteromedial wall	0.28(0.12–0.45)	0.46(0.24–67)	<0.05

Values are presented as median and range unless otherwise stated. * presented as significant differences in some segments in the same groups. *LDL* low density lipoprotein cholesterol; *HDL* high density lipoprotein cholesterol; *IMT* intima-media thickness; *VTTP* Time to the peak of velocity; *Vmax* maximum of velocity; *Smax* maximum of strain; *SRmax* maximum of strain rate

describe “deformation” [11]. The concept of strain is complex. For a one dimensional (1D) object the only possible deformation is lengthening or shortening and the linear strain (amount of deformation) can be defined by the formula:

$$\varepsilon = \frac{L - L_0}{L_0} = \frac{\Delta L}{L_0} \quad (117.1)$$

where ε is strain, L_0 = baseline length and L = instantaneous length at the time of measurement.

Strain rate (SR) is the rate by which the deformation occurs (deformation or strain per time unit). The unit of strain rate is s^{-1} and the local rate of deformation or strain per time unit equals velocity difference per unit length:

$$\dot{\varepsilon} = \frac{\Delta \varepsilon}{\Delta t} = \frac{(\Delta L/L_0)}{\Delta t} = \frac{(\Delta L/\Delta t)}{L_0} = \frac{\Delta V}{L_0} \quad (117.2)$$

Where ΔV is the velocity gradient in the segment studied. Where $L(t)$ is the length at the time instance t and $L(t_0) \equiv L_0$.

During recent years, VVI, a novel echocardiographic imaging technique based on routine two-dimensional grayscale echocardiographic images that is independent of the angle of the transducer have emerged as valuable echocardiographic tools for more comprehensive and reliable assessment of myocardial function [12, 13]. In order to improve the tracking results, a new algorithm applies a carefully designed sequence of intermediate passages to accurately follow tissue motion through that combines ultrasound speckle tracking, the tissue-blood border detection, global motion coherence, and the consistency of motion between cardiac cycles, thus providing more information than qualitative data. All these passages are calculated with the aid of Fourier techniques that ensure a higher accuracy using the periodicity of the artery motion [13, 14]. In this way, the displacement information of the tracked points is obtained. The intima border is visually identified by the user and manually outlined. For feature-tracking technique, the pixel features like grey-scale are tracked based on cross-correlation algorithm which incorporates the image grey-scale normalization process [15]. Using such an algorithm, the level of grey-scale intensity will affect slightly the result of feature-tracking method as long as the ultrasonic imaging parameters are kept unchanged during the cineloop acquisition. Such kind of speckle tracking technique has proven accurate at a minimal temporal resolution of 30 Hz [16] which is far below the frame rate used in the current study. Similarly, the accuracy of VVI measurements has been validated by a sonomicrometry technique and a high reproducibility of VVI measurements has been reported by Pirate et al. [14] and confirmed by the present study. The method provides a measure of the vascular changes of type 2 diabetes mellitus and early detect vascular wall dysfunction through a new algorithm, combining tracking of ultrasound speckle, the vascular blood interface, and the vascular shape.

Type 2 DM is a disease with both metabolic and vascular components [17]. To our knowledge this is the largest study to date, in which velocity vector

imaging has been used to evaluate the biomechanical elasticity of large artery, and the change of common carotid elasticity in early stage of AS in patients with type 2 diabetes. By using this technique we report the quantified elasticity change of differences between the early stage of atherosclerotic plaque on the CCA in patients with type 2 diabetes mellitus (DM2) and health volunteers underwent ultrasonography. Our study show the two-dimensional tissue velocity and strain derived by the VVI technique provide important information on tissue character. Age, BP, HR, BMI and SV had no significant differences in the control group and the study group. Although in the early stage of type 2 diabetes mellitus (DM2), the metabolic disorders in patients predominantly included glucose metabolism abnormalities, and dyslipidemia, TG, TC, LDL in the study group were higher than those in control group, HDL in the study group were lower than those in control group, but there were no significant difference between control group and study group. Although IMT of carotid had no significant difference between control group and study group, not surprisingly, patients in the early stage of type 2 diabetes mellitus (DM2) display greater IMT than the healthy subjects because of their systemic atherosclerotic disease. Considering arterial biomechanics, wall shear stress and the circumferential stress acting on the arterial wall traditionally account for vessel wall remodeling during pathological circumstances. There were significant differences in some segments in control group and study group, V_{max} of anterior wall, anterolateral wall and posterolateral wall were higher than those of posterior wall, posteromedial wall and anteromedial wall. VTTP, V_{max} , S_{max} and SR_{max} of corresponding segments of patients in the early stage of type 2 diabetes mellitus (DM2) displayed significantly lower compared to healthy control group. Similarly, the elasticity change of common carotid artery of diabetogenous nephropathy has been reported by Zou et al. [18]. The comparison demonstrated potential biological significance of this measure in this very initial study. So patients with DM should always control their risk factors and recognize the signs and symptoms of potentially fatal complications as early as possible.

Several limitations of this study should be noted. First, two-dimensional speckle tracking algorithms are inherently dependent on image quality and endocardial border definition, any technical factors affecting gray scale may in theory have an impact on the accuracy of VVI measurements, so VVI may be considered semi quantitative at the present time. To reduce these errors to minimum, ultrasonic imaging parameters, such as gain, depth, focus zone, time gain compensation, and etc. were adjusted to the optimal levels initially and kept unchanged during the entire study. Second, this is only a 2-D ultrasonic imaging model, so the blood flow and the pressure drop were not considered; Third, the residual stress in the atherosclerotic vessels was not considered because present imaging techniques do not allow its quantification. However, VVI is a unique technique for measuring Time to the peak of velocity (VTTP), maximum of velocity (V_{max}), maximum of strain (S_{max}), and maximum of strain rate (SR_{max}) of the carotid arteries. Further incorporation of VVI in patients with type 2 diabetes may greatly enhance the capability of clinical intervention treatment in earlier period.

117.5 Conclusion:

Vector imaging can be used to evaluate the change of common carotid elasticity in early stage of AS in patients with type 2 diabetes. Although it is a relatively bigger scale study compared to the previous studies, it is clear that large scale longitudinal studies are necessary to determine the true diagnostic accuracy of these biomechanical characteristics for prediction of clinical risk of future cerebrovascular ischemic events.

References

1. Unwin N, Gan D, Whiting D (2010) The IDF Diabetes Atlas: providing evidence, raising awareness and promoting action. *Diabetes Res Clin Pract* 87:2–3
2. Graham I, Atar D, Borch-Johnsen K et al (2007) European guidelines on cardiovascular disease prevention in clinical practice: full text. Fourth Joint Task Force of the European Society of Cardiology and other societies on cardiovascular disease prevention in clinical practice (constituted by representatives of nine societies and by invited experts). *Eur J Cardiovasc Prev Rehabil* 14(Suppl 2):S1–113
3. Andersson DK, Svardsudd K (1995) Long-term glycemic control relates to mortality in type II diabetes. *Diabetes Care* 18(12):1534–1543. doi:10.2337/diacare.18.12.1534.
4. Burke AP, Farb A, Malcom GT et al (1997) Coronary risk factors and plaque morphology in men with coronary disease who died suddenly. *N Engl J Med* 336:1276–1282
5. Davies MJ (2000) The pathophysiology of acute coronary syndromes. *Heart* 83:361–366
6. Cho JJ, Chi YS, Yang W-I, Kim S-A, Chang H-J, Jang Y, Chung N, Ha J-W (2010) Assessment of mechanical properties of common carotid artery in Takayasu's arteritis using velocity vector imaging. *Circ J* 74:1465–1470
7. Valocik G, Druzbacka L, Valocikova I, Mitro P (2010) Velocity vector imaging to quantify left atrial function. *Int J Cardiovasc Imaging* 26:641–649
8. Lei Z, Yan L, Fei ZP, Xia ZY, Ping JX, Ting LX, Qiang CW, Xi LC, Cheng Z, Yun Z (2010) Peak radial and circumferential strain measured by velocity vector imaging is a novel index for detecting vulnerable plaques in a rabbit model of atherosclerosis. *Atherosclerosis* 211:146–152
9. Wendelhag I, Gustavsson T, Suurkula M, et al (1991) Ultrasound measurement of wall thickness in the carotid artery: fundamental principles and description of a computerized analysing system. *Clin Physiol (Oxford, England)* 11:565–77
10. Störk S, van den Beld AW, von Schacky C et al (2004) Carotid artery plaque burden, stiffness, and mortality risk in elderly men a prospective, population-based cohort study. *Circulation* 110:344–348
11. D'Hooge J, Heimdal A, Jamal F et al (2000) Regional strain and strain rate measurements by cardiac ultrasound: principles, implementation and limitations. *Eur J Echocardiogr* 1(3):154–170
12. Chen J, Cao T, Duan Y et al (2007) Velocity vector imaging in assessing myocardial systolic function of hypertensive patients with left ventricular hypertrophy. *Can J Cardiol* 23:957–961
13. Canesson M, Tanabe M, Suffoletto MS et al (2006) Velocity vector imaging to quantify ventricular dyssynchrony and predict response to cardiac resynchronization therapy. *Am J Cardiol* 98:949–953

14. Pirat B, Khoury DS, Hartley CJ et al (2008) A novel feature-tracking echocardiographic method for the quantitation of regional myocardial function: validation in an animal model of ischemia-reperfusion. *J Am Coll Cardiol* 51:651–659
15. Bohs LN, Trahey GE (1991) A novel method for angle independent ultrasonic imaging of blood flow and tissue motion. *IEEE Trans Biomed Eng* 38:280–286
16. Mailloux GE, Bleau A, Bertrand M, Petitclerc R (1987) Computer analysis of heart motion from two-dimensional echocardiograms. *IEEE Trans Biomed Eng* 34:356–364
17. Friedlander AH, Garrett NR, Norman DC (2002) The prevalence of calcified carotid artery atheromas on the panoramic radiographs of patients with type 2 diabetes mellitus. *J Am Dent Assoc* 133(11):1516–1523
18. Zou C, Huang P, Jiao Y, Jin H, Yang Y, Chen L (2008) Research on the common carotid artery wall of patients with type 2 diabetic by velocity vector imaging. *Chin J Ultrasonography* 17(12):1026–1029

Chapter 118

Rate Control for Multi-View Video Coding Based on Statistical Analysis and Frame Complexity Estimation

Qiaoyan Zheng, Mei Yu, Gangyi Jiang, Feng Shao and Zongju Peng

Abstract This paper presents a novel rate control algorithm for multi-view video coding. The proposed rate control algorithm is performed on three levels, the view level, group-of-pictures (GOP) level and frame level. In the view level, the target bit rates are allocated according to different types of views with a pre-statistical rate allocation proportion. In the GOP level, the total number of bits allocated to each GOP is computed and the initial quantization parameter (QP) of each GOP considering the QP values of B frames is set. In the frame level, the frame complexity is used to regulate the target bits for each frame. Experimental results show that the proposed rate control algorithm can accurately control the bit-rate to satisfy the requirements of multi-view video systems.

Keywords Three dimensional (3D) video · Multi-view video coding · Rate control · Statistical analysis · Frame complexity

118.1 Introduction

With the rapid development and extensive applications of video technique, three dimensional (3D) video contents has become the major driving force governing today's dynamics of the consumer electronic market [1]. Moreover, 3D video services being brought into mobile is also becoming a reality [2] along with the popularization of mobile phone supporting stereoscopic display. In multi-view

Q. Zheng (✉) · M. Yu · G. Jiang · F. Shao · Z. Peng
Faculty of Information Science and Engineering, Ningbo University, Ningbo, China

G. Jiang
National Key Lab of Software New Technology, Nanjing University, Nanjing, China

video coding (MVC) international standard, disparity-compensated prediction together with motion-compensated prediction are exploited to reduce all kinds of redundancy for coding efficiency. However, the bit allocation and rate control (RC) should be solved in orienting to efficient transmission and applications of MVC.

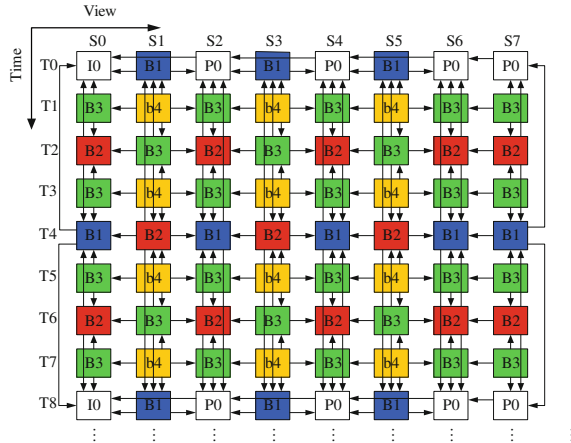
RC problem has been widely studied in video coding standards, such as MPEG-2 TM5 [3], H.263 TMN8 [4], MPEG-4 VM8 [5] and H.264 RC algorithm [6]. Nevertheless, most current RC algorithms concentrate on two dimensional (2D) video communications [7, 8], which are not fit well to MVC. Recently some projects were begun for RC in 3D video research areas [9]. In Yan et al. [10], a RC algorithm for MVC was proposed to reasonably allocate bit-rate among views based on correlation analysis. In [11], some characteristics of MVC was used to allocate proper bit rates for B frames and the quantization parameters (QP) of B frames in MVC were taken into account when deciding the initial QP of GOP. In [12], a joint buffer-related RC algorithm considering the special characteristics of multi-view hypothetical reference decoder for multi-view video plus depth (MVD) MVD is presented. However, the existing rate control algorithms cannot efficiently work for the hierarchical B pictures based MVC. One reason is that the rate allocation of B frames is not carefully considered in the rate control algorithms. Another reason is that in P-view and B-view of MVC, the key frames are not I frames, they are inter-view predicted B frames or P frames.

Based on the different types of views and frame complexity, we propose a novel frame level MVC rate control algorithm. The proposed RC algorithm is performed on three levels, namely view level bit allocation, GOP level bit allocation and frame level rate control. In the view level, the target rates are allocated according to different types of views with a pre-statistical rate allocation proportion. In the GOP level, the total number of bits allocated to each GOP is computed and the initial QP of each GOP considering the QP values of B frames is set. In the frame level, the target bits are allocated according to the frame complexity.

118.2 The Proposed Rate Control Algorithm for MVC

Multi-view video is a new technique in which an object or a scene is recorded by using several synchronous cameras from different positions, and MVC, as a compression technique, makes use of redundancies between adjacent views for high coding efficiency. To control bit rate for MVC and the corresponding bit-stream transmission, we propose a rate control algorithm with three layers: 1) view level rate allocation; 2) GOP level rate allocation; 3) frame level rate control. In the proposed algorithm, the first stage is off-line processed and the latter two stages are online processed.

Fig. 118.1 MVC coding structure with hierarchical B pictures



118.2.1 Target Bit Rate Allocation Scheme for View Level

The HBP coding structure is shown in Fig. 118.1, and the eight original views are encoded by MVC encoder. From the figure, it is seen that there is no inter-view prediction in view S0, thus it is called I-view here. P-view, such as the views S2, S4, S6 and S7, is encoded with unidirectional inter-view prediction from the reconstructed I-view and the previous reconstructed P-view. Likewise, the views S1, S3 and S5, known as B-view, are encoded with bidirectional inter-view prediction from the reconstructed I-view and P-view. Therefore, all the views are divided into three types of views, I-view, P-view and B-view, in the view level of the proposed algorithm.

In order to analyze the rate proportion among views in MVC method, the basis-QPs in the fixed QP coding are set to 22, 27, 32 and 37 for multi-view video. Figures 118.2 and 118.3 show the rate allocation proportion of each P-view and B-view with respect to Ballroom sequence, respectively. The proportion of single P-view is almost the same, from 0.12 to 0.14 under different fixed QP coding, and the average proportion is 0.13, as shown in Fig. 118.2. Likewise, the proportion of single B-view is about from 0.1 to 0.12 under different fixed QP coding, and the average proportion is 0.11, as shown in Fig. 118.3. It can be seen that, the rate allocation proportion for each view in different types of views, the P-view and the B-view, is almost equal.

Figure 118.4 shows the statistical average rate allocation proportion among the I-view, each B-view and each P-view, for Ballroom and Breakdancers. In Fig. 118.4, the proportion among three types of views is approximately equal to 0.155:0.133:0.105 for Ballroom and 0.160:0.126:0.112 for Breakdancers. According to the statistical observation, the rate allocation proportions among three types of views are approximately equal.

Finally, the statistical rate proportion among the I-view, all B-views and all P-views are shown in Fig. 118.5. It can be seen that, the proportions among three

Fig. 118.2 Rate allocation proportion of each view in P-view for Ballroom

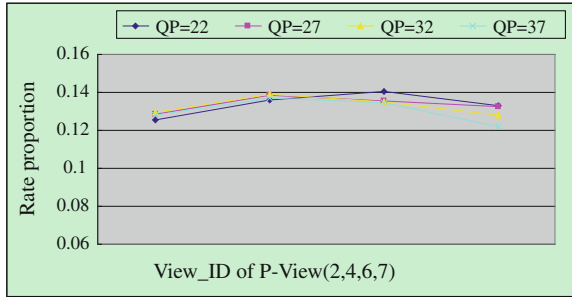


Fig. 118.3 Rate allocation proportion of each view in B-view for Ballroom

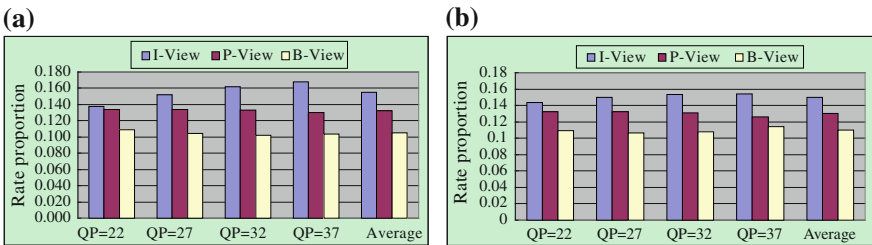
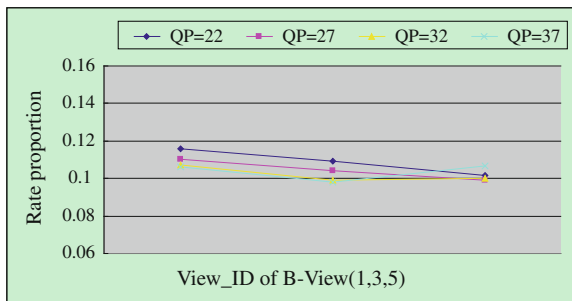
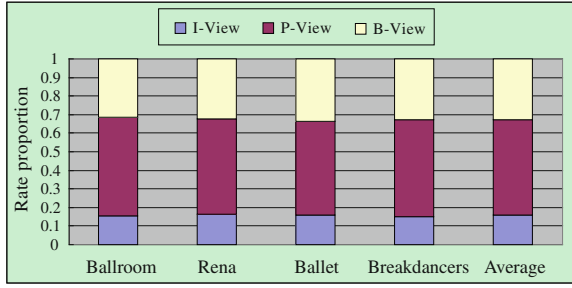


Fig. 118.4 Average rate allocation proportion among different types of views

types of views between different sequences have some slight difference, especially the proportion of I-view which is encoded with intra-view coding technique. The average rate proportion is approximately equal to 0.157: 0.517: 0.326 for the four sequences, Ballroom, Rena, Breakdancers, and Ballet. Under the total rate constraint, the view level rate allocation is needed to be offline performed before the actual encoding.

Fig. 118.5 Rate allocation proportion among the three types of views



118.2.2 Target Bit Rate Allocation Scheme for GOP Level

(1) *Total Number of Bits.*

In the GOP level, the total number of bits allocated to each GOP is computed and the initial QP of each GOP is set. At the beginning of encoding the i th GOP, the total number of bits allocated for the i th GOP is computed as follows:

$$T_r(n_{i,0}) = \frac{u(n_{i,1})}{F_r} * N_{gop} - \left(\frac{B_s}{8} - B_c(n_{i-1, N_{gop}}) \right), \quad (118.1)$$

where N_{gop} denotes the total number of frames in the current GOP, $u(n_{i,1})$ is the available channel bandwidth, F_r is the frame rate, and $B_c(n_{i-1, N_{gop}})$ is the actual buffer occupancy after coding the $(i-1)$ th GOP. The buffer occupancy should be kept at $B_s/8$ after coding each GOP. Since the channel bandwidth may vary at any time, in the case of CBR, T_r is updated frame by frame as follows:

$$T_r(n_{i,j}) = T_r(n_{i,j-1}) - A(n_{i,j-1}), \quad (118.2)$$

where $A(n_{i,j-1})$ denotes the actual encoded bit-rate for the $(i-1)$ th frame.

(2) *Starting Quantization Parameter of Each GOP.*

As shown in Fig. 118.1, MVC referencing structure has a large number of hierarchical B pictures. Hence, the QP values of B frames are taken into account when deciding the initial QP of GOP. The initial QP of each GOP is computed by

$$QP_{st} = \frac{Sum_{BQP}}{N_b} - 1 - \frac{8T_r(n_{i-1, N_{gop}})}{T_r(n_{i,0})} - \frac{N_{gop}}{15}, \quad (118.3)$$

where N_b is the total number of B frame in the GOP and Sum_{BQP} is the sum of quantization parameters for all B frames in the previous GOP. The I frame, the P frame in P-view and the first B frame in B-view in each GOP are coded using QP_{st} .

(3) *Rate Control Scheme for Frame Level.*

In terms of the frame level rate control of JVT-G012, the target bit rate for the j th frame is a weighted combination of $\hat{f}(n_{i,j})$ and $\tilde{f}(n_{i,j})$

$$f(n_{i,j}) = \beta * \hat{f}(n_{i,j}) + (1 - \beta) * \tilde{f}(n_{i,j}), \quad (118.4)$$

where β is a constant and its value is 0.5 when there is no B frame and is 0.9 otherwise in JVT-G012. In this paper, β is 0.5 because of the B frames are treated as P frames. $\hat{f}(n_{i,j})$ is determined according to the number of remaining bits, we propose an optimal bit allocation scheme based on frame complexity as follows

$$\hat{f}(n_{i,j}) = \frac{T_r(n_{i,j})}{N_{b,r}(j-1)} * \frac{XB(n_{i,j-1})}{AXB(n_{i,j-1})}, \quad (118.5)$$

where $N_{b,r}(j-1)$ is the number of the remaining B frames in the i th GOP, $XB(n_{i,j-1})$ is the complexity weight of the $(j-1)$ th B pictures, $AXB(n_{i,j-1})$ is the average complexity weight of the coded B pictures, and the frame complexity $XB(n_{i,j})$ is computed by

$$XB(n_{i,j}) = QP(n_{i,j}) * bits(n_{i,j}), \quad (118.6)$$

where $QP(n_{i,j})$ and $bits(n_{i,j})$ denote the QP and encoded bits, respectively.

Meanwhile, considering the frame complexity, the target buffer level $Tbl(n_{i,j})$ for the B frames is calculated by

$$Tbl(n_{i,j}) = Tbl(n_{i,j-1}) - \frac{Tbl(n_{i,2})}{N_b - 1} * \frac{XB(n_{i,j-1})}{AXB(n_{i,j-1})}, \quad (118.7)$$

Then, $\tilde{f}(n_{i,j})$ is determined based on the buffer level as follows

$$\tilde{f}(n_{i,j}) = \frac{u(n_{i,j})}{F_r} + \gamma(Tbl(n_{i,j}) - B_c(n_{i,j})), \quad (118.8)$$

where γ is a constant and its value is 0.75 when there is no B frame and is 0.25 otherwise. In this paper, γ is 0.75 because of the B frames are treated as P frames. $B_c(n_{i,j})$ denotes the current buffer fullness.

118.3 Experiment Results

In order to demonstrate the effectiveness of the proposed MVC rate control algorithm, many experiments are performed with multiview video sequences of Ballroom, Rena, Ballet, and Breakdancers. Ballroom (640×480) is provided by Interactive Visual Media group of MERL. Rena (640×480) is provided by Nagoya University. The sequence of Ballet and Breakdancers (1024×768) are captured by Microsoft Research. The test conditions of the four sequences are shown in Table 118.1.

In the experiments, we use the revised MVC software JMVC6.0 to implement the rate control algorithm. The eight views are view0, view1, view3, ..., view7 for

Table 118.1 Test conditions

Frame rate	15	Chanel type	CBR
GOP length	8	Search mode	Fast search
Search range	32	Frame's no.	81

Table 118.2 Simulation results of MVC RC

Sequences	Target bit rate (kbps)	Actual bit rate (kbps)	RCE (%)
Ballroom	8,852.496	8,839.686	-0.14
	4,334.225	4,327.584	-0.15
	2,243.433	2,246.327	0.13
	1,248.575	1,249.886	0.11
Rena	4,012.802	4,001.086	-0.29
	1,917.638	1,909.649	-0.42
	969.455	967.689	-0.18
	547.194	546.862	-0.06
Breakdancers	15,233.820	15,236.180	0.02
	5,427.114	5,428.452	0.02
	2,580.990	2,581.440	0.02
	1,463.913	1,464.4815	0.04
Ballet	6,346.296	6,349.655	0.05
	2,848.064	2,846.938	-0.04
	1,577.990	1,579.274	0.08
	957.215	958.9229	0.18

Ballroom, Ballet and Breakdancers, and view38, view39, view40, ..., view45 for Rena. JMVC with the proposed rate control and without rate control are tested, respectively.

We first confirm the accuracy of the proposed multi-view video rate control algorithm. Table 118.2 summarizes the matching accuracy between the controlled bit-rate and the target ones. In Table 118.2, the target bit-rate and the actually controlled bit-rate are with respect to all of the eight views. The rate control error (RCE) is used to measure the accuracy of the bitrate estimation

$$RCE = \frac{|R_{target} - R_{actual}|}{R_{target}} \times 100\%, \quad (118.9)$$

where R_{target} and R_{actual} denote the target bit-rate and the actual coding bit-rate, respectively.

Table 118.2 indicates that the absolute inaccuracy of the proposed multi-view video rate control algorithm is within 0.42%. Through the simple pre-statistical rate allocation proportion method, the rate allocation can approximately adapt to the different types of views and the proposed method can provide a certain degree of rate control accuracy for MVC.

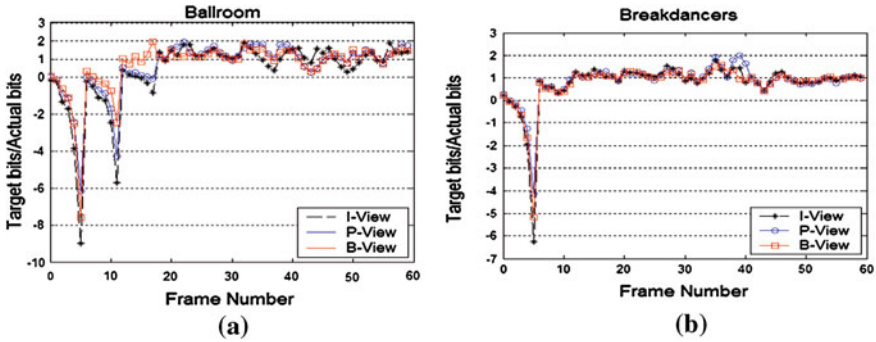


Fig. 118.6 The ratio between target bits and actual bits in three types of views for the proposed RC

Besides the whole rate controlling accuracy, the fluctuations of rate controlling accuracy between the consecutive frames are also taken into account. Fig. 118.6 shows the ratio between target bits and actual bits in three types of views, where the I-view is S0, the P-view is S2 and the B-view is S1. In Fig. 118.6, the I frame and the first B frame do not carry out rate controlling in each GOP, so the number of rate control frame is 60. The value of ratio between target bits and actual bits is closer to 1, the rate controlling accuracy is higher.

Figure 118.6 shows that the target bits are inadequate and the actual coding bits of the encoding frame are possibly higher than the target allocated bits in short frames because of the rate-quantization model is not stable enough in short time. However, this phenomenon will disappear after about two GOPs are encoded, and the value of the ratio between the target bits and actual bits is fluctuated around 1 for the subsequent frames. The proposed rate control algorithm allocates the target bits for one encoding frame with considering the current buffer status. It can be seen that the rate-quantization model is gradually updated and the proposed rate control can maintain suitable rate controlling accuracy in a short time even if the frame is easy or complex to be coded.

118.4 Conclusion and Future Work

This paper presents a novel rate control technique for multi-view video coding, which is performed on three levels, namely view level, GOP level and frame level. At the view level, the rates are discriminatorily allocated according to the special characteristics of MVC. At the GOP level, the bits and the initial QP considering the QP values of B frames of each GOP are computed. In the frame level, the target bits are allocated according to the frame complexity. Experimental results show that the proposed rate control technique can accurately control the bit-rate.

In the proposed RC method, rate allocation proportion is off-line processed in view level, so we will further study the bit allocation relation among the views in order to achieve online processing. In addition, We could investigate just noticeable distortion (JND) as a perceptual criterion in rate control process, so as to keep low bits consumption in each scalable level while maintain optimal encoding results.

Acknowledgments This work is supported by the Excellent Dissertation Foundation of Graduate School of Ningbo University (No. PY20100011), Natural Science Foundation of China (60872094, 60832003, 61071120), the Projects of Chinese Ministry of Education (200816460003, 20093305120002), and Scientific Research Fund of Zhejiang Provincial Education Depart. (Z200909361).

References

1. Merkle P, Müller K, Wiegand T (2010) 3D video: acquisition, coding, and display. *IEEE Trans Consumer Electron* 56(2):946–950
2. Gotchev A, Smolic A et al (2009) Mobile 3D television: development of core technological elements and user-centered evaluation methods toward an optimized system. *Electron Imaging Symp*, San Jose Jan 2009
3. ISO/IEC JTC1/SC29/WG11, MPEG 2 test model 5, Rev. 2, Apr. 1993
4. ITU-T SG16, Video codec test model, near-term, version 8 (TMN8), Portland, June 1997
5. ISO/IEC JTC1/SC29/WG11, Text of ISO/IEC 14496-2 MPEG-4 Video VM-Version 8.0, Doc. W1796, Stockholm, Sweden, July 1997
6. ISO/IEC JTC1/SC29/WG11 and ITU-T SG16/Q.6, Adaptive basic unit layer rate control for JVT, Doc. JVT-G012r1, Pattaya, Thailand, March 2003
7. Xie Z, Bao Z, Xu C, Zhang G (2010) Optimal bit allocation and efficient rate control for H.264/AVC based on general rate-distortion model and enhanced coding complexity measure. *IET Image Process* 4(3):172–183
8. Li JJ, Abdel-Raheem E (2010) Efficient rate control for H.264/AVC intra frame. *IEEE Trans Consum Electron* 56(2):1043–1048
9. Li ZC, Qin SY, Itti L (2011) Visual attention guided bit allocation in video compression. *Image Vis Comput* 29(1):1–14
10. Yan T, An P, Shen L, Zhang Z (2010) Rate control algorithm based on frame complexity estimation for MVC. In: *Proceedings of SPIE: visual communications and image processing*, vol 7741, Huang Shan, Aug 2010
11. Seanae P, Donggyu S (2009) An efficient rate-control algorithm for multi-view video coding. *Digest of Technical Papers—IEEE international conference on consumer electronics*, pp 115–118
12. Liu YW, Huang QM, Ma SW, Zhao DB, Gao W (2011) A novel rate control technique for multi-view video plus depth based 3D video coding (in press)

Chapter 119

Time-Consistent Preprocessing of Depth Maps for Depth Coding in 3DTV

Fengfei Sun, Mei Yu, Gangyi Jiang, Feng Shao,
Zongju Peng and Fucui Li

Abstract Depth-image-based rendering (DIBR) is widely used for view synthesis in three dimensional video applications. However, the quality of the associated depth map is an issue, as the depth channel is usually a result of estimation procedure based on stereo correspondences or comes from a noisy and low-resolution range sensor. Therefore, proper preprocessing of the depth maps is needed before it is used for compression and/or view rendering. In this paper, we propose a novel preprocessing algorithm to refine time-inconsistent depth maps. Experimental results show that the proposed algorithm can improve time-consistency of depth maps and save 22–50% encoding bit-rate, while the proposed algorithm improves the quality of rendered virtual views.

Keywords Three dimensional television system · Depth map · Time-consistency · Depth-image-based rendering

119.1 Introduction

In the recent years, there has been a growing interest in three dimensional television system (3DTV) system, which can enables depth perception for the viewer [1]. Depth-image-based rendering (DIBR) is a key technology in advanced 3DTV

F. Sun · M. Yu (✉) · G. Jiang · F. Shao · Z. Peng · F. Li
Faculty of Information Science and Engineering, Ningbo University,
Ningbo 315211, China
e-mail: yumei2@126.com

M. Yu · G. Jiang
National Key Lab of Software New Technology, Nanjing University,
Nanjing 210093, China

System [2, 3]. As the high computational complexity required for depth estimation [4] is typically not acceptable at the decoder, a depth map should be transmitted. In so-called video plus depth applications, only one view (also named as texture video) with its associated depth map is required for view synthesis.

Depth maps are usually obtained through ‘depth-from-stereo’ or ‘depth-from-multi-view’ type of algorithms based on finding disparities between camera images and converting these to corresponding depth [5]. The quality of the delivered depths also varies from approach to approach. Depth from stereo might suffer from inaccuracies in finding the disparity correspondences, especially when real-time performance is targeted. To utilize limited transmission bandwidth efficiently, depth maps usually undergo compression operations and might suffer from blocky artifacts while compressed by contemporary algorithms such as H.264 [6, 7]. Filtering of depth maps has been addressed mainly from the point of view of increasing the resolution [8–11]. In [9], a joint bilateral filtering has been suggested to upsample low-resolution depth maps. The approach has been further refined in [13] by suggesting proper anti-aliasing and complexity-efficient filters. In [11], a probabilistic frame has been suggested. For each pixel of the targeted high-resolution grid, several depth hypotheses are built and the hypothesis with lowest cost is selected as a refined depth value. The procedure is run iteratively and bilateral filtering is employed at each iteration to refine the cost function used for comparing the depth hypotheses. When accompanying video sequences, the consistency of successive depth maps in the sequence becomes an issue. Time-inconsistent depth sequences may waste extra bandwidth and reduce the encoding efficiency.

In this chapter, the time-inconsistency of the depth maps on the background area is analyzed and compared with color sequences. Then, Time-Consistent Preprocessing algorithm based on color image encoding is proposed.

119.2 Time-Consistent Preprocessing of Depth Map

119.2.1 Analysis on time-consistency of depth maps

Depth maps are usually obtained through ‘depth-from-stereo’ or ‘depth-from-multi-view’ type of algorithms based on finding disparities between camera images and converting these to corresponding depth. The quality of the delivered depths also varies from approach to approach. The depth maps background area may be time-inconsistent. For example, the background area of the same color sequences is invariable, so the depth of background area should not change. Figure 119.1a and b are frames 0 and 1 from ‘Pantomime’ color sequence. Figure 119.1d and e are frames 0 and 1 from ‘Pantomime’ depth sequence. Figure 119.1c is subtraction result for Fig. 119.1a and b. Figure 119.1c is

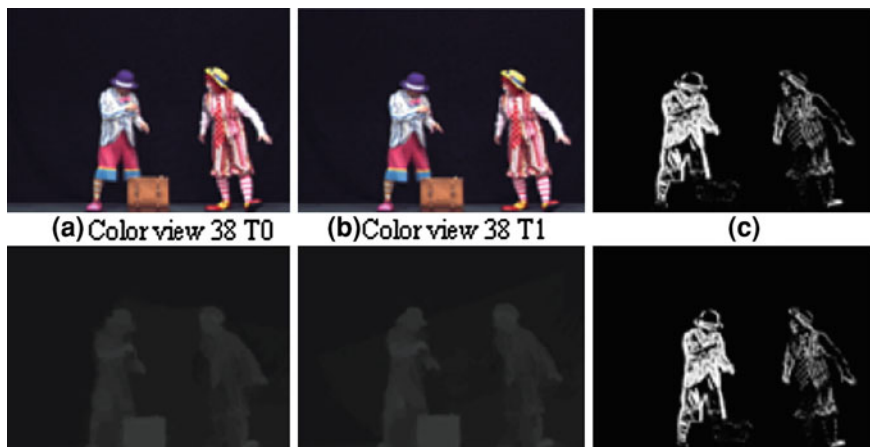


Fig. 119.1 Color and depth frame difference map of Pantomime

subtraction result for Fig. 119.1d and e. Obviously, the color video sequences compared to the depth maps have better time-consistency.

We analyze time-consistency of the color sequences and its corresponding depth maps. Test sequences are ‘Door Flowers’ and ‘Pantomime’. T_C denotes the time-consistency and it is defined by

$$T_C = \frac{N}{M} \quad (119.1)$$

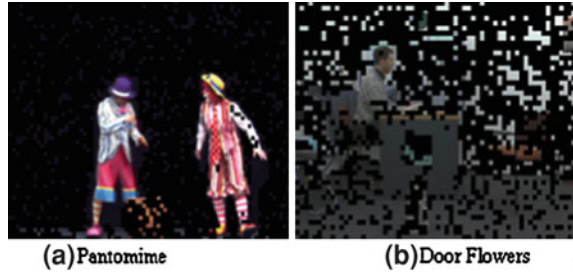
where N represents number of pixels that are zero for binary image. M is number of image pixels. By (119.1), Door Flowers time-consistency of color and depth, respectively, are 0.9054 and 0.6042. Pantomime time-consistency of color and depth are 0.9438 and 0.7140. As can be seen by the above, the time-consistency of depth sequences is relatively poor.

Thus, the depth sequences, relative to the color sequences, have its characteristics. The time-consistency of the depth maps is poor compared with its corresponding color video. It may waste extra bandwidth and reduce the encoding efficiency. Then, time-consistency preprocessing of depth sequences is necessary before depth maps encoding.

119.2.2 Analysis on SKIP Mode Distribution in Color Video Coding

Here, Door Flowers and Pantomime color test sequences are encoded based on H.264/AVC. Then, analyze the distribution of the SKIP mode [12]. Figure 119.2 shows distribution model of SKIP mode and motion vector (0, 0). Black areas in

Fig. 119.2 Distribution of SKIP (0, 0) macro-block.
a Pantomime. **b** Door flowers



this figure represent the current macro-block that encoding mode that is SKIP and motion vector (0, 0), expressed as SKIP (0, 0). As shown in Fig. 119.2, the black areas are basically in the background area. These areas should be time-consistent. By Fig. 119.2a can be seen, some macro-blocks are in motion object, such as the right arms of crown, we can eliminate these macro block that can not meet the requirements through setting threshold. Therefore, we can make use of these areas in accordance with the depth of color image coding sequence model between current frame and reference frame, to enhance the smooth on the depth of the consistency of the temporal sequence.

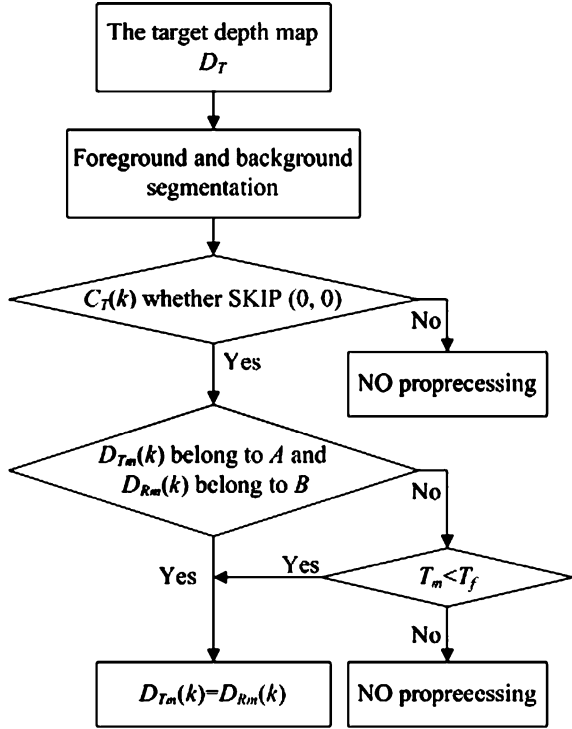
119.2.3 *The proposed time-consistent preprocessing algorithm of depth maps*

SKIP (0, 0) block basically occurs in background region or static areas, and these areas in time should have consistency with reference to target depth image. There we can use the color image encoding information to preprocessing depth map to strengthen the time-consistency of the target and the reference depth map.

This chapter firstly encodes color images and gets encoding information. Then, the foreground area and the background region of the target and the reference depth map are calculated. Next, judge the current macro-block of the target depth map corresponding color image macro block encoding mode, if it is SKIP (0, 0) and the current macro-block of the target and the reference depth map are background area, depth value of the reference depth map assign to the target depth map. Otherwise, judgment target depth and reference depth difference value, if meet the threshold, then the depth value of the reference depth assign to target depth. Figure 119.3 shows the proposed depth map preprocessing algorithm diagram.

In the following diagram, D_T represents target depth map. C_T represents target color image. D_{Tm} and C_{Tm} are macro-block with respect to the target depth map and the target color depth, respectively. D_{Rm} is macro-block of the reference depth map. Figure 119.3 shows the result of foreground and background segmentation. In Fig. 119.4, the black area is background object. A and B are background object with respect to the target depth map and the reference depth map, respectively.

Fig. 119.3 The diagram of the proposed time-consistent preprocessing algorithm of depth maps



T_f and T_m are thresholds in order to determine $D_{Tm}(k)$ and $D_{Rm}(k)$ whether the same scene and They are defined by

$$T_f = \sum_{y=0}^{H-1} \sum_{x=0}^{W-1} \frac{|D_T(x, y) - D_R(x, y)|}{N} \tag{119.2}$$

$$T_m = \sum_{j=0}^{K-1} \sum_{i=0}^{K-1} \frac{|D_{Tm}(i, j) - D_{Rm}(i, j)|}{K^2} \tag{119.3}$$

Where $D_T(x, y)$ and $D_R(x, y)$ are depth value with respect to the target depth map and the reference depth map, respectively. H is the picture height, W is the picture width. N represents number of pixels. $D_{Tm}(x, y)$ and $D_{Rm}(x, y)$ are depth value with respect to D_{Tm} and D_{Rm} , respectively. The value of K is 16.

119.3 Experimental Results and Analyses

In 3D video and Free viewpoint video (FTV), the depth map is mainly used to render the virtual view images. Therefore, we must ensure subjective and objective quality of the virtual view, after preprocessing depth map. The proposed algorithm

Fig. 119.4 Foreground and background segmentation. **a** Pantomime. **b** Door flowers

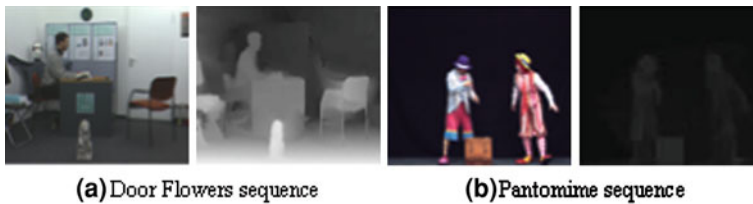
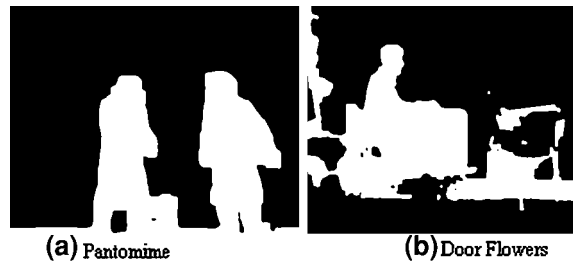


Fig. 119.5 Test sequence of the proposed algorithm. **a** Door flowers sequence. **b** Pantomime sequence

is tested on ‘Door Flowers’ and ‘Pantomime’ sequences, as shown in Fig. 119.5. For ‘Door Flowers’ sequence, view 7 and 10 were selected as reference views and view 8 was set as the virtual view to be synthesized. For ‘Pantomime’ sequence, view 38 and 41 were selected as reference views and view 39 was set as the virtual view to be synthesized. Depth maps were independently encoded using JM14.2 with QPs 18, 22, 26 and 30 with the IBP coding structure.

With the Eq.(119.1), time-consistency of Door Flowers and Pantomime depth maps based on the proposed algorithm are 0.7846 and 0.8856. The proposed algorithm significantly improves the time-consistency of the background region compared with its values 0.6042 and 0.7104 of original Door Flowers and Pantomime depth maps. Figure 119.6 shows an example of time-consistent preprocessing of depth maps based on the proposed algorithm. It improves the time-consistency of the target and the reference depth map.

Tables 119.1 and 119.2 show the results of depth encoding bitrate and Rendered image quality. Figures 119.7 and 119.8 show the corresponding rate-distortion (RD) curves for ‘Door Flowers’ and ‘Pantomime’, respectively. Experimental results show that the proposed algorithm can improve time-consistency of depth maps. From Tables 119.1 and 119.2, it is clear that the proposed algorithm can reduce the encoding bit-rate ranging from 22% to 50%. It can be confirmed that the proposed time-Consistent Preprocessing algorithm not only improves the rate-distortion performance of the depth coding as shown in Figs. 119.7a and 119.8a but also significantly reduces the required depth bit rate for a given rendering quality level as shown in Figs. 119.7b and 119.8b.

Fig. 119.6 Time-consistent preprocessing of depth maps. **a** Door flowers. **b** Pantomime

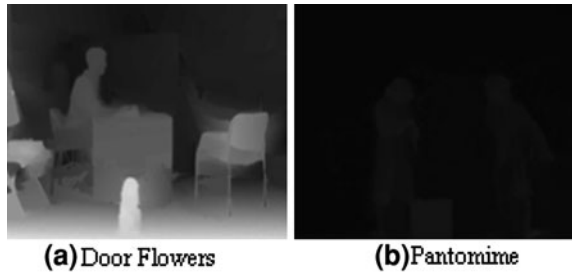


Table 119.1 Experimental results for ‘Door Flowers’

QP	Depth encoding bitrate (kbit/s)		Saved(%)	Rendered image quality (dB)	
	Original	Proposed		Original	Proposed
QP18	7393.63	3776.50	-48.92	26.50	26.53
QP22	4453.82	2355.22	-47.11	26.46	26.49
QP26	2419.22	1339.44	-44.64	26.42	26.45
QP30	1287.65	750.82	-41.72	26.38	26.42

Table 119.2 Experimental results for ‘Pantomime’

QP	Depth encoding bitrate (kbit/s)		Saved(%)	Rendered image quality (dB)	
	Original	Proposed		Original	Proposed
QP18	2788.99	2105.66	-24.49	31.21	31.24
QP22	1387.27	1116.24	-19.53	31.12	31.14
QP26	709.15	567.70	-20.02	31.02	31.05
QP30	415.85	340.73	-18.07	31.02	31.03

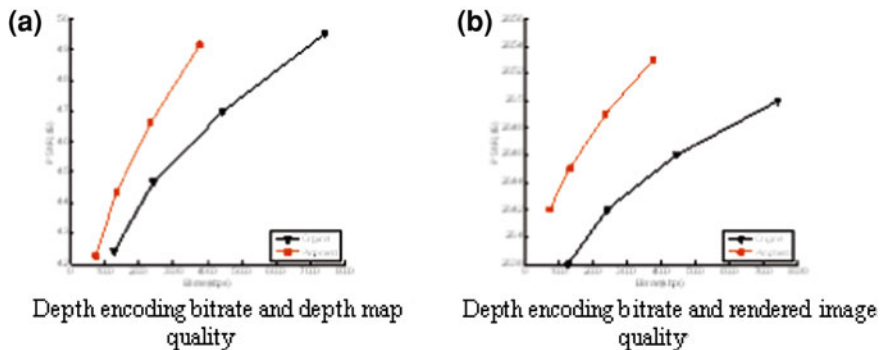


Fig. 119.7 Rate-distortion curves for ‘Door Flowers’. **a** Depth encoding bitrate and depth map quality. **b** Depth encoding bitrate and rendered image quality

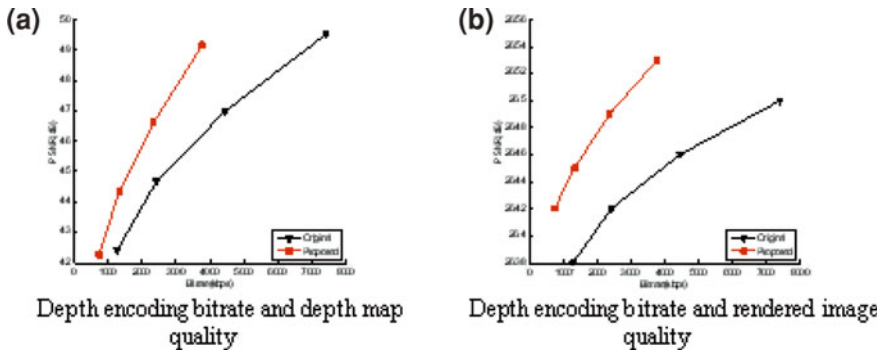


Fig. 119.8 Rate-distortion curves for 'Pantomime'. **a** Depth encoding bitrate and depth map quality. **b** Depth encoding bitrate and rendered image quality

119.4 Conclusion

A depth map preprocessing algorithm is proposed in order to improve time-consistency and high encoding efficiency of depth map. Through statistical results of SKIP (0, 0) macro-block mode in color video coding, the background region extraction and threshold setting, depth maps are preprocessed well to obtain better time-consistency. Experimental results show that the proposed algorithm can achieve 20–50% bitrate saving of encoding depth maps in comparison with JM14.2 because of improvement of depth's time-consistency. It significantly improves the depth encoding performance as well as the quality of rendered views compared with the conventional scheme.

The proposed algorithm based on color image encoding can apply to obtain the depth map, because the depth of depth sequence background area will not change. At the same, the encoding mode of the processed depth map may be close to that of color image. Then, color video coding information can be used to predict the encoding mode of the depth maps in order to speed up the process of depth map encoding.

Acknowledgments This work was supported by Natural Science Foundation of China (60832003, 60872094, 61071120), the projects of Chinese Ministry of Education (200816460003, 20093305120002).

References

1. Xue H, Xin J, Satoshi G (2011) A novel depth-image based view synthesis scheme for multiview and 3DTV. *Comput Sci* 6523:161–170
2. Redert A, Op de Beeck M, Fehn C, IJsselsteijn W, Pollefeys M, Van Gool L, Ofek E, Sexton I, Surman P, (2002) ATTEST advanced three-dimensional television system techniques. In: *Proceedings of 3DPVT' 02*. Padova, Italy, pp 313–319 June 2002

3. Fehn C, Hopf K, Quante Q (2004) Key technologies for an advanced 3D-TV system. In: Proceedings of SPIE three-dimensional TV, video and display III. Philadelphia, PA, pp 66–80 Oct 2004
4. Description of exploration experiments in 3D video coding (2008) ISO/IEC JTC1/SC29/WG11, Doc. W9991, Hannover, Germany
5. Scharstein D, Szeliski R (2002) A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int J Compu Vis* 47:7–42
6. Merkle P, Morvan Y, Smolic A, Farin D, Mueller K, de With PHN, Wiegand T (2009) The effects of multiview depth video compression on multiview rendering. *Sig Process Image Commun* 24(1–2):73–88 Jan 2009
7. S. Sergey, G. Atanas, E. Karen (2010) Multistep joint bilateral depth upsampling. In: Astola JT, Egiazarian KO (eds) *Image processing: algorithms and systems VIII*. Proceedings of the SPIE, vol 7532, pp 75333217–753217-12
8. Qingxiong Y, Ruigang Y, Davis J, Nister D (2007) Spatial-depth super resolution for range images, in *CVPR'07*
9. Kopf J, Cohen M, Lischiski D, Uyttendaele M (2007) Joint Bilateral Upsampling. In: Proceedings of SIGGRAPH conference, *ACM Transactions on Graphics*, vol 26(3):94–1–94-5
10. Riemens AK, Gangwal OP, Barenbrug B, Berretty R-PM (2009) Multistep joint bilateral depth upsampling. In: Majid Rabbani, Robert LS (eds) *Proceedings of SPIE, visual communications and image processing*, vol 7257. pp 72570 M
11. Smirnov S, Gotchev A, Egiazarian K (2009) Method for restorations of compressed depth maps: a comparative study, in *VPQM2009*
12. Merkle P, Wang Y, Muller K, Smolic A, Wiegand T (2009) Video plus depth compression for mobile 3D services. In: 3DTV conference. Potsdam, German

Chapter 120

A Mixture of Gaussian-Based Method for Detecting Foreground Object in Video Surveillance

Xun Wang, Jie Sun and Hao-Yu Peng

Abstract This chapter presents a novel MoG based method for foreground detection and segmentation in video surveillance. Normal MoG is difficult to deal with the foreground objects stay in the scene for a long time and segment different foreground objects from one blob. We improve MoG by adopting posterior feedback information of object tracking based on statistics to robustly modeling the background and to perfect the foreground segmentation result. Experiments and comparisons show that our method is robust and accurate for detecting foreground object in video surveillance.

Keywords Video surveillance · Mixture of Gaussian · Posterior information · Statistics · Tracking

120.1 Introduction

In automated surveillance systems, foreground object detection and segmentation from a video stream is crucial task, it exerts a great influence in follow-up operation, such as object tracking, object classification and etc. Considering the

X. Wang · J. Sun · H.-Y. Peng (✉)
College of Computer Science and Information Engineering, Zhejiang Gongshang
University, Hangzhou 310018, China
e-mail: hypeng@mail.zjgsu.edu.cn

X. Wang
e-mail: wx@mail.zjgsu.edu.cn

J. Sun
e-mail: sj.cadcg@gmail.com

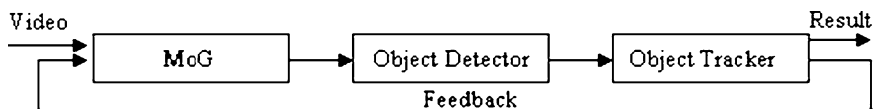


Fig. 120.1 The block diagram of the algorithm

character of video surveillance, e.g., an object comes into the scene, and stays in the scene for a while, the system should be able to recognize the object as foreground instead of classifying it to background rapidly; another problem is that two foreground objects come into the scene, then combine into one blob at a moment due to shelter or shadow, it is necessary to segment the blob to separate foregoing foreground objects instead of a new object. Our goal is to create a robust, adaptive system that can handle the situations related above.

In this chapter, we propose effective algorithms to deal with the prior-mentioned problems. Mixture of Gaussian algorithm is used to modeling background firstly. After the background is established, the algorithm begins to detect foreground objects and its property, e.g., its position, motion direction, speed, area, contour, and color histogram. Those information are used to tracking objects in later frames. According to the posterior information gained by tracking, the MoG background modeling algorithm can perfect the background to detect accurate and integrated foreground objects. The block diagram of the proposed algorithm is shown in Fig. 120.1. The structure of this chapter is as follows: Sect. 120.2 describes related work. Section 120.3 describes the Gaussian background model algorithm [1]. Section 120.4 describes the statistics-based tracking algorithm. Finally, Sect. 120.5 shows the experimental result and Sect. 120.6 concludes the paper.

120.2 Related Work

The simplest and straightforward approach to segment foreground from background is background subtraction, or frame differencing, large changes are regarded as foreground. Based on frame differencing, running average uses a parameter α to maintain the background frame, it is more adaptive than the previous method. The two methods are not robust enough to detect foreground as they are sensitive to subtle environment changes.

The more advanced approach is to build a representation of the background to compare against new frames. Those approaches can be further classified into two classes: parametric methods and nonparametric methods. For parametric methods, Mixture of Gaussian (MoG) [1] is the most common and significant one, it employs statistics by using a Gaussian model for each pixel, every Gaussian model maintain the mean and the variance of each pixel, the assumption then is that the pixel value follows a Gaussian distribution. In their work, they use an on-line approximation to

update the model. Considering the swaying tree and flickering monitor, MoG can handle those multi-model changes well. Many improvements on Gaussian mixture model are presented in [2, 3]. Toyama et al. presented a linear Wiener filter to learn and predict color changes, the linear predictor coefficients are updated each frame. However, this method is still hard to handle moving background objects and shadows. Hidden Markov models (HMMs) have been used in [4, 5]. In [4], a three-state HMM has been used in traffic monitoring, the three states correspond to the background, shadow, and foreground. In [6], edge features are used for modeling background to make the background invariant to illumination changes.

Compared with parametric methods, Elgammal et al. [7] introduce non-parametric model of the background, the advantage is that it can achieve greater accuracy under similar computational cost, another advantage of this approach is the incorporation of spatial constraints into the formulation of foreground classification. The model is robust and can handle the situations where the background of the scene is cluttered and not static but has subtle motions such as swaying trees. Many other nonparametric background model algorithm are presented in [8, 9].

Another excellent approach is taken by Oliver et al. [10], a Bayes decision rule for classification of background and foreground from selected feature vectors is formulated. This method uses posterior probability to identify foreground, the result is better than others in integrality.

In [11], the author analyzes the temporal variation of intensity or color distributions, instead of either calculating at temporal variation of pixel statistics, or the spatial variation of region statistics. This enables slower background updates, and minimizes the probability that the foreground object be incorporated into the background.

The nonparametric methods have the superiority that they do not need to assume the pixel matching a particular model and estimate its parameters. Therefore, this method can adapt to arbitrary unknown data distribution. But nonparametric methods are not as efficient as the parametric methods, it is computationally relatively.

The algorithm presented in this paper falls into parametric category, we did many work based on [1] in order to improve the foreground integrality and accuracy.

120.3 Mixture of Gaussian Background Modeling Algorithm

In this section, we will simply introduce the Mixture of Gaussian algorithm [1].

Each pixel in the image is modeled by a mixture of K Gaussian distributions. The probability of observing pixel value is

$$P(X_t) = \sum_{i=1}^K w_{i,t} * \eta \left(X_t, \mu_{i,t} \sum_{i,t} \right) \quad (120.1)$$

Where $w_{i,t}$ is the estimate of the weight of the i th Gaussian at time t , $\mu_{i,t}$ is the mean value of the i th Gaussian at time t , $\sum_{i,t}$ is the covariance matrix of the i th Gaussian probability density function.

$$\eta\left(X_t, \mu, \sum\right) = \frac{1}{(2\pi)^{\frac{n}{2}} |\sum|^{\frac{1}{2}}} e^{-\frac{1}{2}(X_t - \mu)^T \sum^{-1} (X_t - \mu)} \quad (120.2)$$

The covariance matrix is assumed to be of the form:

$$\sum_{k,t} = \sigma_k^2 I \quad (120.3)$$

The K distributions are ordered based on the fitness value w_k/σ_k .

Every new pixel value, X_t , is checked against the existing K Gaussian distributions, until a match is found. A match is defined as a pixel value within 2.5 standard deviation of a distribution.

The Gaussian model will be updated by the following update equations,

$$w_{k,t} = (1 - \alpha)w_{k,t-1} + \alpha(M_{k,t}) \quad (120.4)$$

$$\mu_t = (1 - \rho)\mu_{t-1} + \rho X_t \quad (120.5)$$

$$\sigma_t^2 = (1 - \rho)\sigma_{t-1}^2 + \rho(X_t - \mu_t)^T (X_t - \mu_t) \quad (120.6)$$

Where

$$\rho = \alpha \eta(X_t | \mu_k, \sigma_k) \quad (120.7)$$

$$M_{k,t} = \begin{cases} 1; & \text{if the model matched} \\ 0; & \text{if the model not matched} \end{cases} \quad (120.8)$$

If the pixel value does not match each of the K Gaussian models, the least probable model will be replaced by a distribution with the current value as its mean, an initially high variance, and a low weight parameter.

Then the first B distributions are chosen as the background models, where B is estimated as

$$B = \arg \min_b \left(\sum_{k=1}^b w_k > T \right) \quad (120.9)$$

The T is the minimum fraction of the background model, it is set to 0.7 experimentally.

MoG is based on statistical modeling of pixel intensity, while this method has a drawback that the foreground object always has some holes, e.g., the object is not integrated. Another drawback is that if a running car stopped in video stream, the car may be turn to be background soon according to the update strategy presented by [1].

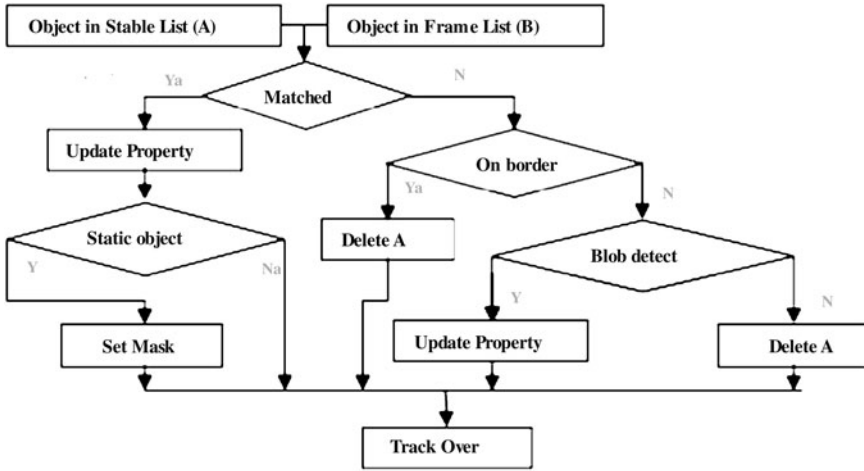


Fig. 120.2 Tracking algorithm diagram

Considering the efficiency in video surveillance, we select $K = 3$ in MoG, that means we build only three models for each pixel, and set History-Window to 1000 to slow the update speed.

120.4 Statistics-Based Tracking Algorithm

120.4.1 Object Detector

After the background is established, we compute the property of the object extracted by using statistical method. In this chapter, position, contour and its area, motion direction, motion speed, color histogram of the object is computed.

120.4.2 Object Tracker

The track diagram is shown in Fig. 120.2 as below. Where the stable list maintains the stable objects and the frame list maintains the new frame objects. In matching step, considering the displacement of the object in adjacent two frames is subtle, it is necessary to reduce the search scope to a small area around the object. In this area, we compare the object’s motion direction, area, speed, and color histogram features to find the best match. The tracking algorithm describes as follow.

For each object in the stable list do:

For each object in the frame list do:

- Compare the position of the two object, if it is too far, skip to the next object in frame list;
- Compare the motion direction of the two object, if the dot product of the two objects is negative, skip to the next object in frame list;
- Compare the speed of the two object, if the result exceed a threshold, skip to the next object in frame list;
- Compare the area of the two object, if the result exceed a threshold, skip to the next object in frame list;
- Compare the color histogram among the remaining objects in frame list, find the best match with the object in stable list.

If a best match is found:

- Update the object's property in stable list, e.g., the position, the direction, the area, the speed, and the color histogram.
- Do static foreground object detect, if the position of the two objects is nearly the same in continuous N frames, we consider the object is stay in the scene, then update the mask image to notice the MoG algorithm do not update the weights of Gaussian models of this object.

Else

- If the object's position is on the border of the frame, and the direction is outward, delete the object in stable list.
- Else do blob detection, compare the foreground pixels of the stable object with the foreground image, if the superposition area is over a threshold, segments the object from the blob.

Where the mask image is used to set the region that should be updated in new frame.

Particularly, in comparing color histogram stage, the algorithm compares the distance between the maximum color components of the two objects, in view of efficiency, the color component is set to 64, and every color component has four gray levels.

While the tracking is over, the Gaussian modeling algorithm will use the feedback information to perfect foreground. In this way, the foreground objects' pixels and contour is perfect than before, for it is using of posterior information.

120.5 Experimental Result

We implement our algorithm with the OpenCV library. We demonstrate the performance of our algorithm through a number of video sequences and compared with [1] and [10], the results is promising.

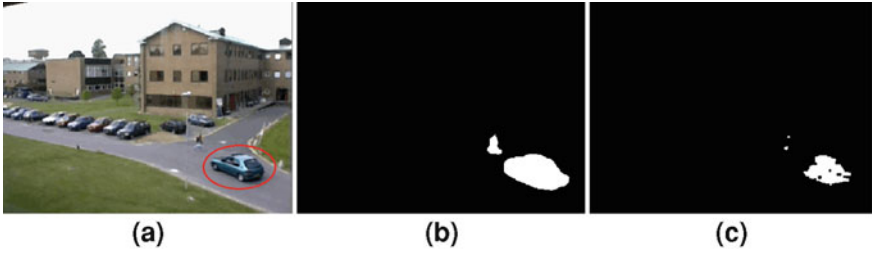


Fig. 120.3 A car come into the scene and its foreground image at frame 538

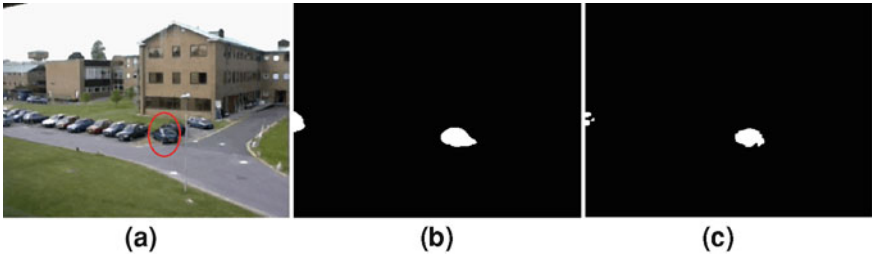
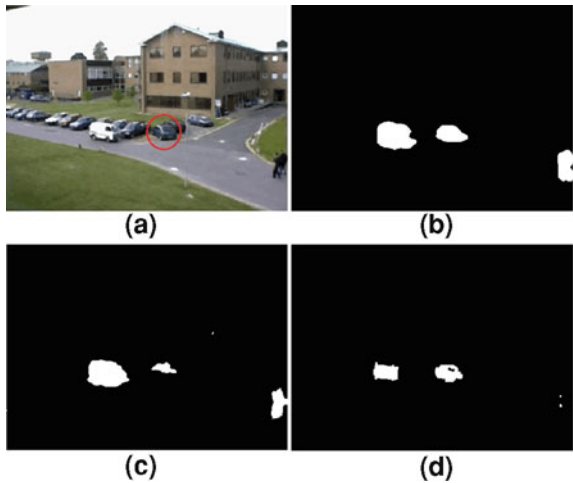


Fig. 120.4 The car stayed in a carport and its foreground image at frame 683

Fig. 120.5 The foreground of the car at frame 782



In the experiment, we track a car in a campus. The car come into the scene, then stayed at a carport. The resolution of the test video is 384*288.

Fig. 120.6 The foreground of the car used our algorithm at frame 2050

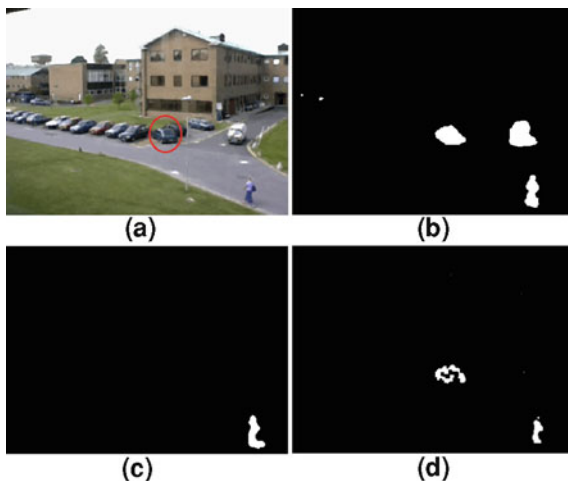


Fig. 120.7 The tracking result

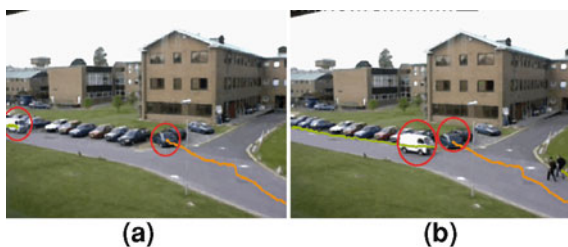


Figure 120.3a shows the car come into the scene, Fig. 120.3b shows the foreground result used our algorithm. Figure 120.3c shows the result used [10].

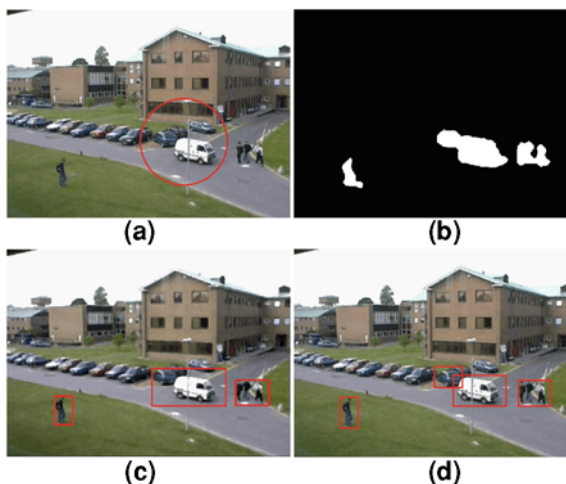
Figure 120.4a shows the car stayed in a carport, Fig. 120.4b shows the foreground result with our algorithm. Figure 120.4c shows the result used [10].

Figure 120.5 shows the difference between our algorithm and [1] and [10], Fig. 120.5c shows the foreground of the same car in Figs. 120.3 and 120.4 used normal MoG at frame 782, Fig. 120.5d shows the result used [10], Fig. 120.5b shows the result used our algorithm at the same frame. We can see that the foreground of the car has fade away in (c) rapidly, and the foreground is not integrated in (d), while the foreground of the car remains integrated in (b).

Figure 120.6a shows the same car at frame 2050, the foreground of the car is still integrated. And Fig. 120.6b shows the result used [1] and Fig. 120.6c shows the result used [10]. Figure 120.7 shows the tracking result.

Figure 120.8 shows the result of segmenting blob into objects. The blob cars are drawn in red circle in Fig. 120.8a. The Fig. 120.8b shows the foreground image of (a), it is obviously that the two cars combined to one blob, and Fig. 120.8c shows

Fig. 120.8 The result of segmenting one blob into different objects



the result used normal MoG, the algorithm detect the two objects to a blob, while in Fig. 120.8d, our algorithm segment the blob into two different objects.

120.6 Conclusion

In this chapter, we propose a robust and accurate algorithm to perfect foreground object in video surveillance. A statistical-based algorithm is proposed to achieve object tracking, and utilize the posterior information to perfect foreground object. Experimental results show that the algorithms we proposed can extract foreground objects efficiently and accurately.

Acknowledgments This work is supported in part by Zhejiang natural science key foundation (Grant No.Z1101243 and No.Y1111017), national natural science foundation of China (Grant No.61003188), and Zhejiang Prior Themes and Social Development Project (Grand No.2009C03015-2).

References

1. Stauffer C, Grimson W (1999) Adaptive background mixture models for real-time tracking. *Comput Vis Pattern Recogn* 2:246–252
2. Li L, Huang W, Gu IY, Tian Q (2002) Foreground object detection in changing background based on color co-occurrence statistics. *IEEE Workshop on applications of computer vision*, Orlando, 269–274

3. KaewTraKulPong P, Bowden R (2001) An improved adaptive background mixture model for real-time tracking with shadow detection. In: European workshop on advanced video based surveillance systems, Kluwer Academic, Dordrecht
4. Rittscher J, Kato J, Joga S, Blake A (2000) A probabilistic background model for tracking. In: Proceedings of 6th European conference computer vision, vol 2, pp 336–350
5. Stenger B, Ramesh V, Paragios N, Coetzee F, Bouhman J (2001) Topology free hidden markov models: application to background modeling. In: Proceedings of IEEE international conference computer vision pp 294–301
6. Yang Y-H, Levine MD (1992) The background primal sketch: an approach for tracking moving objects. *Mach Vis Appl* 5:17–34
7. Elgammal A, Harwood D, Davis L (2000) Non-parametric model for background subtraction. In: European conference computer vision 751–767
8. Askar H, Li X, Li Z (2002) Background clutter suppression and dim moving point targets detection using nonparametric method. In: International conference on communications, circuits and systems and West Sino expositions, vol 2, pp 982–986
9. Thirde D, Jones G (2004) Hierarchical probabilistic models for video object segmentation and tracking. In: International conference on pattern recognition, vol 1, pp 636–639
10. Oliver NM, Rosario B, Pentland AP (2000) A bayesian computer vision system for modeling human interactions. *IEEE Trans Pattern Anal Mach Intell* 22:831–843
11. Ko T, Soatto S, Estrin D (2008) Background subtraction on distributions. In: ECCV, vol 3, pp 276–289

Chapter 121

Video Deformation Based on ASM

Xiaofei Feng and Yujia Ge

Abstract For the edit issues of movement object, facial check and ASM are proposed to extract the feature points of human face. And then model of tracking facial movement object in video sequence is established. In order to assure the consistency, the mean value coordinate is applied to realize the mapping transform between model and feature points. And then deformation technology based on feature segments is integrated to automatically realize the facial caricature for video sequence.

Keywords Caricature · ASM · Video deformation

121.1 Introduction

A caricature has been defined as an exaggerated likeness of a person made by emphasizing all of features that make the person different from everyone else. A caricature makes a more powerful impression than a portrait because of the way that it stresses individuality. It is not easy for person who does not have the necessary ability to draw a caricature. How to accomplish caricature generation of human face is becoming an important issue. In order to apply caricature generation, it needs to check the facial features and obtain the exact positions, including the eyes, nose, mouth, facial contour and other features needed.

X. Feng (✉) · Y. Ge
College of Computer and Information Engineering, Zhejiang Gongshang University,
Hangzhou, China
e-mail: xffeng@mail.zjgsu.edu.cn

Y. Ge
e-mail: yge@mail.zjgsu.edu.cn

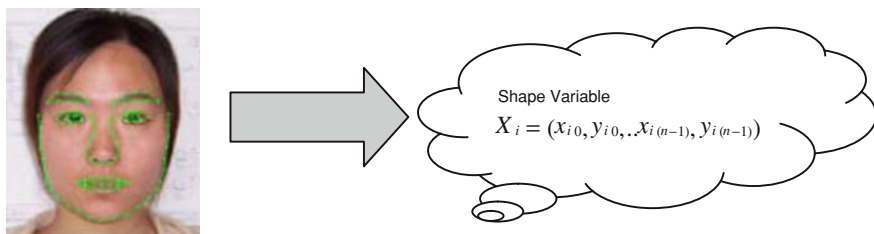


Fig. 121.1 Mathematics expression for demarcated image

It is difficult for computer to automatically analyze facial features [1], because the face has complex 3D structure. There are apparently distinguishing for the 2D images got in different environments. Such differences include facial gesture, expression, and illumination conditions. If we cannot track the feature points of video sequences, it is difficult to realize the exaggerated deformation for facial features. We adopt Active Shape Model (ASM) to realize the automatic track and orientation, then combined with the deformation technology based on feature segments, and automatically realize the facial caricature for video sequence.

121.2 ASM

ASM [2–4], a method for searching features based on statistic model, is involved in two steps. First, the model is trained and studied based on training data. Second, search the object based on the training model. The statistic model includes two kinds of information. One is shape statistic model which points out the changing rule of the 2D shapes, the other is local texture model expressing the variational rule of grayscale.

121.2.1 Shape Model

For a demarcated facial image, the demarcated coordinate can be used to realize the shape's mathematics expression. For many demarcated facial images, the shape's mathematics expression will be random vector, which can be applied to constitute facial shape model by principal component analysis.

(1) Mathematics expression for demarcated image

According to Fig. 121.1, for a demarcated facial image, the shape vector, X_i , can be expressed by each demarcated feature point as (1). Where, x_{ij} , y_{ij} , is the horizontal and vertical coordinate value, i is the ordinal number of images, j is the ordinal number of feature points, n is the amount of feature points.

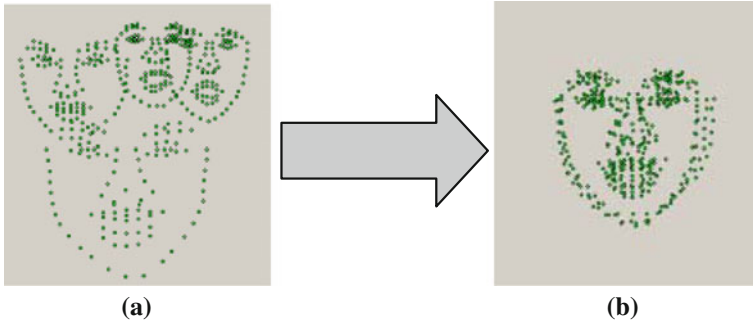


Fig. 121.2 Shape adjustment. **a** Original demarcated shape. **b** Adjustment result

$$X_i = (x_{i0}, y_{i0}, \dots, x_{i(n-1)}, y_{i(n-1)}) \tag{121.1}$$

(2) *Adjustment for demarcated images*

For those demarcated images, it is unreasonable to set up model directly using statistics method because of the different absolute position, the different image size, and the different orientation shown in Fig. 121.2a. The adjustment is to approximate the shape of the images according to rotation, scaling, and transformation. The result is shown in Fig. 121.2b

The process is as follows. First, choose a good shape vector as the original sample, and then adjust all vectors to approximate the chosen sample. Then calculate the average shape and then regularize it. Choose the average and regularized shape as the sample, and then all the shape vectors to approximate it. Repeat such process until the distinction between the near two average shape vectors is less than a set value.

(3) *Shape modeling*

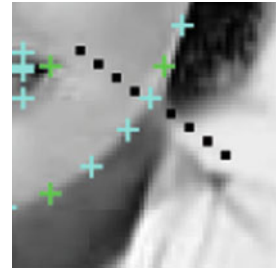
Set the N shape vectors as $X_i (i = 1, \dots, N)$, and calculate the average value, and the covariance matrix according to (121.2), (121.3).

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i \tag{121.2}$$

$$C = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(X_i - \bar{X})^T \tag{121.3}$$

And then compute the eigenvalue $(\lambda_0, \lambda_1, \dots, \lambda_{2n-1})$ eigenvector $(P_0, P_1, \dots, P_{2n-1})$. Because the eigenvector corresponding to the great eigenvalue has much information about shape diversification, so we choose some eigenvectors corresponding

Fig. 121.3 Sketch map along normal direction for demarcated points



to some great eigenvalue to approximately express the random shape vector. Choose the greatest eigenvalues, which can satisfy the formula (121.4), where $t = 98\%$.

$$\frac{\sum_{k=0}^j \lambda_k}{\sum_{k=0}^{2n-1} \lambda_k} < t \tag{121.4}$$

In this way, we can obtain the eigenvectors, $P_s = (P_0, P_1, \dots, P_{2n-1})$. P_s can be recognized as the new orthonormal basis. Each shape vector can be expressed as (121.5). Where b_s is the shape parameter of a shape vector, and it can be calculated according to (121.6). Different b_s represents the different diversification. In this way we create the shape model and can create the new shape samples if b_s is changed.

$$X = \bar{X} + P_s b_s \tag{121.5}$$

$$b_s = P_s^T (X - \bar{X}) \tag{121.6}$$

121.2.2 Local Texture Model

The local texture model is used to do statistic for the grayscale changing situation of each feature point along the normal direction, which can help to get the grayscale distributing rule of the corresponding feature pint. First find the normal direction of each feature point for every training image shown in Fig. 121.3, and take n_p pels pints centered on the given feature pint. The vector that is composed of grayscale values is shown as (121.7). Where k is the number of images, and $k = 0, \dots, N-1$, i is the number of feature points, and $i = 0, \dots, n-1$. The differential coefficient along the normal direction, the regularized result, the average value of N training images, and the covariance matrix are obtained as (121.8)–(121.11).

$$h_{ki} = (h_{ki0}, \dots, h_{ki(n_p-1)}) \tag{121.7}$$

$$dh_{ki} = \left(h_{ki1} - h_{ki0}, \dots, h_{ki(n_p-1)} - h_{ki(n_p-2)} \right) \quad (121.8)$$

$$g_{ki} = \frac{dh_{ki}}{\sum_{q=0}^{n_p-2} |h_{ki(q+1)} - h_{kiq}|} \quad (121.9)$$

$$\bar{g}_i = \frac{1}{N} \sum_{k=1}^N g_{ki} \quad (121.10)$$

$$C_i = \frac{1}{N} \sum_{k=0}^{N-1} (g_{ki} - \bar{g}_i)(g_{ki} - \bar{g}_i)^T \quad (121.11)$$

We can think that the normalized grayscale derivative vector satisfies the Gauss distribution, so the Mahalanobis distance between g_i of each pint and the averaged value is shows as (121.12). Such distance also reflects the probability of the feature point, and we can judge the best point from the corresponding probability while searching the feature points.

$$d = (g'_i - \bar{g}_i)^T C_i^{-1} (g'_i - \bar{g}_i) \quad (121.12)$$

121.2.3 Process of Searching Object

We can begin to search by using the statistic shape model and statistic facial model. The detailed process is as follows.

- (1) Locate the model originally. For an unknown image, we have to locate the object originally according to the approximate position of the object by other image processes. Initialize the ASM model by such position, which can make the initialized shape approximate to the object. For the facial image, we can adopt some algorithms of facial checking to judge the approximate position of face. According to the original position, we can make the average shape model rotate, θ , zoom, s , and displace, t . And we can get the original shape, X , as (121.13).

$$X = M(s, \theta)\bar{X} + t \quad (121.13)$$

- (2) Search along the normal direction of the feature points. Take advantage of the local grayscale model, we can search along the normal direction for each feature point, calculate the Mahalanobis distance, and select the best match point. Search the best match point for each feature point, which can help to get a new shape vector, X' .
- (3) Because the new shape vector cannot be expressed by the new orthonormal basis directly, it only can be approximately expressed under the minimum

error. So we first make affine transformation from X to X' , and get the four parameters, $1 + ds$, $d\theta$, dx , and dy . In this way we can obtain the diversification value of shape parameter as (121.14), where $dx = X' - X$

$$db = P^T P P^T (M((s(1 + ds))^{-1}, -(\theta + d\theta)) [M(s, \theta) [\bar{X}] + dx - dy] - \bar{X}) \quad (121.14)$$

- (4) In order to assure the validity of the shape parameter, we need to make some restriction. Original set $b = 0$. From the shape model we know that the diversification range of the shape parameter, b_i , ought to be among $\pm 3\sqrt{\lambda_i}$. So we judge the value of $b + db$, if it is out of the given range, we should do some transformation and make it among the given range. And take advantage of the final shape parameter to calculate the reconstruct shape.
- (5) The above four steps compose a cycle search. If Euclidean distance between the two shape vector got by twice cycle is great than some value, repeat the 3, 4 step, otherwise we can think it convergences.

121.2.4 The Result of Video Tracking

In order to assure the algorithm validity, we verify it by images and video. The result is shown in Fig. 121.4. The experiment results show that the ASM can effectively capture and track the feature points.

121.3 Deformation Technology Based on Feature Segment

In order to realize image caricature, we need to use contorting distortion to transform the original image into a caricature according to some rules. The image deformation technology based on feature segment is proposed by Beier and Neely [5]. It takes the representative segments of the original image and the objective image to define the feature coordinate mapping, and the other points are ensure the corresponding relationship according to the distance between the point to the segment. It usually uses reverse mapping to valuate the image deformation. Scan each pixel to find the corresponding pixel, each pixel of the objective image can be filled properly. For complicated image, many feature segments are used. Set the weight by using the distance between each pixel to feature segment and the length of the segment, and make weighted average according, which can ensure the displacement of the pixel of original image. Such deformation result are shown in Fig. 121.5.

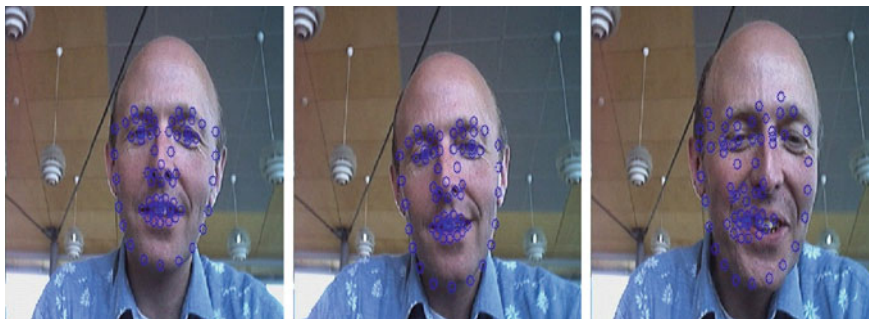


Fig. 121.4 The result of video tracking

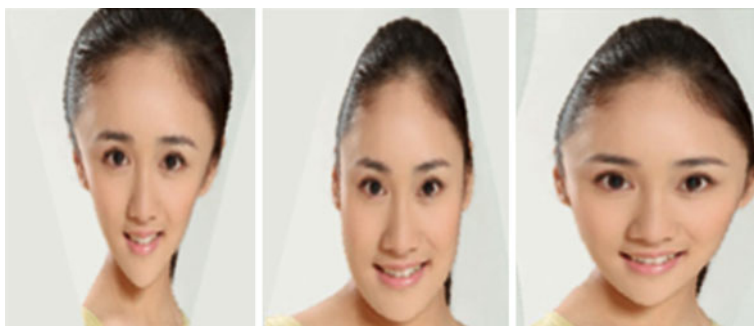


Fig. 121.5 The result of deformation

121.4 Realization of Video Deformation

The process of video deformation is described as follows. First the user realizes the interactive edition for the original frame. Create the mapping relationship between the deformation point and AAM, and obtain the deformation feature position of the continuous frame from mapping relationship. The key problem of above flow is how to effectively track the deformation feature points. First we set the interactive frame be the first frame. The set of the contour feature points from AAM is Pr , $Pr = (Pr,1, Pr,2, \dots, Pr,i, \dots, Pr,N)$, where N is the number of the contour points. The set of deformation control points is Cr , $Cr = (Cr,1, Cr,2, \dots, Cr,i, \dots, Cr,M)$, where M is the number of the control point that are used to construct the feature segment. The set of control points after deformed is $C'r$, $C'r = (C'r,1, C'r,2, \dots, C'r,i, \dots, C'r,M)$. Apparently, in order to realize the algorithm above, the key issue is how to effectively track the deformation feature points.

We solve this problem by mean value coordinates. The basic principle of mean value coordinates is to define one close polygon gridding. Each point of the plane can be expressed by the linear combination of peaks of the polygon gridding, which is shown as (121.15). Here $f_i(v_i)$ is the function value of v_i , m_i is the weight



Fig. 121.6 The result of video deformation

of v_i relative to the inserted point x . The function value can be the color value, texture value, and coordinate value. For caricature deformation, function value f_i is the coordinate value v_i , so the above formula can be expressed as (121.16), (121.17). It is clear that arbitrary point beyond the peaks can be expressed by mean value coordinate. We can translate each position of deformation point into the linear combination of feature points. For the comparative stability of the feature points in video frequency, the stability of such linear combination is good. For the following frame sequence, we can automatically compute the position of deformation point, and than output the caricature by using the deformation method. The caricature effect is shown in Fig. 121.6.

$$\begin{cases} f(\mathbf{x}) = \sum_i m_i f_i(v_i) \\ \sum_i m_i = 1 \end{cases} \tag{121.15}$$

$$\mathbf{x} = \sum_i m_i v_i \tag{121.16}$$

$$\begin{cases} m_i = w_i / \sum_{t=1}^n w_t \\ w_i = |v_i - v_t| \end{cases} \tag{121.17}$$

Acknowledgments The work was supported by the Natural Science Foundation of Zhejiang under Grant No. Y1100824 and Zhejiang Public Interest Project of Technology Bureau under Grand No. 2010C31088.

References

1. Zhou R, Zhou JI (2006) Caricature generation based on facial feature analysis. *J Comput Aid Design Comput Graph* 18(9):1362–1366
2. Kass M, Winkin A, Terzopoulos D (1987) Snakes: active contour models [A]. In: *Proceedings of first international conference on computer vision [C]*. IEEE Computer Society Press, London, pp 259–268
3. Yan F, Fei GZ (2006) A generation algorithm of caricature portrait. *J Comput Aid Design Comput Graph* 19(4):442–447
4. Lee E-J, Kwon J-Y, Lee I-K (2007) Caricature video. *Comput Animat Virtual Worlds* 18:279–288
5. Beier T, Neely S (1992) Feature-based image metamorphosis [A]. *Comput Graph Proc [C]*. ACM Press, Chicago pp 35–42

Chapter 122

An Enhanced Hybrid Image Watermarking Algorithm Using Chaotically Scrambled Technology

Niansheng Cheng

Abstract Image encryption algorithm is of high iteration, and low privacy so we proposed an image watermarking method in discrete cosine transform embedded with chaos scrambling. This algorithm encrypts the watermark image and chooses embed-position randomly for the sake of security. By this way, watermarking can be detected and extracted without the original image, and coding based on chaos keeps secrets better. The experiment results indicate that the proposed method can be realized easily, which keeps and hides the watermarking information better. The proposed method performs well in encountering with gaussian and salt & pepper noise attacks.

Keywords Image watermarking · Chaos scrambling · Discrete cosine transform

122.1 Introduction

In recent years, with the rapid development of Internet and multimedia technology globally, the storage, edition, copying and communication of digital works become very easy thanks to the Internet and computer. As a potentially effective means for solving the issue of protection of digital product copyright, digital watermarking technology has drawn extensive attention. This technology inserts information (watermarking) with certainty and confidentiality to digital products (e.g. static image, voice, document, video, etc.) to enable them to become a part of the raw

N. Cheng (✉)
Department of Information Engineering, Hunan Urban Construction College,
Xiangtan, China
e-mail: 13327322502@189.cn

data and be saved so as to realize various functions including concealed transmission, storage, marking, identification, copyright protection. While, the watermarking algorithm is regarded as a digital communication theory in which assisted information is invisibly spread through host data. Most watermarking algorithms insert the watermarking to medium and high frequencies in the transformation domain. Transforms selected for watermarking embedding mainly include discrete cosine transformer (DCT) [1, 2] and discrete wavelet transform (DWT) [3], etc. The research on implantation of digital watermarking on the basis of DCT according to the static image compression standard of JPEG has become a highlight; the literature [4] has come up with two algorithms suitable for space domain and frequency domain. This kind of approach has better utilized the characteristics of human visual system. The watermarking information is hid in the frequency spectrum part that is relatively crucial to digital image perception after going through certain modulation process, so as to resist lossy compression and other digital image processing operations.

However, the abovementioned digital watermarking technology fails to adopt encryption measure. Those who have no authority can easily obtain the digital watermarking inserted and modify it so as to influence the safety of watermarking. This chapter has proposed a digital watermarking algorithm based on chaotically scrambled DCT domain to handle this problem. This algorithm quantizes the watermarking image in DCT domain according to JPEG quantization table. The first ten coefficients of each image after quantization will be used to generate watermarking signals after modulation processing. Also, two-dimensional chaotic sequence will be used to encrypt the watermarking so as to improve safety. The experimental result shows this algorithm features satisfactory invisibility and outstanding anti-attack performance.

122.2 Watermarking Embedding Algorithm

Due to massive image data, the watermarking image must be compressed to reduce the data volume inserted. The calculation amount is relatively small based on discrete cosine transformer (DCT). Also, it is compatible with international data compression standards (JPEG, MPEG and H261/263), and thus is easily realized in compression domain. Therefore, two-dimensional DCT will be used to transform the original image before the embedding of watermarking. Then, Arnold scrambled chaotic binary sequence to modulate the watermarking information to enhance the transparency and robustness of watermarking system. The basic block diagram of watermarking embedding algorithm is shown in Fig. 122.1:

Watermarking Image Preprocessing. The watermarking image X is a matrix with M lines and N columns. The information data volume is relatively big. For example an image with the dimension of $64 \times 64 \times 8$ has totally 4,096 bytes. In order to reduce calculation amount and weaken or eliminate image data dependency, two-dimensional DCT is used to analyze it and divide the watermarking

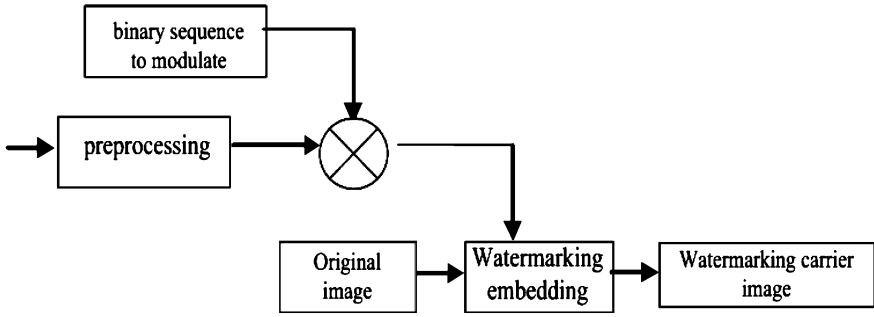


Fig. 122.1 Basic block diagram of watermarking embedding algorithm

image to 8×8 blocks marked as $B_i = f_i(x,y), i = 0, 1, \dots, I-1$. These blocks do not cover with each other. Then, carry out DCT transformation for B_i to obtain:

$$B'_i = Y(k, l) = DCT\{f_i(x,y), 0 \leq x, y \leq 7\} \quad i = 0, 1, 2, \dots, I-1 \quad (122.1)$$

where

$$Y(k, l) = \frac{2}{\sqrt{MN}} c(k)c(l) \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} X(m, n) \cos \frac{(2m+1)k\pi}{2M} \cos \frac{(2n+1)l\pi}{2N} \quad (122.2)$$

where: $m, k = 1, 2, \dots, M-1$; $k = 0, 1, \dots, M-1$; $n, l = 1, 2, \dots, N-1$. Function $c(k)$ and Function $c(l)$:

$$c(k) = \begin{cases} 1/\sqrt{2} & k = 0 \\ 1 & k = 1, 2, \dots, M-1 \end{cases}$$

$$c(l) = \begin{cases} 1/\sqrt{2} & k = 0 \\ 1 & k = 1, 2, \dots, N-1 \end{cases}$$

Quantize and adjust the DCT coefficient in B'_i through the quantization table of JPEG and obtain:

$$B''_i = \text{int}(F(k, l)/Q(k, l)) \quad 0 \leq k, l \leq 7, \quad i = 0, 1, 2, \dots, I-1 \quad (122.3)$$

round () is the integral function. At last, take 10 DCT coefficients from B''_i and mark as S_i , i.e.

$$S = \bigcup_{i=0}^{I-1} S_i = \{d_{kl}, d_{kl} \in B''_i, \quad k = 0, 1, 2, 3 \quad l = 0, \dots, 3-k\} \quad (122.4)$$

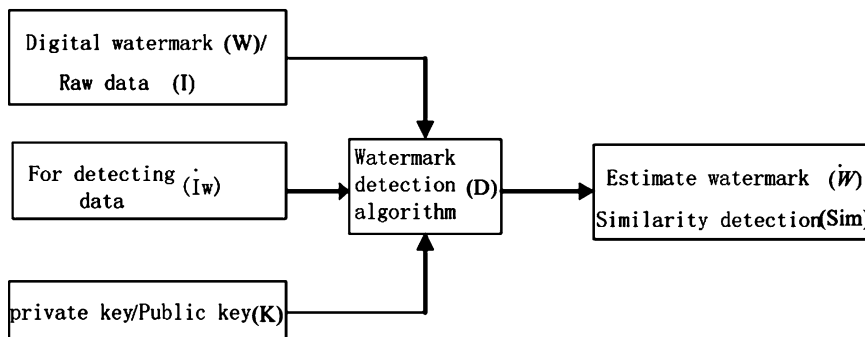


Fig. 122.2 Watermarking detection process

The signal S we obtain from the abovementioned processing is much smaller than the data volume of the original image.

Generation of Watermarking Signal. After converting S to binary code (each figure in S is represented by using 7-bit binary code), we obtain the watermarking sequence W that only contains 0 and 1. Use numerical sequence to represent $W = \{W_1, W_2, \dots, W_I\}$. Use tow-value chaotic sequence to modulate each element of watermarking sequence W [5, 6].

The method for generation of binary chaotic sequence is as follows: define real-value sequence first, i.e. $\{x(k); k = 0, 1, 2, \dots\}$ which is a sequence formed by the track point of logistic chaotic mapping through definition of one threshold function Γ :

$$\Gamma(x) = \begin{cases} 0 & -1 \leq x < 0 \\ 1 & 0 \leq x \leq 1 \end{cases} \tag{122.5}$$

We can obtain the binary chaotic sequence of $\Gamma(x(k))$ from the abovementioned real value $x(k)$, i.e. $\{\Gamma(x(k)); k = 0, 1, 2, \dots\}$. The watermarking signal with chaotic scrambling is obtained from the dot product of watermarking sequence W and binary chaotic sequence $\Gamma(x(k))$ [7, 8].

122.3 Watermarking Detection and Extraction Algorithm

Figure 122.2 shows the detection process of watermarking.

From Fig. 122.2, a general formula of the watermarking detection process can be defined as:

When there is original carrier data:

$$\hat{W} = D(\hat{I}_w, I, K) \tag{122.6}$$

When there is original watermarking W :

$$\widehat{W} = D(\widehat{I}_W, W, K) \quad (122.7)$$

When there is no original information:

$$\widehat{W} = D(\widehat{I}_W, K) \quad (122.8)$$

where: \widehat{W} represents estimated watermarking; D is watermarking detection algorithm; \widehat{I}_W represents the watermarking carrier data after being attached during the transmission process. The approaches for watermarking detection can be classified to two types: one is to abstract or carry out dependence verification for the inserted signal under the condition of existence of original information; the other is to conduct full-search or distribution hypothesis inspection, etc. as a must under the condition of no original information. Similarity test is usually adopted to detect whether the signal is watermarking signal. The general formula of watermarking similarity test is as follows:

$$Sim = \frac{W * \widehat{W}}{\sqrt{W * W}} \text{ or } Sim = \frac{W * \widehat{W}}{\sqrt{W * W} \sqrt{\widehat{W} * \widehat{W}}} \quad (122.9)$$

where: \widehat{W} represents estimated watermarking; W represents original watermarking; Sim represents the similarity of different signals.

Watermarking abstraction algorithm is the inverse process of embedding algorithm. First, we utilize the binary chaotic sequence in embedding algorithm to modulate the watermarking signal detected. Thus, the watermarking signal is recovered as W^* . Then, the value in W^* will go through inverse mapping. In other words, -1 in W^* is mapped to 0 to obtain W' ; W' will be divided into 8×8 image sub-blocks, and inverse quantization will be carried out. Finally, inverse DCT transformation will be carried out through the following formula, and we can recover watermarking image X .

$$X(m, n) = \frac{2}{\sqrt{MN}} \sum_{K=0}^{M-1} \sum_{L=0}^{N-1} c(k)c(l)Y(k, l) \cos \frac{(2m+1)k\pi}{2M} \cos \frac{(2n+1)l\pi}{2N} \quad (122.10)$$

122.4 Experimental Results and Performance Analysis

The original image adopted in the experiment is 512×512 standard grayscale image with Lena as the carrier image. The watermarking image is the binary image with the dimension of 64×64 . Similarity is used to evaluate the difference between carrier image and image containing watermarking.

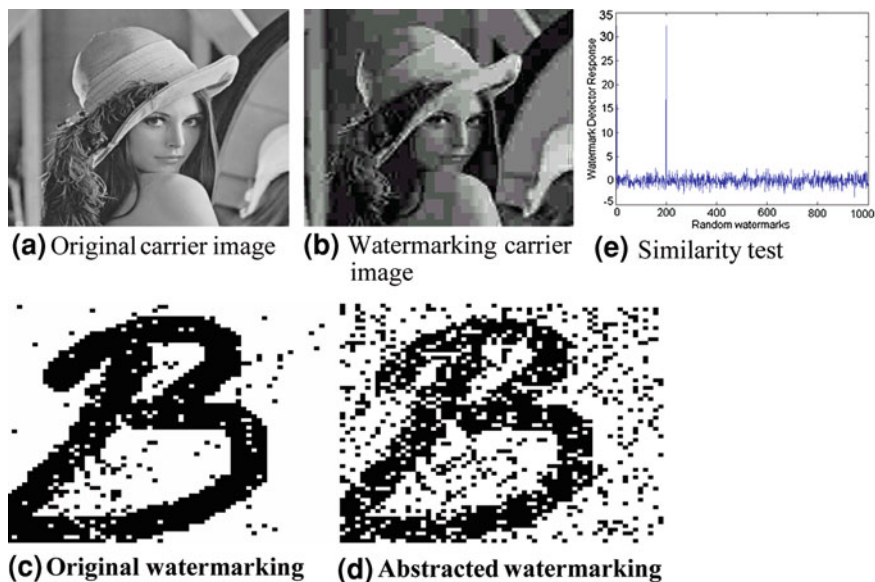


Fig. 122.3 Simulation Result of Watermarking Image with JPEG Processing. **a** Original carrier image. **b** Watermarking carrier image. **c** Original watermarking. **d** Abstracted watermarking. **e** Similarity test

Watermarking Detection after JPEG Processing. Carry out JPEG compression coding for the watermarking image. The compression quality is 10%. Also, two-dimensional DCT and binary chaotic sequence are used to modulate to obtain watermarking carrier image. In Fig. 122.3 (a) is the original Lena image; (b) represents watermarking carrier image; (c) represents original watermarking; (d) represents abstracted watermarking; (e) is the result of similarity detection. The simulation result shows this algorithm is of outstanding visual effect. The watermarking inserted changes the original image in a minimum way and cannot be easily noticed. From the response value, we can see the watermarking abstracted is of uniqueness, and the similarity detection value reaches approximately 32.

Watermarking Detection after Added with Noise. The watermarking is detected after the image inserted with watermarking is added with Gaussian and salt & pepper noise. The response value is around 20. The simulation result is shown in Fig. 122.4 where (a) represents the watermarking carrier image added with Gaussian and salt & pepper noise while; (b) represents the result of similarity detection. We can see this kind of algorithm is of outstanding anti-attach capacity for this high-frequency noise.

Figure 122.3 shows the JPEG compression coding for watermarking image when the compression quality is 5%. From Fig. 122.2, we can see the block effect of the original watermarking image is already very obvious. At this point, the response can reach 10 if the similarity detection is conducted. Thus, it can still well detect the existence of watermarking signal.

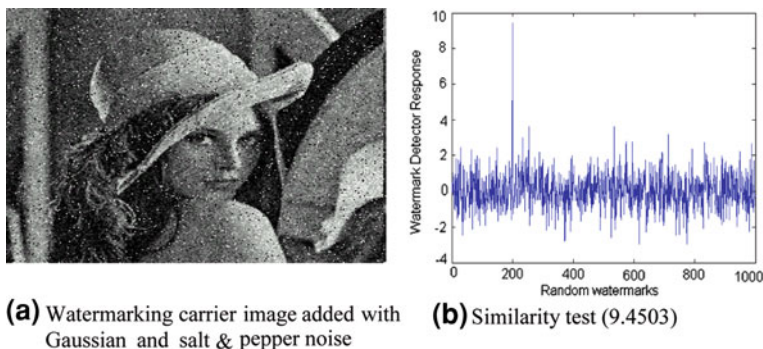


Fig. 122.4 Simulation result of watermarking image added with Gaussian and salt & pepper Noise. **a** Watermarking carrier image added with Gaussian and salt & pepper noise; **b** Similarity test (9.4503)

122.5 Conclusion

This chapter has provided a new digital image watermarking embedding algorithm based on DCT. Since the chaotic scrambling is introduced, the confidentiality of the whole algorithm becomes completer. When the secrete key of the chaotic sequence is unknown, it is basically impossible for reproduction of watermarking image, which makes sure that secrete key owner is able to correctly abstract watermarking information. It is very practical for copyright protection by adding watermarking in the static image.

References

1. Jiwu H, Shi YQ, Weidong C (2000) Image watermarking in DCT: an embedding strategy and algorithm[J]. *Acta Electron Sin* 28(4):57–60
2. Jian Z, Tian Z, Mingquan Z et al (2006) Software realization of DCT watermarking scheme[J]. *Comput Appl Softw* 23(8):52–53
3. Tsai M, Yu KY, Chen YZ (2000) Joint wavelet and spatial transformation for digital watermarking[J]. *IEEE Trans Consumer Electron* 23(8):237–246
4. Liang X, Zhihui W, Wu H (2004) Research of image content based digital watermarking model and algorithm in ridgelet transform domain[J]. *J Electron Inf Technol* 26(9):1440–1448
5. Kelan Y, Jianming Z, Desheng X (2006) Binary image encryption algorithm based on chaotic sequences[J]. *Comput Technol Dev* 16(2):148–151
6. Yongqiang C, Huaning S (2004) A digital image encrypting algorithm based on two dimensional chaotic mapping[J]. *J Wuhan Polytech Univ* 23(4):45–47
7. Yongwei Y, Qinghua Y, Yingyu W (2003) Magic cube encryption for digital image using chaotic sequence[J]. *J Zhejiang Univ Technol* 32(2):173–176
8. Li M, Yong F, Li L (2008) An image encryption approach based on a two-dimensional reversible map[J]. *Comput Simul* 25(2):227–231

Chapter 123

A Digital Watermarking Technology Based on Wavelet Decomposition

Song Yijun and Yang Gelan

Abstract With the rapid spread of computer networks and the further progress of multimedia technologies, copyright protection of digital media has been one of the most important issues in information technology field. In this paper, we analyse a new method for embedded digital watermarking in digital image based on the wavelet transform. We perform the wavelet multi resolution for digital image and determine the coefficients for embedded watermarking by computing self adaptation value, we choose random numbers produced by Gauss distribution as watermarking information, the watermarking are embedded in seven subbands of three stages of DWT decomposition except for HL1,LH1,HH1 subbands. The experiments show that the watermarking is perceptually invisible and robustious.

Keywords Digital watermarking · Wavelet transform · Random number

123.1 Introduction

Along with the fast development of computer Internet technology, much attention have been paid to the problems of the protection of digital media like digital audio, digital image and digital video. The watermarking technology has become one of the most active researches since 1995 [1]. It can prevent the digital products from

S. Yijun

Department of Computer, Hunan Hengyang Normal Nniversity,
Hengyang 421002, China
e-mail: haihai757@163.com

Y. Gelan (✉)

Department of Computer, Hunan City University, Yiyang 413000, China
e-mail: glyang@mail.ustc.edu.cn

copying and spreading when combined with cryptography, which has been proven to be accessible and efficient in the articles written by Cox and other scientists [2].

Generally, the digital watermarking technology is composed of embedding algorithm and detection algorithm [3]. Through the key, the former forms the semi-fragile mark to hide the watermark in the original data; the latter detect or recover the watermark from the semi-fragile mark. While without the key, others cannot obtain, delete or even find the watermark from the semi-fragile mark. With the research going deep, the digital watermarking technology has developed rapidly and extends in other areas such as copyright protection, implicit signals, certification and secret communications. An efficient digital watermarking technology must meet the following requirements:

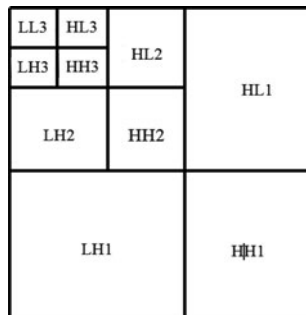
- (1) Imperceptivity: the embedding methods of digital watermarking would not obviously lower the quality of raw data and imagery; also, it must be hard for the attackers who could use statistics methods to find and delete the digital copyright symbol;
- (2) Robustness: it must be difficult for the attackers who can use signal processing, distortion and watermarked several times to break watermark or falsify the digital copyright symbol;
- (3) Safety: it is available only to certain authorized illegal users who can detect, extract and even alert the watermark.

This passage has put forward a new method which digital watermark is added to the wavelet coefficients. In this way, we choose random numbers produced by Gauss distribution as watermarking information, the watermarking were embedded in seven sub-bands of three stages of DWT decomposition except for HL1,LH1,HH1 sub-bands and determine the coefficients for embedded watermarking by computing self-adaptation value.

123.2 Embedding Watermarks

The general mind of image compression realized by means of integer wavelet transform is to use multi-resolution decomposition with image, decompose it into different space and different frequency and then encode coefficients to the sub-image. According to the tower type decomposition algorithm, the image is divided into 4 different bands though wavelet multi-resolution, and the low-frequency part can be decomposed deeply. The image energy mostly focuses on the low-frequency parts. The following Fig. 123.1 is the three stages decomposition of the image, here the L (H) means low-frequencies sub-bands, the data in the lower symbol means decomposition stages. When the image is several stages decomposed, the important information would centered in the sub-bands in the top left corner, through which the image can be compressed and embedded watermark.

Fig. 123.1 The image's three stages of DWT decomposition



The algorithm adopts random numbers produced by Gauss distribution as watermarking information. By Box–Muller transformation of random variant which satisfies average distribution, it produces random numbers produced by Gauss distribution [4, 5].

We performed the resolution of DWT for regional icon and determine the coefficients of the former seven sub-bands for embedded watermarking by computing self-adaptation value, for the sub-band of LL3, traverse it and find out the biggest coefficient calls C, using the Eq. (123.1) to calculate the value of sub-band, and it also can calculate the rest class of value. And then embedded in watermark and applying IDWT technology finally results in watermark picture.

$$T_i = 2^{\lfloor \log_2 C_i \rfloor - 1} \tag{123.1}$$

In this equation, i is for resolution level, $\lfloor X \rfloor$ stands for rounded down. The algorithm below using this equation for value is only for the level LL3, the six rest sub-bands use one-third of the biggest coefficient for value, in order to increasing the quantities of embedded in watermark. As to the selected coefficient, we embedded in watermark by Eq. 123.2.

$$V'_i = V_i + \alpha_i V_i X_i \tag{123.2}$$

Here V_i Stands for selected coefficient vector, X_i is watermark vector, α_i is embedded in strength.

In experiment, there are 1,000 embedded in LL3 and others embedded in their own sub-band of the 5,000 watermark information which was putted out by the watermark generator. It did not loaded with watermark information as the energy of the sub-bands HL1,LH1,HH1 is about 1%.

There are three values of the embedded in strength in experiment, $\alpha = 0.007$ for sub-band LL3, and the value of in the first and second level are respectively 0.1 and 0.2, thus can not only ensure it did not have obviously degeneration but also make sure the robustness of watermark is in good condition.

Table 123.1 PNSR of watermark images

Tested images	Lena	Pepper	Fishboat	Mandrill
PNSR	46.64	46.60	43.61	42.36

123.3 Watermark Identifying

Watermark identifying is the inverse process of watermark embedding and the algorithm will need original image. The specific process is: the original image, firstly, is treated with three-ply wavelet decomposition by the detection end and then obtaining a primitive matrix; the watermark image which needs detecting is done the same wavelet decomposition and then achieving a matrix too. Use the Eq. (123.2) and you will get the digital watermark consisted in the image. After extracting watermark X^* , adopt X (the original watermark) and the cross-correlation value calculation to operate approximately detecting:

$$\text{sim}(X, X^*) = \frac{X \cdot X^*}{\sqrt{X^* \cdot X^*}} \quad (123.3)$$

If the similarity value is larger than the domain value which is defined by Eq. (123.4), we can conclude that the image is embedded with watermarks. Or, it is not.

$$S = \frac{\alpha_i}{2M} \sum |V'_i| \quad (123.4)$$

Here, the M stands for the quantity of the embedded watermarks, V'_i stands for the coefficient received after the image having been embedded with watermarks and A_i stands for the embedding strength used during the watermark embedding process.

123.4 Experiment Result

This algorithm adopts the biorthogonal Antionini97 wavelet to deal with the original image by conducting three-level discrete wavelet decomposition. After supplying this algorithm to embed random data watermarks in four kinds of different gray images with 256 grades, the peak signal-to-noise ratios of the images including watermarks are shown as Table 123.1.

Figure 123.2 is the original image of “Lena”. Figure 123.3 is the embedded watermark image and it shows that the image makes no differences in visibility after being embedded with watermarks. Figure 123.4 presents the result the watermark discriminator respond to 1,000 data which are generated randomly.

In order to validate the robustness of the embedding algorithm in this paper, the image of “Lena” containing watermarks is implemented three kinds of attacks including JPEG compression, center shear and collusion attack. We use the

Fig. 123.2 The original image of “Lena”



Fig. 123.3 The image of “Lena” containing watermarks

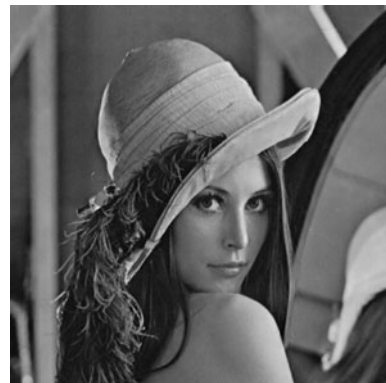
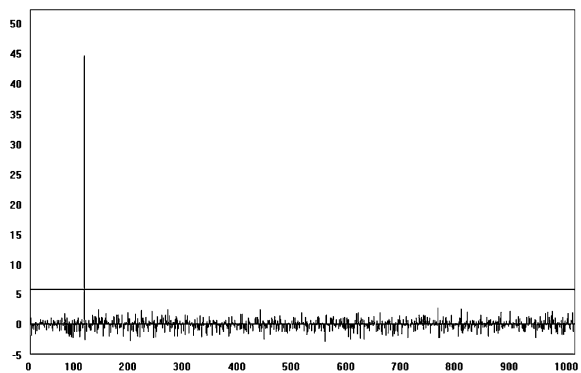


Fig. 123.4 The response the watermark discriminator responds to 1,000 watermark data generated randomly (not attack)



Eq. (123.3) to calculate its similarity value and the results are shown as Figs. 123.5, 123.6 and 123.7 which indicate that the response value the algorithm responds to JPEG compression is smaller, whereas, the response value the algorithm responds to center shear is much better.

Fig. 123.5 Watermark discriminator response (JPEGQ25 compression attack)

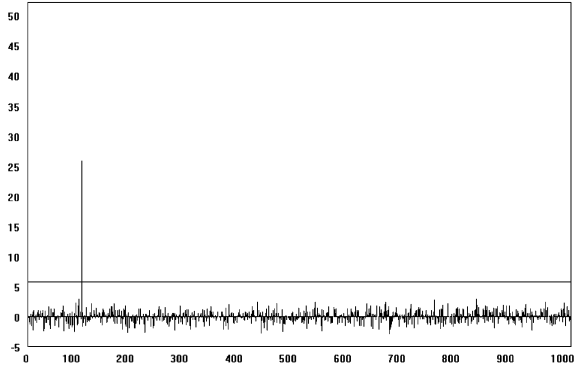


Fig. 123.6 Watermark discriminator response (center shear attack)

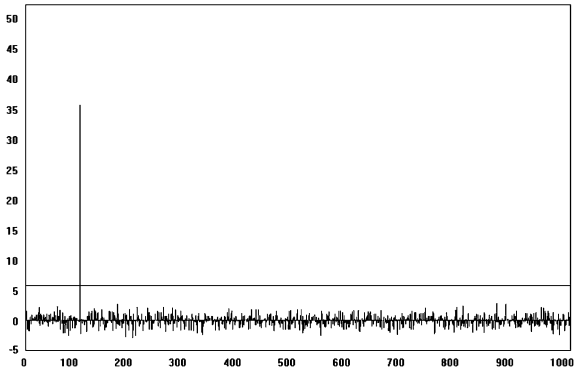
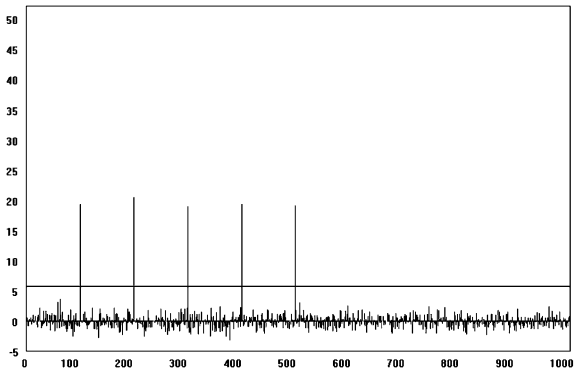


Fig. 123.7 The response that the watermark discriminator responds to collusion attack (Five embedded watermarks can all be identified clearly)



As to collusion attack, we use five watermark images produced independently and do them with average process to simulate collusion attack simply.

Figure 123.7 shows the response that the water discriminator responds to 1,000 watermark data generated randomly, among which there are five watermarks have

been embedded in the image. The five peak values shown in the Fig. 123.7 illustrate clearly that the image contains five watermarks. Meanwhile, it also indicates that it is in vain to treat the watermark image in average process and then to conduct collusion attack.

123.5 Conclusion

This chapter proposes a digital watermark algorithm based on wavelet transform. The algorithm chooses important coefficient attaching to the will-be embedded watermark by selecting domain value freely at every decomposition level and then according to different sub-bands from every decomposition level it can adopt different embedding strength coefficients and improve the imperceptibility and robustness of watermark.

Acknowledgments This work was supported by Hengyang Normal University Science Foundation research funding youth projects under Grant (No. 09A35).

References

1. Andrews HC, Patterson CL (1976) Singular value decomposition (SVD) image coding. *IEEE Trans Comm* COM-24:425–432
2. Barni M, Bartolini F, Cappellini V, Piva A (1998) A DCT-domain system for robust image watermarking. *Signal Process* 66:357–372
3. Barni M, Bartolini F, Cappellini V, Piva A (2001) Improved wavelet-based watermarking through pixel-wise masking. *IEEE Trans Image Process* 10:783–791
4. Celik M, Sharma U, Saber GE, Tekalp AM (2002) Hierarchical watermarking for secure image authentication with localization. *IEEE Trans Image Process* 11:585–595
5. Chang CC, Hwang KF, Hwang MS (2003) A digital watermarking scheme using human visual effects. *Informatica* 24:505–511

Chapter 124

An Improved Rate Control Algorithm of H.264/AVC Based on Human Visual System

Meifeng Liu and Ling Lu

Abstract The spacial and temporal features of video pictures are analyzed based the characteristics of Human Visual System (HVS). Then a analysis model of video pictures is proposed, which will abstract the features of motion and texture of video pictures. Based on the moving region detector and texture analysis model, the perceptual importance of MBs is assigned as four grades. The Lagrange coefficient λ_{mode} and Quantization step length QP are adjusted according to the grades. A adaptive rate-control algorithm of H.264/AVC is proposed at last. Simulation results show that the proposed algorithm can efficiently reduce bit rates without visual quality degradation and gain better subjective video quality under the condition of rate-restricted.

Keywords HVS · H.264/AVC · Rate-control · Perceptual importance

124.1 Introduction

H.264/AVC is the newest international video standard. The rate control in this standard, has become one of the video communication research focuses in this years. Researchers would pay more attention to whether rate-control can increase video

M. Liu (✉)

College of Machine and Electronics Engineering, East China Institute of Technology,
Fuzhou, 344000 Jiangxi, China
e-mail: mliu666@163.com

L. Lu

College of Information Engineering, East China Institute of Technology, Fuzhou,
344000 Jiangxi, China
e-mail: luling@ecit.cn

quality as more as possible in the given channel condition and target rate. The traditional video quality assessment standard is peak signal-to-noise ratio (PSNR) which is always considered as a objective standard. However, with recent research on Human Visual System (HVS), some researchers prefer to a subjective video quality assessment standard which is based on HVS. Tsai et al. [1] and Yu et al. [2] abstract some characteristics of video pictures and adjust some rate control coefficients with simple mathematic models. Among them, Tsai distinguishes the motive regions with random texture from video pictures and improves their values. In Yu' algorithm, one frame picture is divided into three parts of edge, texture and smooth background based on edge covered effect. Sullivan et al. [3] and Ma et al. [4] has analyzed the difference complexity between two adjacent frames in the course of rate assignment. However, Tsai and Yu's algorithms actually only realize spacial covered effect of HVS. The difference of bit waste of pictures brought by different characteristics is not optimized in algorithm of Sullivan and Ma. This chapter is trying to build a more comprehensive and reasonable HVS model, then optimize the rate distortion model of H.264 by adjusting λ_{mode} and finally improve the subjective video quality.

124.2 Analysis of Assessment Effect on Video Quality Produced by Human Visual System

Analysis of temporal features. According to Human Visual System, human visual perceptual are selective when they are observing video pictures [5]. More attention will be paid to motive objects in pictures and distortions are less noticeable because of strongly covered effect in complicated texture regions. So how to distinguish motive regions from a picture accurately is important pre-condition. We use a method of motion degree detection based on the entropy of Histogram Of Differences (HOD) between two adjacent frames which was proposed by Jungwoo in [6].

Firstly, the current frame is divided into macroblocks (MB) of 16×16 pixels and for each MB, we would calculate the HOD between the current MB and its corresponding position in the previous frame by:

$$H_{i,n} = h(e_{i,n} - e_{i,n-1}) \quad (124.1)$$

where $e_{i,n-1}$ is the i th MB of the n th frame and the function of $h()$ is HOD function.

Secondly, a cumulative probability distribution function (CPDF) is calculated by:

$$CPDF = \frac{1}{N} \sum_{q=1}^r H_{q,i,n} \quad (124.2)$$

where N and q represent the number of pixels and the bins in the HOD respectively.

When the CPDF is known, entropy of MD will be ascertained in the following step:

$$E_{i,n} = - \left(H_{q,i,n} / \sum_{r=1}^q CPDF_{r,i,n} \right) \times \log_2 \left(H_{q,i,n} / \sum_{r=1}^q CPDF_{r,i,n} \right) \quad (124.3)$$

The threshold T_{motion} adjudging the state of a given MB as static or moving will be obtained by a simple method. In every frame, a local threshold T_{motion} will be found in each window of 5×5 MB. This local threshold T_{motion} is equal to the average of five MBs entropies in this window. Therefore, only MBs entropies greater than $T_{5 \times 5}$ are selected. All of global threshold T_{motion} of MBs are calculated as follows:

$$T_{\text{Motion}} = \text{mean}(E_{i,n} > T_{5 \times 5}) \quad (124.4)$$

Finally, a given MB with entropy greater than T_{motion} is considered as a moving one

Spatial characteristics classification analysis. A picture is divided into three types of regions: smooth, random texture and fixed texture. Human attention are more likely to be drawn by texture parts in a picture than that smooth one. Meanwhile, the random texture regions are less noticeable than that fixed one. We would distinguish above three types of regions as following method based on the picture features analyzed by Rajashekar et al. and Tang et al. apart in [7] and [8].

Firstly, the texture regions will be separated from a picture. In this process, both the average edge intensity and the edge pixel density of an MB are considered. The edge image will be obtained due to Sobel operator. The edge intensity of $B_n(X, Y)$ is then calculated by:

$$\bar{I}_n(u, v) = \frac{1}{N} \sum_{(i,j) \in B_n(u,v)} EI_n(i, j) \quad (124.5)$$

where $EI_n(i, j)$ notes the intensity value of the pixel at site (i, j) in edge image n . The edge pixel density of $B_n(U, V)$ is calculated by:

$$\bar{ED}_n(u, v) = \frac{1}{N} \sum_{(i,j) \in B_n(u,v)} EP_n(i, j) \quad (124.6)$$

where the value of $EP_n(i, j)$ is 1 when $EI_n(i, j)$ is larger than the threshold of edge pixel (experimental value is 50), otherwise is 0. When $EP_n(i, j)$ is 1, it is looked as an edge pixel. The threshold T_{texture} for texture judging is defined by:

$$T_{\text{texture}} = \beta \sum_{B_n(u,v) \in S_n} \bar{I}_n(u, v) + \gamma \sum_{B_n(u,v) \in S_n} \bar{ED}_n(u, v) \quad (124.7)$$

where β and γ are weighting factors and their experimental value are set as 0.6 and 1.0 respectively. An MB with the sum of edge intensity and edge density larger than T_{texture} is classified as a texture one. Otherwise, it is a smooth one.

Once the MB has been determined as a texture one, the following step is to decide whether it is a random texture or a fixed one. Since random texture regions

Table 124.1 The perceptual importance levels

Temporal characteristics	Static			Moving	
Spatial characteristics	Random texture	Smooth	Fixed texture	Random texture	Fixed texture/ smooth
Perceptual importance level	1	2	3	2	4

usually have lots of small edges with random directions, it is rational to think that the edge intensities in these regions distribute more uniformly than those in fixed texture regions. Thereby, the 16×16 MB is divided into small 4×4 blocks. The $b_n(u, v)$ indicates the block at location (u, v) of S_n and the edge intensity of $b_n(u, v)$ is calculated as:

$$Eb_n(u, v) = \frac{1}{N} \sum_{(i,j) \in b_n(u,v)} EI_n(i, j) \tag{124.8}$$

The normalized edge deviation $Md_n(U, V)$ is calculated by:

$$Md_n(U, V) = \sum_{b_n(u,v) \in B_n(U, V)} |Eb_n(u, v) - \overline{Eb}_n(U, V)| / \overline{Eb}_n(U, V) \tag{124.9}$$

where $\overline{Eb}_n(U, V)$ means the mean block edge intensity of $B_n(U, V)$.

Let $N_n(U, V)$ denote the 4×4 blocks in a 8×8 range neighboring $B_n(U, V)$.

The normalized edge deviation of $N_n(U, V)$ is calculated by:

$$Nd_n(U, V) = \sum_{b_n(u,v) \in N_n(U, V)} |Eb_n(u, v) - \overline{Nb}_n(U, V)| / \overline{Nb}_n(U, V) \tag{124.10}$$

Decision scheme: if both $Md_n(U, V)$ and $Nd_n(U, V)$ are small the mean values of the texture MBs in S_n , it is classified as a random texture MB. Otherwise, it is a fixed texture one.

Perceptual importance map based on spacial and temporal. According to the analysis of spacial and temporal features, the perceptual importance of MBs is assigned as shown in Table 124.1.

HVS have high sensitivity to the MBs slowly moving regions of smooth/fixed texture. This slowly moving regions gain more visual importance in this chapter. On the contrary, the static regions with random texture are assigned to lowest level, since distortions in these regions are less noticeable than those in other moving regions.

124.3 Adjust Rule Enactment of Rate-Controls Parameter

Ascertainment of λ'_{mode} . To reach the target rate as near as possible, λ'_{mode} is adjusted by:

$$\lambda'_{\text{mode}}(i) = \lambda_{\text{mode}}(i) \times \eta_t(i) = 0.85 \times 2^{OP_i/3} \times \eta_t(i) \quad (124.11)$$

where the value of $\eta_t(I)$ are set as 4.0, 2.0, 1.0 and 0.7 respectively according to perceptual importance level.

Adjusting of QP (Quantization Parameter). On the basis of perceptual importance and rate adjusting rule, we have adjusted QP as follows:

$$QP'_i = QP_i + \Delta QP_1(n) + \Delta QP_2(t) \quad (124.12)$$

where $\Delta QP_1(n)$ ($n = 1, 2, 3, 4$) are 2, 1, 0 and -2 respectively, $\Delta QP_2(t)$ reflects increment of code rate and its value is confirmed as follows:

$$\Delta QP_2(t) = \begin{cases} -1, & t \leq \gamma_1 \\ 0, & \gamma_1 < t \leq \gamma_2 \\ 1, & \text{others} \end{cases} \quad (124.13)$$

124.4 Experimental Results

The proposed scheme has been implemented on the reference software JM 12.2 of H.264/AVC to demonstrate the effects. Primary profile has been used in experiment. All of modules keep the same except for rate control. The test sequences with CIF 4:2:0 format and 30 frames per second include Claire, Hall_monitor and highway. QPs are set for 28, 32 and 36 respectively. The format of GOP is IPPP and RDO is on highly complex mode. Hadamard transform and CABAC entropy coding mode are selected. Reference frame and searching range are 2 and 16 respectively.

To assess the subjective quality, we adopt the Double Stimulus Continuous Quality Scale (DSCQS) method. Ten observers (most students with image processing knowledge) have been asked to evaluate video quality under normalized conditions and vote with 5 grades: excellent (5), good (4), fair (3), poor (2) and bad (1).

The target bit rates, PSNR and valuation grades etc of mobile, Claire, Hall_monitor and highway sequences are listed in Table 124.2. It shows that the proposed method with different target bit rates gain almost the same mean opinion scores (MOS) as those of Yu and JM. The proposed algorithm even gain slightly higher scores in the case of low bit rates. This indicates our scheme is more propitious to enhance the reconstructed visual quality under low bit rates condition. From the experiment results, we can see that both Yu and the proposed

Table 124.2 Coding performance comparisons under different target bit-rate

Sequence	QP	Algorithm	MOS	PSNR (dB)	Rate (Kbps)	Rate-saving (%)
Claire	28	JM10.2	4.1	35.53	1,301	0
		YU	4.1	35.09	1,274	2.1
		Ours	4.1	33.98	1,158	11.0
	32	JM10.2	3.8	32.61	621	0
		YU	3.8	31.27	603	2.9
		Ours	3.9	30.03	536	13.7
	36	JM10.2	3.4	29.82	261	0
		YU	3.5	28.69	252	3.4
		Ours	3.5	28.13	223	14.6
Hall_monitor	28	JM10.2	4.0	36.37	1,394	0
		YU	4.0	35.58	1,366	2
		Ours	4.0	34.65	1,225	4.9
	32	JM10.2	3.7	32.76	693	0
		YU	3.7	31.83	672	2.7
		Ours	3.8	31.74	603	12.9
	36	JM10.2	3.3	29.16	321	0
		YU	3.4	28.43	316	1.6
		Ours	3.5	28.14	286	10.9
Highway	28	JM10.2	4.1	35.37	1,678	0
		YU	4.2	34.76	1,646	1.9
		Ours	4.2	33.80	1,513	9.8
	32	JM10.2	3.9	32.64	834	0
		YU	3.9	31.93	812	2.6
		Ours	4.0	31.35	751	9.9
	36	JM10.2	3.5	28.14	412	0
		YU	3.5	27.65	399	3.2
		Ours	3.6	27.53	368	10.7

algorithm results in PSNR degradations. As compared to JM, the PSNR degradation of proposed algorithm is about 1 dB. This is understandable because fewer bits are allocated to the less perceptual importance regions by adjusting the Lagrange coefficient in RDO process. But it do not lead to visual quality degradation since distortions in those regions are less noticeable. This has been well proved by the comparative results of MOS in Table 124.2.

124.5 Conclusion

We analyzed the spacial and temporal features of neighboring frames of video pictures on the basis of HVS characteristics. Video pictures were disparted two types of moving and static in temporal field and three types of smooth, random and fixed textures in spacial field. Following, the perceptual importance of images were assigned as four grades based on integrating spacial and temporal features. In the

last, the Lagrangian coefficient λ_{mode} and Quantization step length QP of H.264/AVC rate control were adjusted according to the grades. Experimental results show our algorithm can efficiently reduce bit rates without visual quality degradation and gain better subjective video quality under the condition of rate-restricted.

References

1. Tsai CJ, Tang CW, Chen CH, Yu YH et al (2004) Adaptive rate-distortion optimization using perceptual hints [C]. In: IEEE international conference on multimedia and expo, vol 1, pp 667–670
2. Yu HT, Pan F, Lin ZP, Sun Y et al (2005) A perceptual bit allocation scheme for H.264 [C]. In: IEEE international conference on multimedia and expo, Amsterdam, July 2005, pp 4–7
3. Sullivan G, Wiegand T, Lim KP et al (2003) Joint model reference encoding methods and decoding concealment methods [C], section 2.6: rate control JVT-I049, San Diego, pp 24–31
4. Ma SW, Gao W, Wu F, et al (2003) Rate control for JVT video coding scheme with HRD considerations [C]. In: Proceedings of IEEE ICIP 2003. IEEE Signal Processing Society, Barcelona, pp 793–796
5. Itti L, Koch C, Niebur E et al (1998) A model of saliency-based visual attention for rapid scene analysis [J]. *IEEE Trans Pattern Anal Mach Intell* 20(11):1254–1259
6. Jungwoo L (1999) A fast frame type selection technique for very low bit rates coding using MPEG-1 [J]. *Real-time Imaging J* 5(2):83–94
7. Rajashekar U, Cormack LK, Bovik AC et al (2003) Image features that draw fixations [J]. *ICIP* 3: 13–16
8. Tang CW (2006) Spatiotemporal visual considerations for video coding [J]. *IEEE Trans Multimedia* 9(2):231–238

Chapter 125

Contour Lines Extraction from Color Scanned Topographical Maps with Improved Snake Algorithm

Huali Zheng and ZhouWei Guo

Abstract Contour lines are the most important features to characterize three dimensional terrain on color scanned topographical map sheet. Based on the analysis of characteristics of contour lines, a novel snake model for automated tracking is proposed. With the variable force control and gradient vector flow, contour lines can be vectorized automatically. The tracking algorithm is significant for its flexible form and combination of global field information with local tracking. Furthermore, the algorithm has general reference to other kinds of graphical vectorization.

Keywords Contour lines • Topographical map • Snake algorithm

125.1 Introduction

Automatic vectorization of archival maps is a promising and strenuous problem in image processing and pattern recognition domain [1]. As the majority of data acquisition in topological maps, the automatic extraction and interpretation of contour lines plays a fundamental role for Geographic Information System (GIS) purposes. Traditional vectorization methods can be classified as various thinning-based methods [2], pixel-tracking based methods [4, 5] and contour-based

H. Zheng (✉)

Institute of Special Electromechanical Technology, Beijing, China
e-mail: easylife8@163.com

Z. Guo

Artillery Institute of PLA, Hefei, China
e-mail: guo_zw@126.com

methods [5, 6]. However, these methods have difficulties in the interpretation of contour lines. As a result of the performance of scanners and the original map, plenty of transitional color and color distortion exists in scanned map images [7], the phenomenon of broken or touching contour lines appear inevitably. Intersections and coverage of different elements each other generate more break points and discontinuous contour lines. An efficient algorithm is need for digitization of contour lines in topological images.

125.2 Snake Model

Snake model is a classical energy-based algorithm in deformable models. It was originally proposed by Kass [8] and has been used in a variety of computer vision applications recently. This framework matches a deformable model to images by means of energy minimization and therefore it exhibits dynamic behavior. The conventional energy function in the model is defined:

$$E_{snake} = \int_l |E_{int}(v(s)) + E_{ext}(v(s))| ds \quad (125.1)$$

where $v(s) = (x(s), y(s))$ is deformable curve segment. E_{int} and E_{ext} represent the internal energy and the external energy of the contour respectively.

125.3 Improved Algorithm of Tracking

Traditional snake models are always applied in the detection of closed contours of objects in images. However, it still has some weakness. First, a set of initial control points is demanded before the implementation and must be close to the true object. Second, the model cannot handle well the case of high curvature or turning point in graphic images and likely converge to wrong result. To overcome the problems and make it have the property of tracking graphical curves as well, a novel improved model for tracking is presented.

The main idea of the improved snake algorithm of tracking simulates the function of human vision while tracking contours (Fig. 125.1). Generically speaking, the global information comes from local analysis but the interpretation of local information depend on global direction. It seems to be a basal inconsistency in computer vision, The global information is obtained by statistical computation.

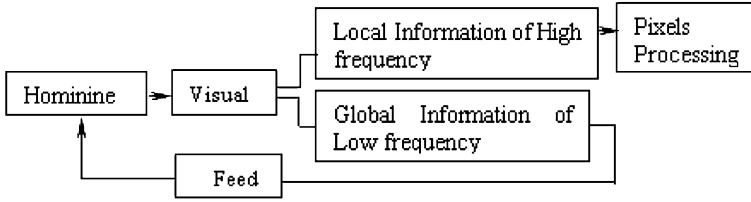


Fig. 125.1 Information flow of the improved snake algorithm

125.3.1 Seed Segments Detection

As the initial form of snake algorithm, a seed segment is composed of several projecting matrix. N is the number controllable points in it. h denotes the average distance between them. A projecting matrix is defined by

$$\mathbf{K}_{(w,h,x,y,a)} = \begin{bmatrix} k_{11} & k_{12} & \cdots & \cdots & k_{1h} \\ k_{21} & k_{22} & \cdots & \cdots & k_{2h} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ k_{w1} & k_{w2} & \cdots & \cdots & k_{wh} \end{bmatrix} \tag{125.2}$$

$$k_{ij} = I[(j - 1) \cos a - \left(\frac{w + 1}{2} - i\right) \sin a + x, (j - 1) \sin a + \left(\frac{w + 1}{2} - i\right) \cos a + y] \quad (1 \leq i \leq w, 1 \leq j \leq h) \tag{125.3}$$

$I(i, j)$ is the membership of brown color in pixel (i, j) that can be obtained from the algorithm presented in Ref. [7]. \mathbf{K} is a matrix with slant angle a and the coordinate of its midpoint on the bottom line is (x, y) . The norm of \mathbf{K} is defined:

$$\|\mathbf{K}\| = \mathbf{K} \otimes \mathbf{M} \tag{125.4}$$

\mathbf{M} is the weighting matrix of the pixels in \mathbf{K} . The process of seed segment detecting is described as follows:

- (1) Let the width of the contour lines be w , h is the distance of controllable points. Scans the entire image using a mesh. On encountering a brown pixel, searches the connected pixel area for a projecting matrix satisfied: $\|\mathbf{K}\| \geq threshold$
- (2) Continue searching for the next new projecting matrix from the point $A(x + (h - 1) \cos a, y + (h - 1) \sin a)$ with the slant angle $(a - \pi/2, a + \pi/2)$. If $\mathbf{K}_n = \underset{a+\pi/2 > ; a > ; a+\pi/2}{Max} \|\mathbf{K}_a\| > ; threshold$, then the Mark the point A as controllible point, else goes to (1). for detection attempts.
- (3) Find N consecutive controllible points along the curve and label one of two extreme points as starting point.

125.3.2 Internal Energy of Snake Algorithm

Minimizing an energy function is a typical way to for snake model. The internal energy in the traditional snake model is dominated by two kinds of forces: continue constraint and smooth constraint. The consecutive constraint is the squared distance between the neighboring controllable points in digital images. During the process of energy minimum, it often trend to shorten the space of snake model and will shrink different points into one. Since the image energy and projecting matrix have ensured the continuity. The index of continual constraint in the tracking model α is zero. The principle feature of snake model explored in this chapter is the tracking. The efficiency of the model depends on the accordance of the model with the contour lines in the image. So the internal energy mainly comes from image. Since the tracking model should cross small gaps in the maps, on the other hand it must keep up the continuity of curves, the internal energy should have the geodesic and statistical property. Unlike the traditional snake model that compute the internal energy around controllable points only, the internal energy of tracking model refined:

$$E_{\text{int}} = -\gamma \sum_{m=1}^N \|K_m\|. \quad (125.5)$$

125.3.3 External Energy of Snake Algorithm

The contour lines in the topographic map are winding and have high curvature sometimes. As a tracking algorithm, the smoothness of the model behaves as its accordance with the original curves. So it is not practical to have fixed smooth constraint while tracking. The external force must be variable with the contour orientation analysis. Orientation analysis is to make using of global information to direct local tracking.

Firstly, with color segmentation the map of brown color is obtained by selecting a threshold. Despite of lots of broken line and noise appearing in the image, the direction of contours could still be recognized. To degrade the influence of noise, 2D convolution operation is needed:

$$S(x, y, \sigma) = I(x, y) \otimes G(x, y, \sigma) \quad (125.6)$$

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} I(m, n) \exp\left[-\frac{(x-m)^2 + (y-n)^2}{2\sigma^2}\right] dmdn \quad (125.7)$$

$I(x, y)$ denotes the original image, $G(x, y, \sigma)$ is Gauss function with distribution variance σ . When the distance between two neighbor contours is smaller than σ , they will be combined. Let σ be of half of the width of contour generally.

The gradient of pixels in the map is computed.

$$f_x = I(x, y) \otimes H_x(x, y) f_y = I(x, y) \otimes V_y(x, y) \quad (125.8)$$

$H_x(x, y)$ and $V_y(x, y)$ are the horizontal and vertical gradient kernels respectively. They could be Prewitt or Sobel kernels. The gradients of a field cannot be averaged in some local neighborhood directly since opposite gradient vectors will then cancel each other. This is caused by the fact that local line structures remain unchanged when rotated over 180° . So we square the gradient and then get of average value.

$$\begin{aligned} (f_x + j \cdot f_y)^2 &= (f_x^2 + f_y^2) + j \cdot (2f_x f_y) \\ &= f_{sx} + j \cdot f_{sy} \end{aligned} \quad (125.9)$$

The average gradient of A is obtained:

$$\begin{bmatrix} \overline{f_{sx}} \\ \overline{f_{sy}} \end{bmatrix} = \begin{bmatrix} \sum_A f_{sx} \\ \sum_A f_{sy} \end{bmatrix} = \begin{bmatrix} \sum_A (f_x^2 - f_y^2) \\ \sum_A 2f_x f_y \end{bmatrix} = \begin{bmatrix} F_{xx} - F_{yy} \\ 2F_{xy} \end{bmatrix} \quad (125.10)$$

The flow orientation is defined as:

$$\theta = \begin{cases} \frac{1}{2} \angle(2F_{xy}, F_{xx} - F_{yy}) + \frac{1}{2} \pi & \angle(2F_{xy}, F_{xx} - F_{yy}) > 0 \\ \frac{1}{2} \angle(2F_{xy}, F_{xx} - F_{yy}) - \frac{1}{2} \pi & \angle(2F_{xy}, F_{xx} - F_{yy}) \leq 0 \end{cases} \quad (125.11)$$

$$\angle(x, y) = \begin{cases} \tan^{-1}(y/x) & x > 0 \\ \tan^{-1}(y/x) + \pi & \text{for } x < 0 \cap y > 0 \\ \tan^{-1}(y/x) - \pi & x > 0 \cap y < 0 \end{cases} \quad (125.12)$$

To improve the precision of the direction calculation, the area calculated in the method is adaptive and alterable. The coverage of field must ensure the quantity of gradients in it.

The new external energy is defined as follows:

$$\xi_{ij} = \begin{cases} \left| \theta_i - \frac{\vec{v}_j}{|\vec{v}_i|} \right| & \pi/2 > \left| \theta_i - \frac{\vec{v}_j}{|\vec{v}_i|} \right| \geq 0 \\ \left| \left| \theta_i - \frac{\vec{v}_j}{|\vec{v}_i|} \right| - \pi \right| & 3\pi/2 > \left| \theta_i - \frac{\vec{v}_j}{|\vec{v}_i|} \right| \geq \pi/2 \end{cases} \quad (125.13)$$

$$E_{ext} = \beta(\xi_{i,i} + \xi_{i,i+1})/2 \quad (125.14)$$

\vec{v}_i denotes the vector from the $(i-1)$ th point to i th point.

Fig. 125.2 Original map



125.3.4 Process of the Snake Algorithm

The implementation of improved snake algorithm of tracking is described as follows:

Step 1: Initialize threshold, β , γ , h , w , N . Find the initial seed segments in the map.

Step 2: Select a circle whose center is the starting point of seed segment with the parameter of h . m is the number of ridge points intersecting with the circle and select one of them as the ending point of seed segment. The method of getting ridge points see [9].

Step 3: If no ridge point detected, $h = 3/2 \cdot h$, go back to 2. If there are several ridge points, select one by one.

Step 4: For each of the controllable point of seed segment adjust its position within the projecting matrix and move to the new position with less energy.

Step 5: Compute the general energy and compare with the former one. Stop the loop if there is little change. Else go back to step 4.

Step 6: Compare the model energy E_i with different ridge points and select the minimum one E_{min} . If $E_{min} < threshold$, select the ridge point as the new starting point. Erase the ending point and its projecting matrix, then go back to step 2.

Step 7: If several ridge points $E_i \leq threshold$, it is identified that the tracking encounters cohering lines or characters and then execute the module of disposal of broken or touching curves or alert.

Step 8: If $E_i > threshold$ for all ridge points, search for controllable point within the domain along the original direction and annex it. Go back to step 1.

Step 9: Execute the module of disposal of broken or touching curves for the residue break points on the initial result of vectorization.

Fig. 125.3 Segments detection

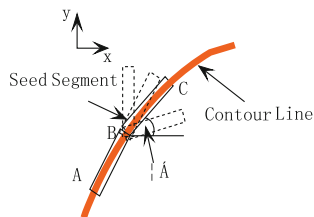


Fig. 125.4 Vectorization results. **a** Traditional tracking algorithm. **b** Improved Snake algorithm

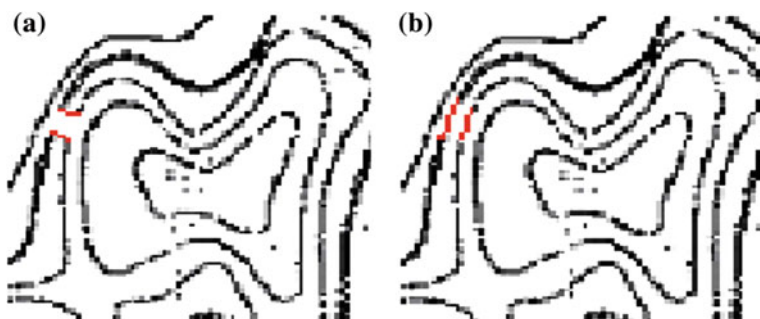


Fig. 125.5 Comparison of tracking algorithm. **a** Traditional tracking algorithm. **b** Improved snake algorithm

125.4 Experiments and Conclusion

To demonstrate the power of our proposed algorithm, we have applied the improved snake algorithm of tracking to different topological maps of large scale and have satisfying results. Figure 125.2 is a part of topographic map of scale 1:500,000. Figure 125.3 represents seed segments detection. The final result of vectorization of contour lines is shown in Fig. 125.4 with parameter settings: $w = 3$, $h = 10$. The comparison of traditional tracking algorithm and improved snake algorithm is shown in Fig. 125.5. It can be seen that the improve snake algorithm of tracking behaves accurately. Its flexible form and global field information direction ensure its advantages over traditional tracking algorithms.

References

1. Nagy G (2000) Twenty years of document image analysis in PAMI. *IEEE Trans Pattern Anal Mach Intell* 22(1):38–62
2. Qian J (2006) A study of contour lines extraction from color scanned topographical maps. *J Image Graph* 11(6):1443–1448
3. Dori D, Liu WY (1999) Sparse pixel vectorization: an algorithm and its performance evaluation. *IEEE Trans Pattern Anal Mach Intell* 21(3):202–215
4. Pan W (2009) A new threshold segmentation of identification and pick-up for contour line in paper military map. *Comput Technol Autom* 24(2):35–37
5. Mun SL, Yatim SM (2004) A-extracting contour lines from scanned topographic maps. In: *Proceedings of the international conference on computer graphics, imaging and visualization, Aachen*, pp 187–192
6. Li S (2007) Research on contour line automatic extraction from scanned topographic map. *Bulletin of Surveying and Mapping* (1) pp 65–67
7. Zheng HL, Zhou XZ (2003) Research and implementation of automatic color segmentation algorithm for scanned color maps. *J Comput Aided Des Comput Graph* 15(1):29–33
8. Kass M, Witkin A (1987) Snakes: active contour model. In: *Proceedings of first international conference on computer vision, London*, pp 259–268
9. Zheng HL, Zhou XZ (2003) Automatic color segmentation of topographic maps based on the combination of spatial relation information and color information. *J Image Graph* 8:353–340

Chapter 126

The Aberration Characteristics of Wave-Front Coding System for Extending the Depth of Field

Zhang Rong-fu, Wang Liang-liang and Zhou Jinbo

Abstract Wave-front coding involves the insertion of an aspheric phase mask to the pupil plane of an optical imaging system so as to achieve a final image with extended depth of field, when combines with decoding of intermediate image by digital processing. By using method of the stationary phase, an approximate expression of modulation transfer function (MTF) for wave-front coding system with cubic phase mask is derived while multiple aberrations are taken into account, and aberration characteristics are comprehensively analyzed. The analysis results show that spherical aberration will bring MTF degradation and vibration, but existence of defocus, astigmatism, or field curvature will improve such condition.

Keywords Wave-front coding · Cubic phase plate · MTF

126.1 Introduction

Wave-front coding [1] involves the insertion of an aspheric phase mask to the pupil plane of an optical imaging system, an intermediate image which is insensitive to defocus is gained. Digital processing is used to restore intermediate focus-invariant image with a very large depth of field.

Z. Rong-fu (✉) · W. Liang-liang · Z. Jinbo
School of Optical-Electrical and Computer Engineering,
University of Shanghai for Science and Technology, Shanghai 200093, China
e-mail: zhangrongfu@usst.edu.cn

W. Liang-liang
e-mail: science140311@126.com

Z. Jinbo
e-mail: zjbusst@163.com

While the depth of field is extended, a number of aberrations, such as spherical aberration, coma, astigmatism, chromatic aberration, field curvature are controlled [2]. Defocus suffering from temperature variation and fabrication tolerance is also minimized. Previously, the analysis of aberration characteristics based on single aberration is widely carried out [3, 4].

By using method of stationary phase [5–7], an approximate expression of modulation transfer function (MTF) for wave-front coding system with cubic phase mask is derived while multiple aberrations are taken into account, and aberration characteristics are comprehensively analyzed.

126.2 MTF with Aberration

126.2.1 Wave Aberration

Wave aberration means optical path difference (OPD) between Gaussian reference sphere and image capture plane. The one-dimensional (1-D) expression [8] in orthogonal coordinate is given by

$$w(x) = w_{11}x + (w_{20} + w_{22})x^2 + w_{31}x^3 + w_{40}x^4. \quad (126.1)$$

where x denotes the normalized coordinate of pupil, the meaning of wave aberration coefficients [9] are as follows: W_{31} represents the sum of vertical defocus, the primary and secondary distortion; $W_{20} + W_{22}$ represents the sum of axial defocus, primary field curvature and primary astigmatism; W_{31} represents the sum of primary coma and secondary coma produced by field; W_{40} represents the sum of primary spherical aberration and secondary spherical aberration produced by field, respectively.

Phase difference φ can be shown by the product of wave number k and wave aberration

$$\varphi = kw(x) = \frac{2\pi}{\lambda} w(x) = \frac{2\pi}{\lambda} [w_{11}x + (w_{20} + w_{22})x^2 + w_{31}x^3 + w_{40}x^4]. \quad (126.2)$$

Pupil function with aberration can be considered as generalized pupil function, according to the difference of wave aberration in Eq. 126.2, the effect of aberration to MTF can be analyzed by using generalized pupil function.

126.2.2 The Derivation of MTF With Aberration

The pupil function of imaging system with cubic phase mask along one dimension [1] is given by

$$q(x) = \begin{cases} \frac{1}{\sqrt{2}} \exp(j\alpha x^3) & |x| \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (126.3)$$

where α is the modulation coefficient of cubic phase and $\alpha \geq 20$. The generalized pupil function suffering from defocus (or astigmatism, field curvature), coma and spherical aberration [10] can be described by

$$Q(x) = \begin{cases} \frac{1}{\sqrt{2}} \exp[j(\alpha x^3 + \varphi_1 x^2 + \varphi_2 x^3 + \varphi_3 x^4)] & |x| \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (126.4)$$

where φ_1 , φ_2 , φ_3 represent the coefficient of defocus (or astigmatism, field curvature), coma and spherical aberration respectively, and

$$\varphi_1 = \frac{2\pi}{\lambda} W_{20}, \varphi_2 = \frac{2\pi}{\lambda} W_{31}, \varphi_3 = \frac{2\pi}{\lambda} W_{40}.$$

According to the derivation in [10], we drive an approximate expression of MTF with various aberrations for the wave-front coding system by using method of stationary phase. The MTF is

$$|H(u)| = \begin{cases} \sqrt{\frac{\pi}{2} \left[\frac{1}{|\mu''(x_{01})|} + \frac{1}{|\mu''(x_{02})|} \right] + \frac{1}{2} \sqrt{\frac{2\pi}{|\mu''(x_{01})|}} \sqrt{\frac{2\pi}{|\mu''(x_{02})|}} \cos \left[\frac{\pi}{2} + \mu(x_{01}) - \mu(x_{02}) \right]} & 0 < u \leq 2 \\ 1 & u = 0 \end{cases} \quad (126.5)$$

where u is the normalized spatial frequency,

$$x_{01} = \frac{-3(\alpha + \varphi_2) + \sqrt{9(\alpha + \varphi_2)^2 - 24\varphi_1\varphi_2 - 12\varphi_3^2 u^2}}{12\varphi_3},$$

$$x_{02} = \frac{-3(\alpha + \varphi_2) - \sqrt{9(\alpha + \varphi_2)^2 - 24\varphi_1\varphi_2 - 12\varphi_3^2 u^2}}{12\varphi_3},$$

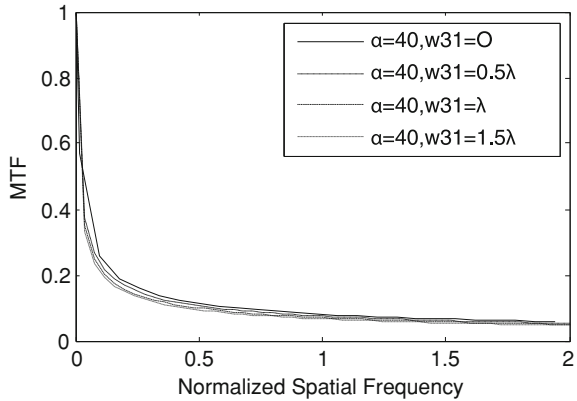
$$\mu(x) = (\alpha + \varphi_2) \left(3ux^2 + \frac{u^3}{4} \right) + 2\varphi_1 xu + \varphi_3 (4ux^3 + u^3 x),$$

$$\mu''(x) = 24\varphi_3 x + 6(\alpha + \varphi_2).$$

126.3 Aberration Characteristics

MTF is one of the most effective tools to characterize the imaging quality of an optical system. The imaging quality of the cubic phase coding system with single aberration or multiple aberrations can be analyzed by setting corresponding aberration coefficient in Eq. 126.4.

Fig. 126.1 Variation of MTF suffering from coma W_{31}



126.3.1 MTF with Coma

A system with coma is produced if the coefficient of defocus (or astigmatism, field curvature) φ_1 and spherical aberration φ_3 equal to 0, and meantime φ_2 does not equal to 0 in Eq. 126.4. Compared with the system suffering from defocus [1], the increasing of coma coefficient gives MTF a falling trend, as showed in Fig. 126.1, which illustrates that the ability of controlling coma for the system is slightly weakened.

126.3.2 MTF with Spherical Aberration

A system with spherical aberration is generated if φ_1 and φ_2 equal to 0, and meantime φ_3 does not equal to 0 in Eq. 126.4. Concussion appears even in low frequency domain and the magnitude of MTF decreases, which indicates that the ability of controlling spherical aberration for the system is greatly weakened. In addition, invariant modulation coefficient α and increasing spherical aberration result in a slight growing trend of concussion frequency.

126.3.3 MTF with Coma and Spherical Aberration

A system with coma and spherical aberration is created if φ_1 equals to 0, and meantime φ_2 and φ_3 do not equal to 0 in Eq. 126.4. Figure 126.3 shows that concussion of MTF is weighted because of the existence of coma, compared with Fig. 126.2.

Fig. 126.2 Variation of MTF suffering from spherical aberration W_{40}

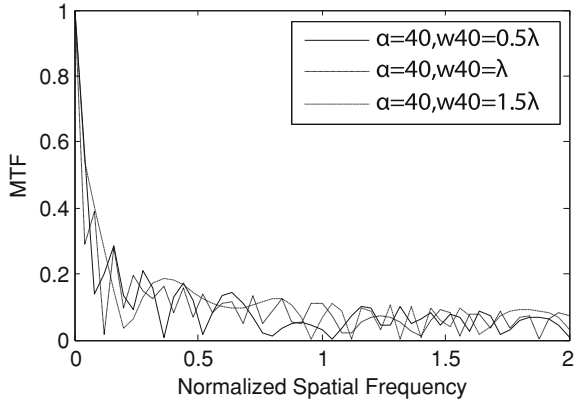
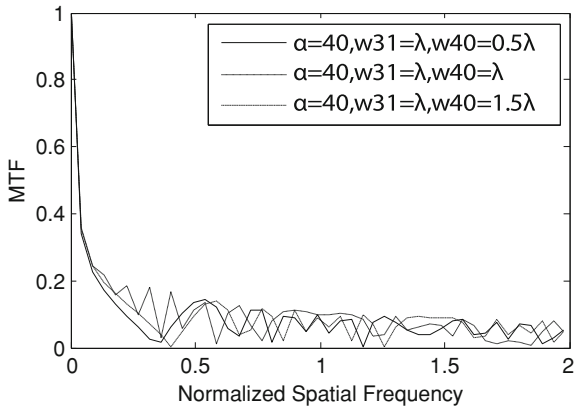


Fig. 126.3 Variation of MTF suffering from coma W_{31} and spherical aberration W_{40}

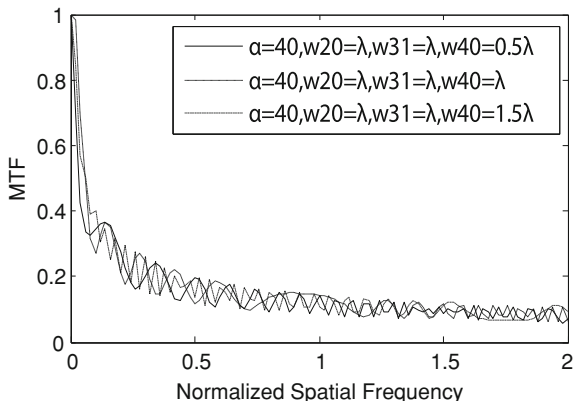


126.3.4 MTF with Defocus, Coma and Spherical Aberration

A system with defocus, coma and spherical aberration is yielded if φ_1 , φ_2 and φ_3 don't equal to 0 in Eq. 126.4. While spherical aberration increases, concussion of MTF is slightly aggravated, but it is seen from Fig. 126.4 that the magnitude of MTF becomes largely rising in the whole frequency domain, compared with Fig. 126.3.

Comprehensive comparison of Figs. 126.2, 126.3 and 126.4 indicates that a system with spherical aberration alone will cause concussion and degradation for MTF, which will be slightly aggravated for a system with coma and spherical aberration. However, to a great extent such situation will be improved for a system with defocus, coma and spherical aberration.

Fig. 126.4 Variation of MTF suffering from defocus W_{20} (astigmatism W_{22}), coma W_{31} and spherical aberration W_{40}



126.4 Conclusion

We have presented an approximate expression of MTF for wave-front coding system with cubic phase mask by using method of stationary phase when multiple aberrations are taken into account, and aberration characteristics are comprehensively analyzed. The conclusions are as follows: (1) The ability of controlling defocus (or astigmatism, field curvature) for the cubic wave-front coding system is the best, followed by coma, finally for spherical aberration; (2) Concussion and attenuation of MTF result from spherical aberration, but such condition can be largely improved if defocus (or astigmatism, astigmatism) exists. Therefore, when correcting aberration for cubic wave-front coding system, the following should be paid special attention to: (1) When taking on-axis monochromatic meridian or arc imaging into account merely for the system, the quality of out-of-focus image is better than that on focal plane as there are not other aberrations except defocus and spherical aberration;(2) Astigmatism or field curvature rather than coma can greatly contribute to the improvement of degraded MTF resulted from spherical aberration, while taking off-axis monochromatic meridian imaging into account merely for the system.

References

1. Dowski ER, Cathey WT (1995) Extended depth of field through wavefront coding. *Appl Opt* 34:1859
2. Dowski ER Jr, Cathey WT, Braddum SC (1996) Aberration invariant optical/Digital incoherent systems. *Opt Rev* 3:429
3. Castro A, Ojeda-Castaneda J (2005) Increased depth of field with phase-only filters: ambiguity function. *Proc SPIE* 5827:1
4. Wach HB, Dowski ER, Cathey WT (1998) *Appl Opt* 37:5359
5. Cathey WT, Dowski ER (2002) New paradigm for imaging systems. *Appl Opt* 41:6080

6. Hua Lei, Hua-jun Feng, Xiao-ping Tao et al (2006) Appl Opt 45:7255
7. Mezouari S, Harvey AR (2002) Phase pupil functions for reduction of defocus and spherical aberrations. Proc SPIE 4768:21
8. Information on http://www.eefocus.com/html/dict_129551_dd61d74bed43944fd3edbc1f24f38147.html
9. Wang Z-J, Gu P (2007) In: Handbook of practical optical technology, Mechanical Industry Press, Beijing
10. Chen Y-P, Zhang W-Z, Zhao T-Y (2007) Opt Exp 15:1543

Part XII
Green Computing

Chapter 127

Two Improved Nearest Neighbor Search Algorithms for SPH and Their Parallelization

Zheng Dequn, Wu Pin, Shang Weilie, Cao Xiaopeng, Zhang Xia and Chen Wei

Abstract Efficient search algorithm is very important for Smoothed Particle Hydrodynamics (SPH). In this Chapter, the entire search process is divided into pre-search process and main searching process. On this basis, two improved search algorithms based on sliding windows and background grid are proposed. Application was made for three-dimensional shear cavity flow. Results show that the sequential program can greatly reduce the computational cost while maintaining the same accuracy. Furthermore, the improved search algorithms have been implemented on a cluster, the experimental results demonstrate that the parallel program has a higher speedup.

Keywords SPH · Nearest neighbor search algorithm · Parallel computing · MPI

127.1 Introduction

The basic Smoothed Particle Hydrodynamics (SPH) method was created by Lucy (1977) and Gingold and Monaghan (1977) in order to study fission in rotating stars [1]. Because of its Lagrangian nature, SPH presents some clear advantages over more traditional grid-based methods for calculations of particle interactions. Most importantly, fluid advection is accomplished without difficulty in SPH, since the particles simply follow their trajectories in the flow [2]. SPH is also very computationally efficient, since it concentrates the numerical elements (particles) where the fluid is at all times, not wasting any resources on empty regions of space [3].

Z. Dequn · W. Pin (✉) · S. Weilie · C. Xiaopeng · Z. Xia · C. Wei
Department of Computer Science and Engineering, Shanghai University,
Shanghai, China
e-mail: wupin@shu.edu.cn, dequn6688@163.com

However, SPH as an interpolation algorithm, needs to search for interacting particles, then calculate the kernel function and their spatial derivatives, which need a very large computational cost. The most common methods for searching are all-pair search and Linked list algorithm. All-pair search is simple but inefficient. Linked list algorithm is efficient but not dynamic and requires a large list [4]. In order to reduce the computational cost and improve the algorithm adaptive, further optimization of the search algorithm is necessary. With the continuous development of high-performance computers, in the parallel system architecture, study how to improve the efficiency of the parallel algorithm is also necessary.

127.2 Basic Principles of SPH

Firstly, SPH assumes that one field variable can be approximate by the weighted sum of variables in the variable’s support domain, weights of these variables in the support domain are decided by a kernel function that is actually a interpolation function defining how each variable contributes to the field variable [5].

Then, SPH discretizes the computational field by making a series of particles taking on the property variables. Any field variable at the point i can be estimated by Eq. (127.1), and the gradient of the field variable can be estimated by Eq. (127.2).

$$\langle f(x_i) \rangle = \sum_{j=1}^N \frac{m_j}{\rho_j} f(x_j) \cdot W(x_i - x_j, h) \tag{127.1}$$

$$\langle \nabla f(x_i) \rangle = \sum_{j=1}^N \frac{m_j}{\rho_j} f(x_j) \cdot \nabla W(x_i - x_j, h) \tag{127.2}$$

N is the number of particles in the support domain, J is the subscript of particles within the support domain, m is the particle mass, ρ is the particle density, x is the particle position, h is the smoothing length (if $|x_i - x_j| \leq \kappa h$, j is neighbor particle of i), W is the kernel function.

Finally, SPH converts the continuous equation into discrete equation. The discrete form of Navier–Stokes equation is shown in (127.3).

$$\left\{ \begin{aligned} \frac{d\rho_i}{dt} &= \rho_i \sum_{j=1}^N \frac{m_j}{\rho_j} v_{ij}^\beta \cdot \frac{\partial W_{ij}}{\partial x_i^\beta} \\ \frac{dv_i^\alpha}{dt} &= \sum_{j=1}^N m_j \frac{\sigma_i^{\alpha\beta} + \sigma_j^{\alpha\beta}}{\rho_i \rho_j} \frac{\partial W_{ij}}{\partial x_i^\beta} \\ \frac{de_i}{dt} &= \frac{1}{2} \sum_{j=1}^N m_j \frac{p_i + p_j}{\rho_i \rho_j} v_{ij}^\beta \frac{\partial W_{ij}}{\partial x_i^\beta} + \frac{\mu_i}{2\rho_i} \varepsilon_i^{\alpha\beta} \varepsilon_j^{\alpha\beta} \end{aligned} \right. \tag{127.3}$$

m is the particle mass, ρ is the particle density, x is the particle position, v is the particle velocity, σ is stress tensor, p is pressure, e is internal energy, μ is viscosity, ε is Strain rate, $\alpha\beta$ is coordinate direction.

127.3 Algorithm Description

The core of SPH is nearest neighbor search subroutine. In order to improve the efficiency of the subroutine, two improved search algorithms are proposed for searching neighbor particles.

We divide the entire search process into pre-search process and main searching process in two improved search algorithms. Pre-search process searches precisely neighbor particles of a small part of the particles while reducing the search range of most of the remaining (not search) particles. The main searching process is to find the nearest neighbor of each remaining particle within the range which we get from the pre-search process. Before searching, we defined two kinds of data structures: tag array and main searching array. Tag array is used to distinguish in which process the neighbor particles of particle are searched precisely. For example, if the tag of a particle is 0, its neighbor particles are precisely found in the pre-search process. If the tag of a particle is 1, its neighbor particles are precisely found in the maining search process. All particles are initially marked as 0. Main searching array is generated during pre-search process, which gives the search range of each remaining particle, and is the search object of main searching process. Algorithm is briefly summarized as follows:

Firstly, the pre-search process traverses each particle according to particle number and do selective search. If the tag of a particle is 0, search its nearest neighbor particles till all nearest neighbor particles are marked as 1. For all these particles which had just been marked as 1, each particle (such as i) calls all the other particles as “quasi-neighbor particles”, and put the particle numbers of “quasi-neighbor particles” into the corresponding positions (i th row) of main searching array, waiting to search exactly its(i) nearest neighbor particles in the main searching process. If the particle is marked as 1, not search its neighbor particles in the pre-search process. After the pre-search process is completed, all “quasi-neighbor particles” of every particle whose tag is 1 constitutes its own search range of nearest neighbor particles. Example of pre_search process is shown in Fig. 127.1.

Secondly, optimize the main searching array generated in pre-search process, remove repeat numbers of particles, avoid duplicatly considering the interaction between particles.

Finally, the main searching process traverses each row of the optimized main searching array according to particle number, exactly searches its nearest neighbor particles from the “quasi-neighbor particles” of the remaining particles. Now search process is completed, each particle finds its nearest neighbor particles. Pre-search process has greatly reduced the search range of most of the remaining

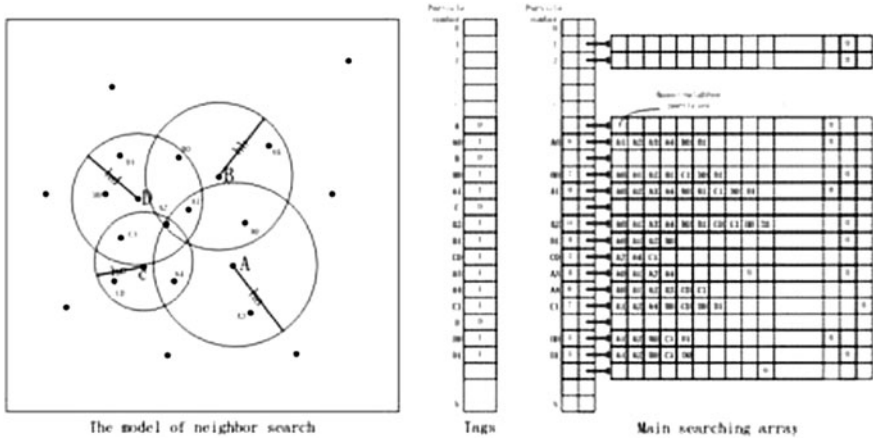
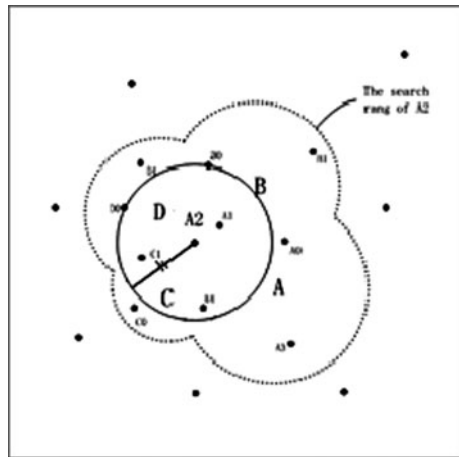


Fig. 127.1 Example of pre_search process

Fig. 127.2 Example of main searching process



particles, therefore, the main searching process is very efficient. In addition, main searching method is suitable for solving the problem which the smooth length is variable in the SPH. Especially, the operating is more efficiency than the previous two methods (All-pair search, Linked list algorithm) in the case of the large number of particles. Example of main searching process is shown in Fig. 127.2.

In this Chapter, we use two different pre-search methods, specifically described as follows.

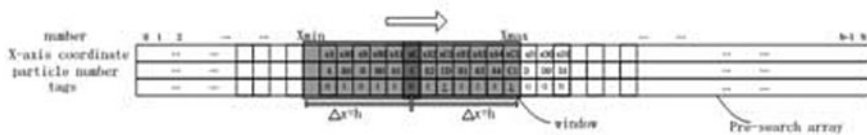


Fig. 127.3 Example of pre-search process based on sliding window

1. Pre-Search Method Based on Sliding Window

This method defines a pre-search array which holds a component of a particle position coordinates (such as the X-axis coordinate), particle number and tags (the initial values are 0). Pre-search process is following: Firstly, the array is sorted according to their X-axis coordinate (using heap sort), Then, all the neighbor particles of every particle will be in a range of array “centering” on this particle. This range of the array is called the window. Each particle corresponds to a window and the width of windows varies due to particle. Secondly, the window serially traverses ordered array of the above. The particle at the center of the window is called the current particle. If its tag is 0, search precisely the neighbor particles within its Window range (determined by their smoothing lengths) then all neighbor particles are marked 1 and put their “quasi-neighbor particles” into corresponding position of the main searching array. If the particle is marked as 1, just slide the window into the next particle. When the center of the window slides into the last position of array, pre-search process is completed. As a result, the particle whose tag is 0 finds precisely its neighbor particles, and the particle whose tag is 1 obtains its “quasi-neighbor particles”. Example of pre-search process based on sliding window is shown in Fig. 127.3.

2. Pre-Search Method Based on Background Grid

Similar to the commonly linked list algorithm, this method splits the computational domain using the background grid. The pre-search process finds the neighbor particles from the current grid of particles located and a number of nearby grid. The detailed pre-search process is similar to above-mentioned. Traverses each particle according to particle number to find precisely neighbor particles of a small part of the particles and obtain “quasi-neighbor particles” of the remaining particles. Example of pre-search process based on background grid is shown in Fig. 127.4.

127.4 Time Complexity of Algorithm

Table 127.1 shows the time complexity of all-pair search algorithm [4], Linked list algorithm [6], search algorithm based on sliding window (A), search algorithm based on background grid (B).

Fig. 127.4 Example of pre-search process based on background grid

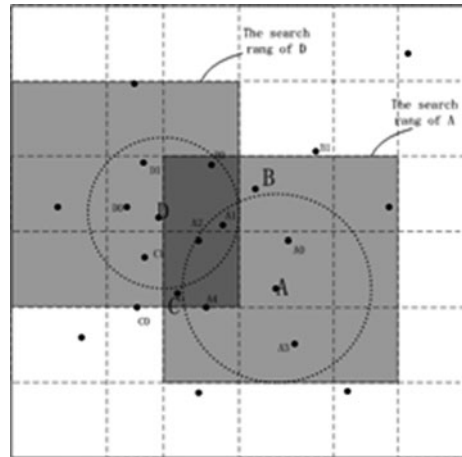


Table 127.1 Time complexity of algorithm

Algorithms	All-pair search	Linked list	A	B
Time complexity	$O(n^2)$	$O(n)$	$O(n \log n)$	$O(n)$

Table 127.2 Computation time of algorithms (4,000 iteration steps)

All-pair search	Linked list	A	B
21908.37 s	8007.72 s	9478.86 s	4708.34 s

127.5 Numerical Simulation and Results Analysis

Simulations are performed to evaluate the effectiveness of the proposed method. The first simulation is a three-dimensional shear cavity flow simulation. Fluid placed in the closed cube containers. In order to drive the fluid flow, top surface of the container move at a constant speed while other surfaces are still. A swirl along the velocity direction is formed when flow is stable. In this experiment, edge lengths of cube is $l = 10^{-3}$ m, viscosity coefficient and density are $\nu = 10^{-6}$ m²/s and $\rho = 10^3$ kg/m³, Container cover is moving along the Y-axis direction and velocity is $v_{top} = 10^{-3}$ m/s. In this case, Reynolds is 1. Initially, 17,602 particles uniformly distributed in container. Time step is set to a constant 5×10^{-5} s. After 4,000 iteration steps, flow is stable.

Table 127.2 shows the computation time of the conventional and the improved algorithms. A is sliding windows algorithm, B is background grid algorithm.

Figure 127.5 shows the result of velocity vector of above-mentioned simulation algorithm.

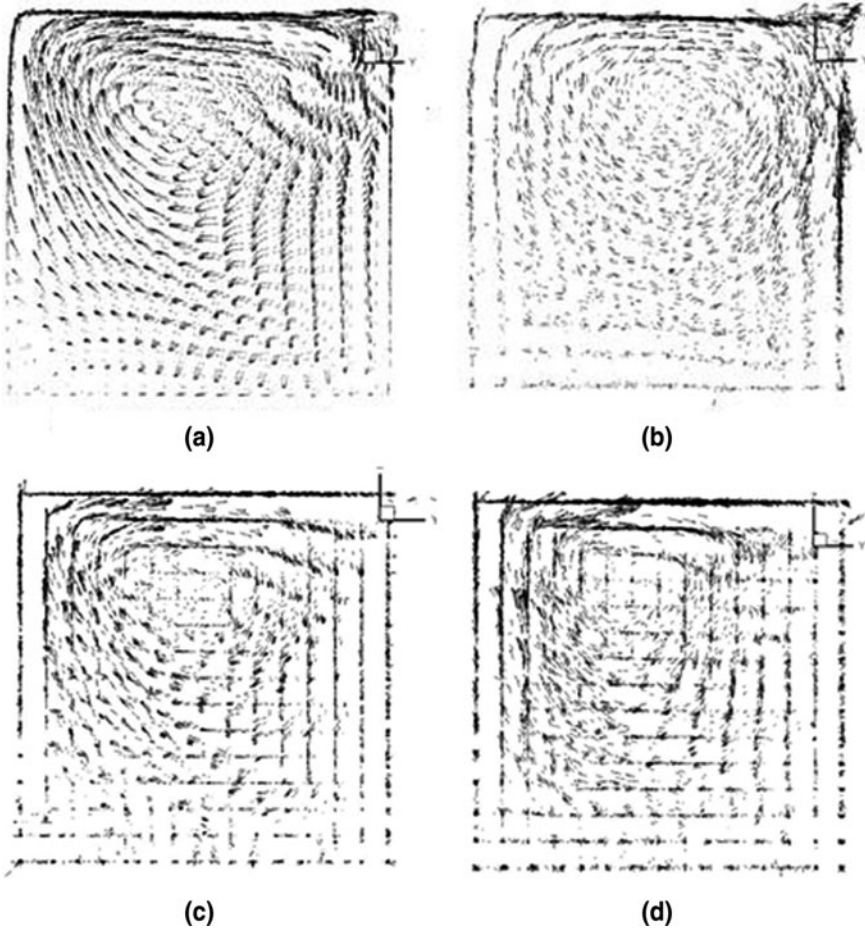


Fig. 127.5 Result of velocity vector. **a** All-pair search algorithm **b** A algorithm **c** Linked list algorithm **d** B algorithm

127.6 Parallel Implementation

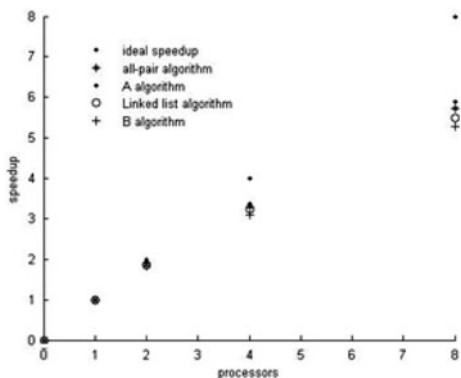
Most small-scale simulation applications are implemented by sequential simulation techniques. As the problem size increases, however, sequential techniques may be unable to reduce the computational cost of the simulation applications adequately. It is natural to consider re-implementing the corresponding large-scale simulations using parallel techniques, which have been reported to be successful in reducing the computational cost for several examples [7].

The second simulation is also a three-dimensional shear cavity flow simulation. We divide up this task among processors in a rather straightforward way, Each

Table 127.3 Computation time of parallel program

Processor	All-pair search			Linked list		
	Time/s	Speedup	Efficiency/%	Time/s	Speedup	Efficiency/%
1	26907.4	1.000	100	10,167	1.000	100
2	14084.5	1.910	95.50	5459.2	1.862	93.10
4	8119.8	3.314	82.85	3143.7	3.234	80.85
8	4707.1	5.716	71.45	1851.6	5.491	68.64
Processor	A			B		
	Time/s	Speedup	Efficiency/%	Time/s	Speedup	Efficiency/%
1	12003.2	1.000	100	6267.5	1.000	100
2	6201.3	1.936	96.80	3400.8	1.843	92.15
4	3540.8	3.390	84.75	2014.5	3.111	77.78
8	2036.5	5.894	73.68	1189.6	5.269	65.86

Fig. 127.6 Speedup of algorithms



processor with approximately the same number of particles [8]. The examples were carried out on a cluster with 8 processors using MPI for message passing.

Table 127.3 shows the computing time of 4,000 iterations for above-mentioned simulation algorithm. Result shows that parallel simulation offer significant reduction in the execution time.

Figure 127.6 shows the speedup values for above-mentioned simulation algorithm. Result shows that improved algorithms have also a good parallelization.

127.7 Conclusion

In this paper, two improved search algorithms based on sliding windows and background grid are applied to three-dimensional shear cavity flow. Results show that the sequential program can greatly reduce the computational cost while

maintaining the same accuracy. In addition, the improved search algorithms have been implemented on a cluster, the experimental results demonstrate that the parallel program has a higher speedup.

Acknowledgments This work is funded by National Natural Science Foundation of China (No. 11002086) and Shanghai Leading Academic Discipline Project (J50103).

References

1. Liu GR (2005) Smooth particle hydrodynamics—a meshfree particle method, Hunan University Press
2. Xiaoyan S, Wang J (2007) Theory and application of SPH method. Water Hydropower Technol
3. Joshua F StarCrash: a parallel smoothed particle hydrodynamics (SPH) code for calculate stellar interactions. Northwestern University
4. Chen J (2009) SPH-based visual simulation of fluid. 4th international conference on computer science and education
5. Song B (2010) A new boundary treatment method for SPH and application in fluid simulation. 3rd international conference on information and computing
6. Chen R (2009) Moving least squares particle hydrodynamics based on leaping search. Acta ArmamentariII
7. Guo L (200) The applications of parallel computing in particle method. 9th computing chemical conference
8. Jia L (2005) Computing of complex molecule dynamics simulation on PC cluster-based system using MPICH technology. Comput Appl Chem

Chapter 128

Comprehensive Evaluation of TPL Using Genetic Projection Pursuit Model with AHP in Supply Chain

Fuguang Bao, Jiandong Si, Lei Zhang and Ertian Hua

Abstract This chapter studies that how to select TPL partner in the supply chain for core enterprise. In order to achieve long-term supplier relationships with TPL, manufacturers are always trying to choose a better TPL partner. We, from the cost, service quality, logistics capabilities, cooperation stability and corporate image, design the multi-index evaluation system of TPL providers. Evaluation model is constructed using projection pursuit and AHP, and the matching genetic algorithm is designed for the model. At last, calculation case analysis is given. The result of evaluation by this proposed method is objective, comprehensive and reasonable.

Keywords TPL · AHP · Projection pursuit · Multi-index · Genetic algorithm

F. Bao (✉) · J. Si · L. Zhang · E. Hua

College of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, China
e-mail: baofuguang@126.com

J. Si
e-mail: sijiandongwu@163.com

L. Zhang
e-mail: zhangleixiaoshi@tom.com

E. Hua
e-mail: huaertian@mail.zjgsu.edu.cn

L. Zhang
East China University of Science and Technology, Shanghai, China

128.1 Introduction

With modern logistics management theory and practice to develop, the optimization problem of supply chain has been getting attention. In order to get efficient supply chain, the tasks of logistics given to the professional TPL Company has become one of the most effective strategies of the manufacturing enterprises' increasing competitiveness. Christopher and Juttner [1] presented an idea to establish cooperation and develop strategic partnership with logistics service providers. Liu [2] studied the transaction cost analysis of the logistics outsourcing. Liu and Wang [3] used Delphi method to establish important evaluation criteria of TPL. But most of the main determinants are much more subjective, which may affect the evaluation and selection. In this chapter, we adopt a fuzzy mathematics that called projection pursuit clustering to achieve evaluation and our goal. The projection pursuit is based on the thinking of making multi-dimensional data projection onto low-dimensional space, and then through the analysis of low-dimensional space data to know the characteristic features of the multi-dimensional data structures. This chapter consists of three different techniques: (1) use fuzzy Delphi method to identify important evaluation criteria; (2) apply AHP to design the multi-index evaluation system of TPL providers preliminary; and, (3) develop a projection pursuit approach for the final selection. Projection pursuit is a useful tool that developed in statistics for seeking "suitable" projections to try to find the intrinsic structure hidden in the high dimensional data. If we could fortunately find a suitable projection direction that gives the clear intrinsic data structure in low dimension space, it is possible for us to further obtain the good fitting results for classification or regression in lower dimension for visual pattern recognition [4].

The proposed method enables decision analysts to better understand the complete evaluation process of 3PL selection. Furthermore, this approach provides a more accurate, effective decision support tool for 3PL provider selection. Finally, an actual industrial application is presented to demonstrate the proposed evaluation and selection.

128.2 Appraisal Target System of TPL Providers

For the indicators of evaluating and selecting TPL providers in the supply chain, there are a number of related researches at home and abroad.

In this article, we summarize the relevant findings at home and abroad [5–7], and based on the research results, combined with practice, using Delphi method, put forward "the level of cost, the quality of services, logistics capability, expected cooperation stability and corporate image" five important areas as the primary level of appraisal target.

Table 128.1 The comprehensive evaluation system of TPL partners

Composite index (target level)	First-level indexes a_m	Second-level indexes	Weight1 a_{mn}	Weight2 b_j	
The comprehensive evaluation of TPL Partners	The level of cost (0.25)	Cost	1	0.25	
	The quality of services (0.3)	On-time shipments and deliveries	0.2	0.06	
		The wastage rate	0.4	0.12	
		JIT responsiveness	0.1	0.03	
		Customer satisfaction	0.2	0.06	
		Capability of specific business requirements	0.1	0.03	
		Logistics capability (0.15)	Logistics equipment	0.2	0.03
			Growth forecasts of logistics equipment	0.2	0.03
			Logistics information system	0.3	0.045
		Expected cooperation stability(0.2)	Financial considerations	0.3	0.045
			Strategies compatibility	0.3	0.06
	Cultural fit		0.2	0.04	
	Historical experience		0.2	0.04	
	Business relevance		0.3	0.06	
	Corporate image (0.1)	Staff quality	0.3	0.03	
		Enterprise cohesive force	0.3	0.03	
General reputation		0.4	0.04		

In first-level index of “quality of services”, we choose “on-time shipments and deliveries”, “the wastage rate”, “JIT responsiveness”, “customer satisfaction” and “capability of specific business requirements” as the second-level indexes. In first-level index of “logistics capability”, we choose “logistics equipment”, “growth forecasts of logistics equipment”, “logistics information system” and “financial considerations” as the second-level indexes. In first-level index of “expected cooperation stability”, we choose “strategies compatibility”, “cultural fit”, “historical experience” and “business relevance” as the second-level indexes. In first-level index of “corporate image”, we choose “staff quality”, “enterprise cohesive force” and “general reputation” as the second-level indexes.

Each index has metric attributes (negative or positive, objective or subjective). This appraisal system should use objective data as far as possible. A comprehensive evaluation system of TPL partners is shown in Table 128.1.

“Level of cost” and “the wastage rate” are negative indicators, the lower the better. The others are all positive indicators, the higher the better. According to the different data sources of various indicators or the different description way of them, these data may be subjective or objective.

The normalization of projection pursuit puts the whole evaluating indicators into equally importance. While, above all various indicators have different

importance for the target level (composite index), i.e. different wight. So we proposed to solve this problem using AHP, through pair-wising comparison of the importance, we establish an appropriate judgment matrix, and then solve the eigenvalues of the matrix.

Let $A = \{a_m\}, m = 1, 2, \dots, 5; a_m \geq 0 (\sum a_m = 1)$ be a weight set of first-level indexes; let $A_m = \{a_{mn}\}, m = 1, 2, \dots, 5; n = 1, 2, \dots, k; a_{mn} \geq 0 (\sum a_{mn} = 1)$ be a weight set of the second-level indexes in each first-level indexes; let $A_{mn} = \{b_j\} = \{a_m \times a_{mn}\}, m = 1, 2, \dots, 5, n = 1, 2, \dots, k, j = 1, 2, \dots, p, x_j \geq 0 (\sum b_j = 1)$ be a weight set of the second-level indexes toward the target level. Due to space limitations and not the focus of this chapter, here is a solution given directly. The weight set of first-level indexes is $A = \{0.25, 0.3, 0.15, 0.2, 0.1\}$, shown in Table 128.1. That will determine the weight of multi-index evaluation system preliminary. But this AHP determined the weight, which depends on the subjective conscious to great extent.

128.3 Modeling Process of Projection Pursuit Clustering

First, we establish the comprehensive evaluation system of TPL Partners. Then, we use the projection pursuit clustering to achieve the modeling of comprehensive evaluation. Let $x_{ij}^* (i = 1, 2, \dots, q; j = 1, 2, \dots, p)$ be the i th TPL provider's j th indicator data (the number of TPL provider: q , the number of index: p). The steps of projection pursuit clustering as follows:

128.3.1 Normalization

For the original data of the TPL service providers, the whitening operation puts the whole evaluating indicators into equally importance and treats each eigenvector as equally importance. The normalization with Eqs. 128.1 and 128.2, Positive indicators are the higher the better, and negative indicators are the lower the better. Normalization for positive indicators and negative indicators:

$$x_{ij} = \frac{x_{ij}^* - \min(x_{ij}^*)}{\max(x_{ij}^*) - \min(x_{ij}^*)}, \tag{128.1}$$

$$x_{ij} = \frac{\max(x_{ij}^*) - x_{ij}^*}{\max(x_{ij}^*) - \min(x_{ij}^*)}. \tag{128.2}$$

128.3.2 Design Weights Preliminary

Apply AHP to design the multi-index evaluation system of TPL providers preliminary. Normalization data of TPL service providers are given different weights, i.e. $x'_{ij} = b_j x_{ij}$.

128.3.3 Linear Projection

The projection is essentially is a useful tool that developed in statistics for seeking “suitable” projections to try to find the intrinsic structure hidden in the high dimensional data. If we could fortunately find a suitable projection direction that gives the clear intrinsic data structure in low dimension space, it is possible for us to further obtain the good fitting results for exploitation in lower dimension. Let $a = (a_1 a_2 a_3 \dots a_p)$ be p-dimensional unit vector, and $\sum_{j=1}^p a^2(j) = 1$, ($0 \leq a(j) \leq 1$). That is the representative of the projection direction of variable indicators. The i th TPL provider’s projection eigenvalues on a linear space: Z_i

$$z_i = \sum_{j=1}^p a_j x'_{ij}, \quad i = 1, 2, \dots, q. \quad (128.3)$$

128.3.4 Fitness of Projection Pursuit

According to the scatter diagram of Z_i , it can cluster the TPL providers. The distance between the classes should be parted as far as possible, which can fully reflect the different characteristics of data; the density of each class should be high as possible, which can make each class be fully representative. Therefore, the fitness of projection pursuit can be defined as

$$Q(a) = S(a) \times D(a). \quad (128.4)$$

where, $S(a)$ is the distance between the classes, and $D(a)$ is the density of each class.

The distance between the classes $S(a)$ can be calculated by the standard deviation of the TPL providers’ projection eigenvalues.

$$S(a) = \left[\sum_{i=1}^q (z_i - \bar{z})^2 / (q - 1) \right]^{1/2}. \quad (128.5)$$

where, \bar{z} is the mean of the entire TPL provider’s projection eigenvalues. The higher $S(a)$ is, the farther the distance is.

Let r_{kl} be the absolute value of the differentials of any two (k, l) TPL providers’ projection eigenvalues.

$$r_{kl} = |z_k - z_l|, \quad k = 1, 2, 3 \dots q; l = 1, 2, 3 \dots q. \tag{128.6}$$

Thus, the density of each class is

$$D(a) = \sum_{k=1}^q \sum_{l=1}^q (R - r_{kl})f(R - r_{kl}). \tag{128.7}$$

- (1) The lower r_{kl} is, the higher the density of each class is. The higher r_{kl} is, the lower the density of each class is.
- (2) R is bandwidth of density, and its value ranges from ‘ $\max(r_{kl}) + p/2$ ’ to ‘ $2p$ ’.
- (3) $f(R - r_{kl})$ is unit speed function, when $R > r_{kl}$, $f(R - r_{kl}) = 1$, otherwise $f(R - r_{kl}) = 0$.

128.3.5 Optimize Projection Direction

The fitness of projection pursuit $Q(a)$ will change with the changes of the projection direction. The different projection directions reflect the different structural features. The best projection direction can make the data in low dimension space reflect the data in high dimension fully. Thus, we establish the equation as follows:

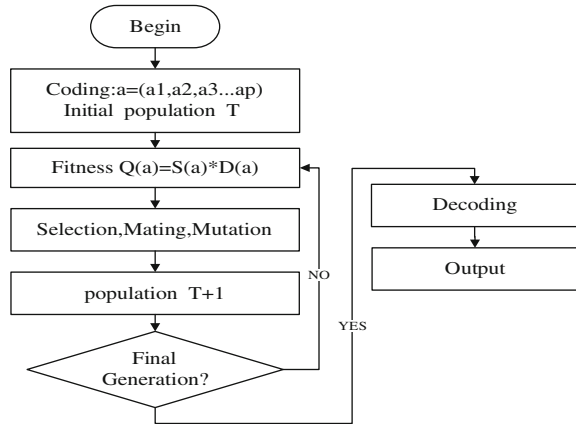
$$\begin{aligned} \text{obj} \quad & \max \quad Q(a) = S(a) \times D(a) \\ \text{s.t.} \quad & \sum_{j=1}^p a^2(j) = 1, (0 \leq a(j) \leq 1). \end{aligned} \tag{128.8}$$

Seen from the objective function and the binding conditions, it is a nonlinear optimization problem. To ensure the accuracy, the value of the projection direction vector is likely to pick up four decimal points. If using method of exhaustion, the amount of calculation can be up to $10,000^p$, which is hard to solve the problem. For the solution and the best values of the projection pursuit, we design a new genetic algorithm based on MATLAB.

128.4 GA to Implement the Projection Pursuit Clustering

Genetic algorithm is a method of searching the best solution, which imitate the natural selection and genetic elimination of biological process. We use genetic algorithms to search for the best direction of the projection, and achieve the projection pursuit clustering (Fig. 128.1).

Fig. 128.1 Flowcharts of genetic algorithm



Step 1: Coding of Chromosome. First of all, coding $a = (a_1 a_2 a_3 \dots a_p)$ with random real array, it called a chromosome or an individual, and $\sum_{j=1}^p a^2(j) = 1$, $(0 \leq a(j) \leq 1)$.

Step 2: Initial population. Random generate generation of an initial population and eliminate some unbound chromosome. Then compute the objective values or fitness of each individual in the population, $fit = Q(a) = S(a) \times D(a)$ and Sort the population by fitness.

Step 3: Selecting. Apply reproduction process and select the solutions as parents by the roulette wheel criterion. Rule: the higher the fitness is, the more probability is, which makes the optimum to inherit.

Step 4: Mating. Apply crossover or mating process with crossover rate pc

Step 5: Mutation. Apply random mutations on the newly generated off-springs. Reduce possibility of a local optimization.

Step 6: final generation, decoding and output. Allocate the resulting individuals into a new population. If the desired population size is not reached, then return to Step 2. Check the stopping condition. Terminate the genetic search process and adopt the best candidate solution as the final solution if the stopping condition is satisfied. Otherwise, proceed to the next generation with the old population to be replaced by the new population, and return to Step 2.

128.5 Calculation Case Analysis

In order to increase the supply chain marketability, core enterprise let logistics outsource. Here are some data of seven TPL providers shown as Table 128.2.

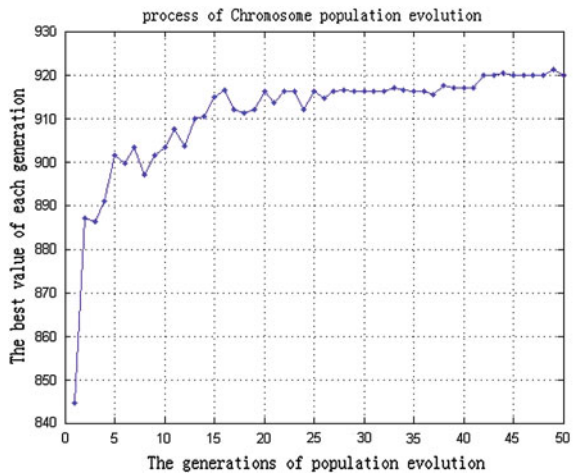
First of all, make normalization data of seven TPL service providers. Then, search for the best projection vector using the genetic projection pursuit model with AHP. Evolution of genetic projection pursuit is shown as Fig. 128.2.

The best projection vector is:

Table 128.2 The original data of seven TPL service providers

Second-level indexes	The original data						
	R1	R1	R3	R4	R5	R6	R7
TPL service providers	R1	R1	R3	R4	R5	R6	R7
Cost	4	3	5	8	6	7	9
On-time shipments and deliveries	80%	85%	86%	95%	93%	90%	98%
The wastage rate	0.1‰	0.5‰	0.3‰	0.2‰	0.3‰	0.4‰	0.6‰
JIT responsiveness	0.8	0.7	0.7	0.6	0.6	0.5	0.7
Customer satisfaction	0.6	0.8	0.4	0.8	0.5	0.7	0.9
Capability of specific business requirements	0.6	0.4	0.5	0.8	0.8	0.9	0.7
Logistics equipment	0.4	0.5	0.6	0.8	0.6	0.7	0.7
Growth forecasts of Logistics equipment	4%	6%	7%	3%	7%	6%	5%
Logistics information system	0.65	0.70	0.40	0.75	0.80	0.50	0.60
Financial considerations	9.0	8.0	8.5	6.5	7.5	6.0	7.0
Strategies compatibility	0.50	0.60	0.40	0.80	0.60	0.55	0.65
Cultural fit	0.20	0.30	0.50	0.40	0.35	0.45	0.30
Historical experience	5.0	7.0	6.0	9.0	8.0	6.5	5.5
Business relevance	0.30	0.50	0.40	0.70	0.50	0.45	0.60
Staff quality	6.0	8.0	7.0	6.5	5.0	3.0	6.0
Enterprise cohesive force	7.0	8.0	8.0	7.5	6.0	5.0	7.0
General reputation	8.0	7.5	7.0	6.0	9.0	5.0	8.0

Fig. 128.2 Evolution of genetic projection pursuit



$$\alpha = (0.35193 \ 0.07610 \ 0.376 \ 0.34902 \ 0.18756 \ 0.27756 \ 0.10796 \ 0.28376 \\
 0.29639 \ 0.30861 \ 0.1575 \ 0.10701 \ 0.37271 \ 0.12078 \ 0.20554 \ 0.17814 \\
 0.20929)$$

Table 128.3 Composite index

TPL	R5	R7	R1	R2	R3	R4	R6
Eigenvalue	1.5863	1.1898	1.1898	0.9915	0.7932	0.3966	0

Then we can compute the eigenvalues of seven TPL providers on this projection vector shown as Table 128.3. The best TPL provider is R5 whose eigenvalue is 1.5863, R1 and R7 are following.

128.6 Conclusion

With modern logistics development, logistics will be outsourced to professional third-party logistics enterprises so that many production enterprises can develop its core business. It concerns how to choose a right, stable and long-term cooperation partner among so many third-party logistics enterprises. This chapter designs an algorithm based on genetic projection pursuit approach to evaluate and select TPL, which combine the strength of AHP on indicators, at the same time, prevent its weakness of subjectivity effectively. We hope this chapter can provide some advices for choosing TPL in the supply chain.

Acknowledgments This research is supported by Ministry of Education's Social Sciences and Humanities Research Project (10YJC630018) and Zhejiang Provincial University Students Scientific Innovative Project.

References

1. Martin C, Uta J (2000) Developing strategic partnerships in the supply chain: a practitioner perspective. *Eur J Purch Supply Manag* 6(2):117–127
2. Liu Y (2005) An analysis of logistics outsourcing in term of transaction costs theory. *Bus Econ Adm* 167(9):13–17 (in Chinese)
3. Liu H, Wang W (2009) An integrated fuzzy approach for provider evaluation and selection in third-party logistics. *Expert Sys Appl* 36(3):4387–4398
4. Feng Z, Zheng H, Liu B (2005) Comprehensive evaluation of agricultural water use efficiency based on genetic projection pursuit model. *Trans CSAE* 21(3):66–70 (in chinese with English abstract)
5. Dickson G (1996) Analysis of venter selection systems and decision. *J Purch* 2:28–41
6. Ackerman K (2000) How to choose a third-party logistics provider. *Mater Handl Manag* 55(3):95–100
7. Wang Y, Sun L, Wang M (2004) Nonparametric frontiers evaluation of eigenvalue attributes of third-party logistics providers. *J Southwest Jiaotong Univ* 39(4):525–530 (in Chinese)

Chapter 129

A Modified PID Tuning Fitness Function Based on Evolutionary Algorithm

Xiao Long Li and Dong Liu

Abstract A modified general PID tuning fitness function based on evolutionary algorithm is proposed. Function is related to several mainly characters of controller response and system robustness. The simulation results show that the function is appropriate for PID tuning with PSO algorithm, the function is also appropriate for unstable time-delayed processes and has good convergence and distinguishability. Compared with ITAE, the PID controllers optimized by the function has better performance.

Keywords Evolutionary algorithm · PID tuning · Fitness function · PSO

129.1 Introduction

PID is the acronym of Proportion, Integral, Differential control, which is a controlling technique developed first. It is widely used in industry control due to its simple arithmetic, better robustness and high reliability. But the effect mainly lies in the three parameters of the controller: i.e., proportion parameter K_p , integral parameter K_i and differential coefficient K_d . So the optimization of parameter has long been a hot issues. In engineering application, common traditional methods include Ziegler-Nichols, Cohen-Coon, improved Ziegler-Nichols, Internal Model Control etc. [1]. These traditional methods have pondered that it's hard to acquire the accurate mathematic model. And the tuning process is semi-experience so as to

X. L. Li (✉) · D. Liu
School of Electrical Engineering, Southwest Jiaotong University,
Chengdu 610031, China
e-mail: halblxl@126.com

simplify the operation, which will hinder the further improvement of its performance. Another genus optimization is based on error rule, like ITAE rule, GISE rule, ISTE rule etc. [2]. PID controller on the basis of these rules often has better performance. But meanwhile we need know a quite accurate model of the controlled object and a better optimization algorithm.

With the rapid development of intelligent optimization technique, many evolutionary algorithm theoretical basis represented by genetic algorithms are maturing and widely using in engineering practice. The PID parameters optimization based on the rule is actually a extreme problem of multivariate function, which has its distinct advantages in solving big-space, non-linear, overall streamlining, which is becoming more and more widely used in many fields, including PID optimizing. Ref. [3] adopts genetic arithmetic to design the optimal PID controller. Ref. [4, 5] adopts improved PSO algorithm to tuning PID parameters.

As far as the essence of algorithm is concerned, genetic algorithm is based on the parameter option of probability. Fitness function evaluates the different parameters. The results are used to indicate the direction of the optimization. A good fitness function should have good selectivity, which will make evolution head in the direction of better parameter performance. Compared to evolutionary algorithm's use in specific situation, the study of evolutionary algorithm is comparatively scarce. Ref. [6] compares the differences of different fitness function in vocal recognition. Ref. [7] compares the differences of different fitness function in antenna bearing chart and puts forward a general fitness function. From the results we can see that proper improvement on fitness function can develop the advantages of evolutionary algorithm.

This paper aims at the time domain response of the control system, puts forward a model of fitness function for the evolutionary algorithm PID parameter. Use closed-loop system's main performance index of step-time domain response to draw the fitness function, considering the robustness of the system at the same time. Optimized PID parameters show good performance on the second-order lag object, which has obvious advantages compared with ITAE rule.

129.2 Fitness Function

129.2.1 Integral Index Fitness Function on the Basis of Optimal Control

PID controller has three parameters: proportion parameter K_p . Integral parameter K_i . Proper parameter can bring good controlling effect for the PID controller. So it vital to optimize these three parameters for the controller. For evolutionary algorithm, to tell the advantages and disadvantages of different PID parameters, we need to evaluate the response of the PID controller which composed by every group of parameters. Ref. [8] pointed out that the two aspects can be evaluated for

the controller's performance: system step-response performance index and system deviation integral performance index. The optimal control always calculate PID parameters with the rule of the minimum system deviation integral performance index. System deviation integral index has five forms [8]: Deviation integral(IE), Integral absolute deviation(IAE), Square deviation integral(ISE), Time absolute deviation product integral(ITAE), Time deviation square product integral(ITSE) and so on. According to the ITAE criterion to get the controller is the best in the response speed and the over compromised effect, In most existing reference of researching in tuning PID parameters of evolutionary algorithm, ITAE rule is used as with the fitness function. ITAE rule as the fitness function, such as (129.1):

$$J = ITAE = \int_0^t t|e(t)|dt. \quad (129.1)$$

where: t is the time of system response, $e(t)$ is system's error. Ref. [2] has joined the error differential, taken the Amplitude margin of system as an index. Ref. [9] add the penalty item to suppress the overshoot on based of ITAE rule. These improvements have not changed the ITAE rule from the essence, the controller response is still compromising between the overshoot and response, the flexibility is insufficient.

129.2.2 Fitness Function Based on the Weighted Sum of Some Performance Indexes of System Time Domain Response

References [9, 10] makes the weighted sum of overshoot of system step-response and adjusting time as the fitness function. Ref. [11] also evaluated the sensitivity function as an index of function. Compared to [10], it is more comprehensive. The fitness function of [11] shows like the equation (129.2).

$$F(k_p, k_i, k_d) = w_1\sigma + w_2t_s + w_3M_s. \quad (129.2)$$

where: σ is the overshoot; t_s is the rise time; M_s is the sensitivity function

129.2.3 A Universal Fitness Function Based on the System Step-response Index of the System

References [2, 7–10] has separately studied the application situation of the fitness function based on the ITAE rule as well as the weighted sums of some performance indexes of system time domain response. However each one has its own limitation, not be able to be suitable with the general PID controller situation.

The fitness function which based on ITAE integral performance index will integrate the product of absolute of the time and errors, the smaller the better results of calculation, which is applicable on many occasions. The disadvantage is that the fitness function is more sensitive to the error itself than the duration of the error, if the system gain is big, the system can complete damped oscillation and eliminate the static error within a very short time. However, due to very short duration, ITAE result is not big. At the same time, ITAE rule is not distinguishable to the overshoot of step-response in controller of different parameters. Not be distinguished is overshoot big while adjustment time is short or overshoot small but a long time to adjust. Rise time and the adjustment time also can not be distinguished. In addition, Using ITAE as the fitness function does not take into account the magnitude margin of system, it is difficult to guarantee to the inaccurate model or the affection of objet control which can not be approximated as FOPDT [2]. ITAE integral performance as the optimization criterion has another drawback that is not directly related to the system output response, not intuitive, and not flexible. When the system is more complex and there is a conflict between the dynamic index, the ITAE-based fitness function can not adjust the weights of different indicators in order to facilitate trade-offs.

By comparison, the time domain response of the controlling system can reflect directly and accurately the performance of the system. The existing fitness function based on the time domain response of the system is not comprehensive. For a closed-loop system, time domain response is including the dynamic processes and static process. Reference [12] pointed out that the rise time, adjusting time and overshoot can evaluate the performance of the dynamic process in system while the steady-state error is used to evaluate the system's static process. The fitness function in type [2] only has two indicators of time domain step-response, not comprehensively response the performance of system, not comprehensive enough, not give full play to the advantages of evolutionary algorithm.

Meanwhile, the system's robustness in the ITAE rule and weighted sum of the time domain indexes are not considered. For the PID controller parameters, there exists a stable region, and the closed-loop system, composed by PID controller of parameters in stable region and controlled object, can be stable. Most of the literatures did not consider the static error because the three parameters of PID are set in stable domain directly. Reference [13] says calculation method of stable parameter domains of PID are mainly DP method, HB method, WangDeJin's method. Since most of the existing reference considered the fitness function of concrete application of the scene, not considered fitness function's commonality, also did not fully exert evolutionary algorithm's advantage of optimization in big extent. Yet there is no publicly report about application of fitness function in the PID parameters optimization, considering both time domain response index and system robustness.

The zero pole of closed-loop control system determines the performance of the system. The distribution of poles determines the stability of system, zero and poles together determine the system response [12]. For the system owning more than two poles, its stability is influenced by the nearest poles [11]. Because the controller is

physically realized, if the system's pole is not in real axis, it must exist in the form of conjugate plural pole, so the root track is real axisymmetric, if there is a stable region, then stable region is also about real axis symmetric. Plural pole corresponds to the time domain oscillation-response, and the imaginary corresponds to the oscillation frequency, the higher the oscillation frequency is, the bigger the imaginary is, so that it inevitably deteriorates the robustness of the system. So this paper uses Concussion number of the system step-response as robustness performance index of controller's parameters. Compared with the Ref. [2], it is simpler and the calculation is very low. In general, the final static error of the optimized PID controller response is very small, so this paper will use the static error as a rule of system stability. General fitness function of this paper, such as the type (129.3a):

$$J = w1Ess + w2\sigma + w3tr + w4ts + w5Cnt. \quad (129.3a)$$

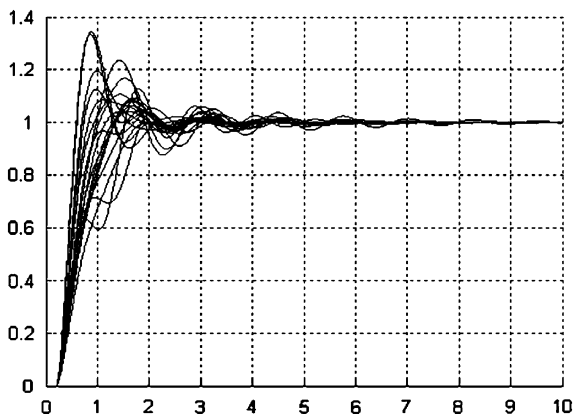
$$Ess = \begin{cases} 1000 & e > 0.02 * SetValue \\ \frac{e}{0.02 * SetValue} & e \leq 0.02 * SetValue \end{cases} \quad (129.3b)$$

where: Ess is a static error, σ is the overshoot, tr is the rise time, ts is the adjustment time, Cnt is the oscillation number of response, in order to evaluate the different parameters uniformly, all of the above parameters were normalized. By the type (129.3a) shows that the fitness function synthesizes the overshoot of system's step-response, rise time, adjustment time, static error and the oscillation number and so on. Taking into account the performance requirements in different control occasions, the fitness function of the various performance indicators can be regulated piecewise or by different weighted parameters. Evolutionary algorithms in real time programming, according to the performance requirements of different applications of the appropriate adjustments to the weights. The fitness function marks the three parameters of PID controller as the optimization goal, the time domain step-response performance as the criteria so as to coordinate dynamical performance indices of system time domain response within a certain range, so to meet the PID controller parameters optimization in different applications.

129.3 Computer Simulation

In order to verify the advantages of improved fitness function model, this paper simulates from two ways: the adapt value curve and the controller response of reality. To the first-order and second-order systems, stability region of PID controller parameters is more easily determined, PID parameters can be tuned easily. But to high-end systems, especially high-end systems with lags, PID parameter optimization is more difficult, considering equation (129.4), the simulation object of high-order, instability and lag [4, 5].

Fig. 129.1 The corresponding response of ITAE group's parameters



$$G(s) = \frac{0.1s + 10}{0.0004s^4 + 0.0454s^3 + 0.555s^2 + 1.51s + 11} e^{-0.2s} \tag{129.4}$$

In order to compare and analyze the relationship between iterations and the change in fitness of the different objective functions, taking the average of 50 experiments to observe. Optimizing by the PSO algorithm of [14], and ITAE rule and the expression showed by (129.3) were used to be the fitness function. When tuning PID parameters, the search range of the Kp, Ki and Kd is [0, 5], the number of particles of PSO algorithm is N = 50, optimization algebra is iter = 50, each 50 times. The w1-w5 in the fitness function in equation (129.3), means that the target of the static error in the system is less than 2% or which is unstable. In stable parameters, the overshoot and the response speed are the main index, considering the robustness of the system when the overshoot and the adjust time are as small as possible.

Adopting ITAE rule as the PSO optimization of fitness function, and in the 50 times tests only 18 times converged, parameters of the corresponding response is shown in Fig. 129.1, and as shown in Fig. 129.2 is average convergence curve; With this fitness function doing PSO optimization of fitness function, there are 21 times convergences in 50 tests, and the response of the corresponding parameter group shows in Fig. 129.3, as shown in Fig. 129.4 is average convergence curve. The comparison results of two groups of the controller step-response evaluation index of different PID parameters shows in Table 129.1, in which weighted group signifies fitness function optimization parameters group in this paper, with ITAE group says the parameters obtained which used fitness functions in ITAE rule. Weighted index 1 corresponding weight value is :w1 = 0.01, w2 = 0.9, w3 = 0.03, w4 = 0.03, w5 = 0.003; weighted index 2 corresponding weighted value: w1 = 0.01, w2 = 0.27, w3 = 0.3, w4 = 0.3, w5 = 0.03.

Line 1 is the name
 Line 2 is ITAE optimal indexes of ITAE group

Fig. 129.2 The changing curve of the average fitness in ITAE group

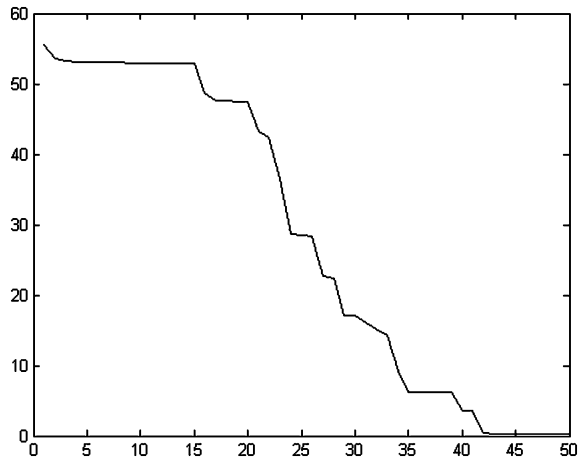


Fig. 129.3 The corresponding response of weighted group parameters

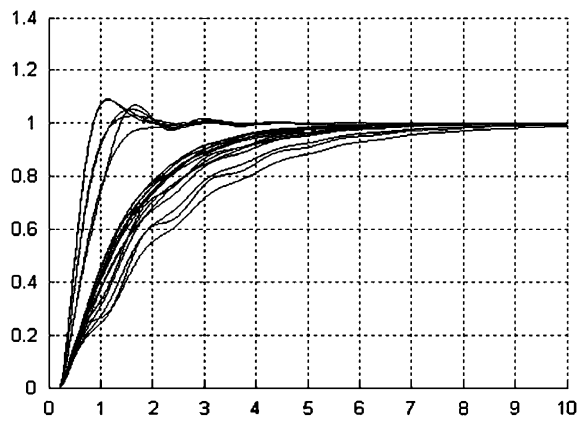


Fig. 129.4 The changing curve of weighted average value

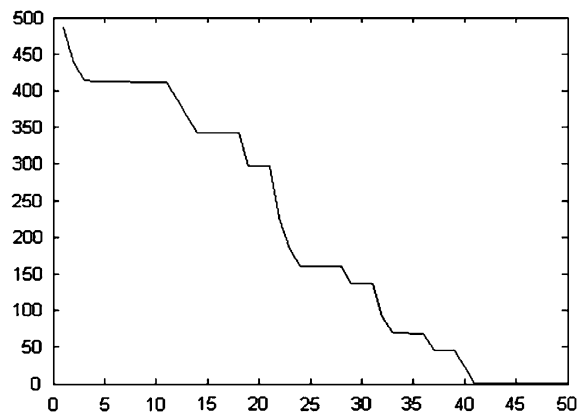


Table 129.1 Comparing the optimal parameters in parameter tuning which optimized from different fitness function

Kp	Ki	Kd	$\sigma(\%)$	Tr[s]	Ts[s]
0.0785	1.8194	0.0797	3.89	0.682	4.498
0.0454	1.6895	0.0810	1.15	1.120	5.074
0.0785	1.8194	0.0797	3.89	0.682	4.498
0.1135	2.3157	0.1065	9.07	0.047	3.680
0.0556	1.411	0.0630	0.43	0.948	2.981
0.0877	1.8554	0.0859	2.82	0.630	2.599

Line 3 is weighted indexes 1 of ITAE group

Line 4 is weighted indexes 2 of ITAE group

Line 5 is ITAE optimal indexes of weighted group

Line 6 is weighted indexes 1 of weighted group

Line 4 is weighted indexes 2 of weighted group

It can be seen from Table 129.1 that this fitness function in this paper can evaluate the performance of the PID parameters flexibly. Weighted index 1 tells that overshoots is the most important evaluation index, and from the parameters obtained from the ITAE optimization, the overshoot of the best PID parameters obtained is 1.15% by weighted index 1, rising time for 1.12 s and adjusting time for 5.074 s; and in weighted index 1, the overshoot of the best PID parameters obtained with fitness function of this paper is 0.43%, in which rising time and adjust time is 0.948 s and 2.981 s respectively. Weighted index 2 says overshoots and response speed must be considered, in the parameters obtained from the ITAE optimization, the overshoot of the best PID parameters obtained is 3.89% by weighted index 2, rising time for 0.682 s and adjusting time for 4.498 s. In ITAE parameters groups, by weighted index 2, the best PID parameters and the parameters obtained from the ITAE rule is consistent. Corresponding to the weighted index 2, the overshoot of the best PID parameters obtained with this fitness function is 2.82%, rise time and adjust time for 0.630 s and 2.599 s respectively; In all obtained, the best PID parameters performance by ITAE rules is obviously worse than the parameters by weighted indexes. This shows the value function to the evaluation of PID parameters in this paper are effective, which is more flexible than ITAE to distinguish different performance indicators.

By the contrast of Fig. 129.1 and Fig. 129.3, the ITAE standards are actually the compromise response time and overshoot, in which the distinction between the parameters is not clear, almost all the of parameters the parameters group have more over shoots, and oscillate more, adjust time is long, and the parameters of the diversity is not big. But the fitness function of this paper have a stronger discernibility ability to the parameters, in which a part of the PID parameters obtained corresponding do not have overshoots, another part in adjust time and overshoots is consist to the PID parameters obtained by ITAE standards and the oscillation of all the parameters is rarely. This shows the parameters with this fitness function

parameters have bigger diversity, better performance and more strong robustness. Simultaneously the optimizing path of parameters of is more able to exert omnidirectional and bring global optimal performance of evolutionary algorithm into play. Just by effect of PSO algorithm performance, particle number and iteration times, it didn't reach the optimal value of weighted index.

129.4 Conclusion

In this paper, under the background of using evolution algorithm to optimize the PID controller, by the comprehensive consideration of each unit and the system robustness of the time domain response, adding other step-time domain indexes to construct fitness functions based on the part of the weighted of the time domain indexes and the fitness functions obtained provides a new idea to further optimization for PID controller. Especially it can consider trade-offs for different performance in different occasions flexible and degree of distinction also have greatly ascend by using the fitness functions of ITAE standards relatively widely. Of course, considering the different algorithms background and different optimized objects and optimize goals, usually demand different definitions for function relationship of corresponding weight coefficient. Therefore, we can further explore the relationship of different weight coefficient and system performance, signing a kind of fitness function of good distinction and better convergent caused by the evolutionary algorithm.

References

1. Xue D (2010) System designing and analyzing in feedback Controlling—MATLAB Language application[M]. Tsinghua University Press, Beijing
2. Zhen-ping Z, Zeng-qiang C, Zhu-zhi Y et al (2004) PID controller optimization based on new error integral criterion [J]. Control Eng 11(1):52–54
3. Yang S, Xu S (2010) Spacecraft attitude maneuver planning based on particle swarm optimization[J]. J Beijing Univ Aeronaut Astronaut 36(1):48–51
4. Zi Wu R, Yu Y, Feng Cheng J (2006) Improved PSO algorithm and the application in PID parameter tuning [J]. Sys Simul 18(10):2870–2873
5. Xu X, Qian F, Wang F (2008) Novel method of system identification based on modified particle swarm optimization algorithm[J]. Syst Eng Electron 30(11):2231–2236
6. Altincay H, Demirekler M (2001) Comparison of different objective functions for optimal linear combination of classifiers for speaker identification [J]. IEEE Acoust Speech, Signal Process 2001(1):401–404
7. Yu F, Ronghong J, Liu B et al (2005) The objective function of genetic algorithm in Array antenna pattern synthesis 27(5):801–804
8. Huang Z (2004) Control system MATLAB calculation and simulation (version 2)[M]. National Defence Industry Press, Beijing
9. Liu J (2007) Advanced PID controller and MATLAB simulation (Version 2)[M]. Electronic Industry Press, Beijing

10. Haijie M, Xiaolin F, Li W, Hui C (2010) Study of intellectual auto-generating test-paper based on adaptive genetic algorithm [J]. 2010 In: Proceedings of the third international conference on education technology and training (ETT), pp 5108–5120
11. Lin Lin O, Gu YY, Zhang WD (2007) The optimal PID parameter tuning method bases on gain margin and phase margin[J]. Control Theory Appl 24(5):837–840
12. Zhang W, Jianqin M (2010) Fuzzy tree model based on adaptive fuzzy control method synovial [J]. Beijing Control Theory Appl
13. Hongbo M, Jing Yuan F (2008) The optimal PID controller design in BUCK power converter[J]. Electr Mach Control 12(6):639–643
14. Chuan L, Quanyuan F (2008) A new particle swarm optimization which can adapt itself[J]. Comput Eng 34(7):181–183

Chapter 130

Analysis of Convergence for Free Search Algorithm in Solving Complex Function Optimization Problems

Lu Li, Zihou Zhang and Xingyu Wang

Abstract Free Search (FS) algorithm is efficient in solving complex function optimization problems. Convergence of FS is analyzed in two different cases: continuous and discrete space. For continuous space, convergence of FS did not exist for all functions, such as functions containing singular points. However, taking advantage of measure theory, convergence of FS could be shown when functions satisfy continuous and Lipschitz conditions. For discrete space, making use of random process theory, convergence of FS was got in finite search space and FS was characterized by Markov property. Simulation of multi-model Shubert function is done. Compared with Genetic Algorithm (GA), FS is superior in convergence speed, accuracy and robustness. The analytic and experimental results on FS provide useful evidence for further understanding and properly tackling optimization problems of complex functions.

Keywords Free search algorithm · Convergence of algorithms · Optimization problems of complex functions · Markov chain · Genetic algorithm

130.1 Introduction

Biologically information transfer can be divided into two types. One is to inherit information genetically. Genetic information is crucial to the survival and evolution of species. A good example is Evolutionary Computation

L. Li (✉) · Z. Zhang
College of Fundamental Studies, University of Engineering Science, Shanghai, China
e-mail: ecnulilu@sohu.com

L. Li · X. Wang
School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China

(EC), which includes major evolutionary theories such as Genetic Algorithms (GA) [1].

Another type is non-genetic information transfer, which mainly refers to the information transfer between individuals and living environment. Swarm Intelligence Algorithm is a random search simulating behaviors of social animals, which includes two main algorithms, that is, Ant Colony Optimization (ACO) [2] and Particle Swarm Optimization (PSO) [3].

Free Search (FS) was proposed by Penev and Littlefair [4, 5], which has similarities with ACO and PSO. It is a novel population-based optimization method, which combines with survival-of-the-fittest mechanism in GA, theory of pheromone and individual search space in ACO, and advantages of swarm memory in PSO. Individuals in FS have not only peculiarities of lower animals but also those of higher animals, namely sense and mobility [6]. Each individual has its own sense of smell, which is defined as sensibility in FS. Sense of smell enables each individual to have distinguishing ability in search space. Unlike GA, ACO and PSO, the behavior of animals is not confined in FS. FS is based on the idea: with an uncertainty can cope other uncertainty, with infinity can cope other infinity.

130.2 Mathematical Model of FS

Consider a minimization problem (X, f, Ω) : where X is the set of (candidate) solutions, f is the objective function, which assigns to each candidate solution $x \in X$ an objective function value $f(x)$, and Ω is a set of constraints, which defines the set of feasible candidate solutions. By using penalty function, optimization of constrained function can be converted into optimization in the following constraint space:

$$\begin{aligned} \min f(x_1, x_2, \dots, x_n) & \quad (130.1) \\ \text{s.t. } x_{i,\min} < x_i < x_{i,\max}, & \quad (i = 1, 2, \dots, m) \end{aligned}$$

In solving the above problem, FS includes three components: initialization, exploration and termination.

130.2.1 Parameter Initialization and Generation of Initial Group

Initial parameters include: amount of individuals M ; individual number j ($j = 1, 2, \dots, M$), coordinate point to mark pheromone k ($k = 1, 2, \dots, M$); variable number of objective function n , namely dimension number of search space; steps in search walk T ; current step of search walk t ; search space of individual j in the i -th dimension space $R_{j,i} \in [R_{\min}, R_{\max}]$; termination function G .

Group initialization has three forms: random value, certain value and single value. Random value is the commonly-used form.

$$x_{0ji} = x_{i,\min} + (x_{i,\max} - x_{i,\min}) \cdot \text{random}_{ji}(0, 1)$$

where $x_{i,\max}, x_{i,\min}$ are the maximum and minimum of the i th dimension variables, $\text{random}(0, 1)$ is a random value between 0 and 1, M is coordinate point in any search space.

130.2.2 Search Behavior

In search walk, individual behavior is described as follows:

$$\begin{aligned} x_{t,j,i} &= x_{0,j,i} - \Delta x_{t,j,i} + 2\Delta x_{t,j,i} \cdot \text{random}_{t,j,i}(0, 1) \\ \Delta x_{tji} &= R_{j,i} \cdot (x_{i,\max} - x_{i,\min}) \cdot \text{random}_{tji}(0, 1) \end{aligned}$$

130.2.3 Termination Condition

Acceptable strategies for termination are: (1) reaching of the optimization criteria: $f_{\min} \leq f_{opt}$, where f_{\min} is minimal achieved solution x , f_{opt} is an acceptable value of the objective function; (2) expiration of generation limit: $g \geq G$, where G is a limit; (3) complex criterion: $f_{\min} \leq f_{opt}$ or $g \geq G$.

130.3 Convergence Proof for FS

130.3.1 FS Convergence in Continuous Space

For (1), suppose that $X = \{(x_1, x_2, \dots, x_n) | x_{i,\min} \leq x_i \leq x_{i,\max}, i = 1, 2, \dots, n\}$, the best-achieved value is x^* , the optimum of each generation in FS is defined as $x^*(t)$, t is generation limits, then $\{x^*(t), t = 0, 1, 2, \dots\}$ consists of a discrete random process.

Definition x^* is the optimum, for any given positive number δ and ε , if there exists a corresponding $N(\varepsilon, \delta)$, when $t > N(\varepsilon)$ satisfies $P(|f(x^*(t)) - f(x^*)| \leq \delta) \geq 1 - \varepsilon$, then $x^*(t)$ is said to be the best-achieved value via probability convergence.

Theorem 1 Assume the optimum x^* in X is singularity of function $f(x)$, i.e. $\exists \varepsilon > 0$, satisfies $\min_{x \neq x^*} \{f(x)\} > f(x^*) + \varepsilon$, then FS cannot converge to the global optima.

Proof Assume solution space is X , its Measure is $|\Omega| = \prod_{i=1}^n (x_{i,\max} - x_{i,\min}) > 0$, its optimum $|\arg\{f(x^*)\}| = 0$, then and $P\{x^*(t + 1) = x^* | x^*(t) \neq x^*\} = 0$, namely FS cannot guarantee to Converges to the global optima.

Theorem 2 Assume $f(x)$ satisfies the following requirements in boundary space X : (1) $f(x)$ is continuous ; (2) $f(x)$ satisfies Lipschitz condition, namely $\exists L$, for $\forall x_1, x_2 \in \Omega$, $|f(x_1) - f(x_2)| \leq L|x_1 - x_2|$; (3) $f(x)$ has the only optimum x^* in space X , then the optimal solution sequence $\{x^*(t)\}$ has the global optimal value via probability convergence in any accurate sense.

Proof Assume solution space is X , $f(x)$ is continuous, and $f(x)$ has the only optimum x^* in space X , there exists a neighborhood $U(x^*)$ of x^* satisfies $\forall x_1 \in U(x^*), x_2 \notin U(x^*)$, and $f(x_1) < f(x_2)$. Where $|X|$ is the measure of collection X , for any x from solution space X , then $P(x \in U(x^*)) = \frac{|U(x^*)|}{|\Omega|}$. Since $f(x)$ satisfies Lipschitz condition, for $x^*(t)$ and x^* , if $L|x^*(t) - x^*| \leq \delta$, namely $|x^*(t) - x^*| \leq \frac{\delta}{L} \triangleq \varepsilon$, then $|f(x^*(t)) - f(x^*)| \leq L|x^*(t) - x^*| \leq \delta$. Therefore only if $|x^*(t) - x^*|$ is small enough, $|f(x^*(t)) - f(x^*)| \leq \varepsilon$ is guaranteed.

Assume $\theta = \frac{|U(x^*)|}{|\Omega|}$, so probability of each individual entering $U(x^*)$ for one time through free search is θ . The probability of individual entering for k times is $1 - (1 - \theta)^k$. When $\theta > 0$, $\lim_{k \rightarrow \infty} (1 - (1 - \theta)^k) = 1$. Therefore in probability sense, FS can search the neighborhood of the optimum, that is $\lim_{k \rightarrow \infty} P(x_k \in U(x^*, \delta)) = 1$, namely $\lim_{t \rightarrow \infty} P(|f(x^*(t)) - f(x^*)| \leq \varepsilon) = 1$.

If population size is M , after one free search, probability of at least one individual in M entering $U(x^*)$ is $1 - (1 - \theta)^M$; after k times free search, the probability is $1 - (1 - \theta)^{Mk}$.

130.3.2 FS Convergence in Discrete Space

Assume the solution space of test problem (1) is

$$X = \{(x_1, x_2, \dots, x_n) | x_i \in X_i = \{x_{i,1}, x_{i,2}, \dots, x_{i,M_i}\}, i = 1, \dots, n\},$$

then $|X| = \prod_{i=1}^n M_i$, that is to say, the solution space is finite, where X is the population formed by the factors in Ω , the population size is $N = |X|$, $|\Omega|^N$ is the probable amount of populations.

If the optimal solution is $x^* = (x_1^*, x_2^*, \dots, x_n^*), x_i^* \in \Omega_i$, the optimal solution sequence in each search process of FS is $\{f(x^*(t))\}$. $\{f(x^*(t))\}$ is a random process. FS bases on individual parents to search, $f(x^*(t + 1))$ is only related to individuals

of generation t and not to the previous generations of populations. So $\{f(x^*(t))\}$ forms a Markov Chain. If the correspondent state set of $\{f(x^*(t))\}$ is E_i , and the correspondent state set of $f(x^*(t + 1))$ is E_j , then p_{ij} is the transfer probability from generation t in the state E_i to generation $t + 1$ in the state E_j . Then

$$p_{ij}(t) = \begin{cases} \frac{1}{|E|}, & f(E_i) < f(E_j) \\ p_{ii}(t) = 1 - \sum_{j \neq i} p_{ij}(t), & f(E_i) < f(E_j) \\ 0, & f(E_i) \geq f(E_j) \end{cases}$$

Theorem 3 *In finite space, FS converge to the optimal solution with probability.*

Proof Arrange each individual according to its fitness, and the transfer probability matrix of generation t in Markov chain is

$$P(t) = (p_{ij}(t))_{|E||E|} = \begin{pmatrix} p_{11}(t) & & & & \\ p_{21}(t) & p_{22}(t) & & & \\ \vdots & \vdots & \ddots & & \\ p_{|E|1}(t) & p_{|E|2}(t) & \cdots & p_{|E||E|}(t) \end{pmatrix} \text{ where}$$

$$p_{ij} = \begin{cases} \frac{1}{|E|}, & i > j \\ 1 - \frac{i}{|E|}, & i = j, \text{ satisfy } \sum_j p_{ij} = 1. \\ 0, & i < j \end{cases}$$

If the transfer matrix of P in k walk is $P^{(k)}$, according to nature of Markov chain,

we know $P^{(k)} = P^k$, namely $\lim_{k \rightarrow \infty} P^{(k)} = \lim_{k \rightarrow \infty} P^k = \begin{pmatrix} 1 & & & & \\ 1 & 0 & & & \\ \vdots & \vdots & \ddots & & \\ 1 & 0 & \cdots & 0 \end{pmatrix}$,

that is $\lim_{k \rightarrow \infty} p_{ij} = \begin{cases} 1, & j = 1 \\ 0, & j \neq 1 \end{cases}$ any state reaches the optimal convergence state.

130.4 Experiments

Choose Shubert function as test function to verify algorithm performance. The Function is

$$f(x, y) = \sum_{i=1}^5 i \cos[(i + 1)x + i] \cdot \sum_{i=1}^5 i \cos[(i + 1)y + i] + (x + 1.42513)^2 + (y + 0.80032)^2,$$

Table 130.1 Results of GA

No.	Optimal solution	Optimal value	No.	Optimal solution	Optimal value
1	(-1.4197, -0.7312)	-176.3255	6	(-1.3867, -0.7813)	-182.4771
2	(-0.7568, -1.4063)	-180.9669	7	(-0.8023, -1.4276)	-185.9264
3	(-0.7352, -1.4737)	-171.4811	8	(-0.8005, 4.8489)	-154.2333
4	(-1.4249, -0.7910)	-186.5396	9	(-1.4969, -0.6851)	-148.9035
5	(-1.4120, -0.70703)	-146.6348	10	(-1.3806, -0.9802)	-115.8060

Table 130.2 Results of FS

No.	Optimal solution	Optimal value	No.	Optimal solution	Optimal value
1	(-1.4253, -0.7989)	-186.7308	6	(-1.4245, -0.8022)	-186.7308
2	(-1.4266, -0.8016)	-186.7309	7	(-1.4258, -0.8011)	-186.7308
3	(-1.4233, -0.8003)	-186.7308	8	(-1.4259, -0.7995)	-186.7308
4	(-1.4217, -0.7990)	-186.7308	9	(-1.4248, -0.8009)	-186.7308
5	(-1.4220, -0.7996)	-186.7309	10	(-1.4269, -0.7997)	-186.7309

where x, y satisfy $-10 \leq x, y \leq 10$. This is a multi-modal function and it has 760 local minima in defined search space, the global minimum $f(-1.42513, -0, 80032) = -186.7309$.

The minimum of Schubert by using GA and FS respectively for 10 times are Table 130.1 as follows:

GA: population size is $M = 101$, search space is $[-10, 10] \times [-10, 10]$, cross-over probability 0.6, mutation probability 0.02, to retain the optimal solution roulette wheel selection, search terminal generation limits $G = 200$. According to the given parameters, the results from 10 continuous operations are as follows:

Choose the following parameters of FS: individuals amount $M = 5$, Steps in Search Walk is $T = 5$, search radius $R_{j,i} = [-10, 10]$, terminal generation limits $G = 200$. According to the given parameters, the results from 10 repetitive operations are Table 130.2 as shown.

As Tables 130.1 and 130.2 demonstrate, the results of GA are: search generation limits $G = 200$ though the population size is 101, the average minimum of 10 repetitive calculations is -164.92942 ; while those of FS: population size is 10, search terminal generation limits $G = 200$ the average minimum of 10 repetitive calculations is -186.73083 , and $\max_j \{ |f(x^*(t) - x_i^*)| \} < 10^{-4}$. This indicates that FS is more effective than GA in solving these optimization problems.

130.5 Conclusion

Solution space is divided into continuous and discrete conditions. The convergence of FS was analyzed in the two different situations. If solution space is continuous, the convergence of FS does not exist for all functions. AS for those functions

satisfying particular conditions, FS has convergence, whose speed is related to the neighborhood measure of the best-achieved value. If solution space is discrete and finite, convergence exists and the optima sequences through FS satisfy Markov property. To understand convergence and its speed, Shubert function is chosen to carry out the simulation by using GA and FS respectively. As a result, FS is superior to GA in terms of optimization outcome. FS produced better results in random selection of initial values, while GA produced unstable results and was sensitive to the selection of initial values. GA is not as good as FS in robustness.

Acknowledgements This paper was supported by Shanghai Leading Academic Discipline Project, No.B504 and Special Fund of Shanghai University of Engineering Science.

References

1. Holland JH (1992) *Adaptation in natural and artificial systems*. MIT Press, Cambridge
2. Dorigo M, Stutzle T (2004) *Ant colony optimization*. MIT Press, Cambridge
3. Kennedy J, Eberhart K (1995) Particle swarm optimization. In: *Proceedings of IEEE international conference on neural networks*, pp 1942–1948
4. Penev K, Littlefair G (2005) Free Search—a comparative analysis. *Inform Sci* 172:173–193
5. Penev K (2006) Novel adaptive heuristic for search and optimisation. In: *IEEE John Vincent Atanasoff 2006 international symposium on modern computing*, pp 149–154
6. Zhou H (2007) A new swarm intelligence algorithm—free search. *J Donghua Univ* 33(5):558–579

Chapter 131

Design of FH Sequences with Given Minimum Gap Based on Logistic Map 1

Wan Youhong and Li Jungang

Abstract Based on logistic map, a novel method by adding binary prefix to generate chaotic frequency-hopping (FH) sequences with given minimum gap was proposed. Theoretical analysis and simulation results show that the FH sequences base on the method have larger FH gap and better sequence generation efficiency than existed methods. In addition, the random-like property, hamming correlation property and bit error rate (BER) of sequences had been analyzed. The simulation results show that the FH sequences based on the method are suitable for FH communication and secure communication.

Keywords Chaos communication · FH sequences · Logistic map · Given minimum gap

131.1 Introduction

Chaotic systems have some special properties such as random-like, non-periodic, non-line. Besides, chaotic systems are characterized by extreme sensitivity to initial condition, a small perturbation eventually causes a large change in the state of the system. They are favorable to generate spread spectrum communication system of better hamming correlation property, more addresses, more

W. Youhong (✉) · L. Jungang
College of Automation, Nanjing University of Posts and Telecommunications,
Nanjing 210046, China
e-mail: wanyh@njupt.edu.cn

L. Jungang
e-mail: lijungang1021@gmail.com

anti-jamming and more anti-intercept capacities when used for pseudo-noise modulation [1]. Based on above various characteristics, chaotic communication systems are widely used in two fields: secure communications and spread spectrum communications.

The performance of FH sequences greatly determines the overall chaotic communication systems performance, especially for chaotic FH communication systems. The study of chaotic FH sequences is generally limited in the hamming correlation property, linear complexity, bit error rate, and hardware implementation etc. In recent years, FH gap widening has been researched and improved in some chapters. The solutions can be broadly grouped into three categories: symmetry band algorithm, shift replace algorithm and drop algorithm.

References [2–6] utilized and improved the symmetry band algorithm, the algorithm utilized a method that chose between up and down bands to realize given minimum gap, by jumping to the symmetry band when FH gap is narrower than given minimum gap. References [7, 8] used and improved shift algorithm respectively. Especially, paper [7] proposed D non-linear mode algorithm, it belonged to shift replace algorithm. The method was improved based on literature [9] that without widen gap process. The detail was to determine the FH gap of two adjacent hopping frequency slot first, and then if the gap was wider than given minimum gap, made $f_{i+1} = f_{i+1}$, else made $f_{i+1} = (f_i + p^r + f_{r+1} \bmod d)$, $d \in [0, p^{r-1}]$. Reference [10] proposed drop algorithm to realize the given minimum gap. According to the distance of two adjacent slots, the method also judged whether current frequency slot should be dropped or not.

However, a process of comparing and judging must be involved no matter which method is used. The comparison point by point would reduce the generation efficiency of frequency hopping sequence in the case of long frequency hopping sequence. This chapter takes an improved logistic map as basic map and proposes a method of adding fixed binary prefix to broad the FH gap effectively. In addition, frequency hopping sequence constructed by the proposed method can get the ideal width of FH interval directly without comparison with next hop point, which improves the efficiency of sequence generation.

131.2 Design of Chaotic FH Sequences

The improved logistic map is defined as:

$$x_{n+1} = f(x_n) = 1 - 2x_n^2 \quad (131.1)$$

where $x_n \in [-1, 1]$, $n = 0, 1, 2, \dots$

Firstly, we produce a series of x_n by iteration with logistic map, and then binarize the values into a binary sequence using the threshold function (131.2), expressed as:

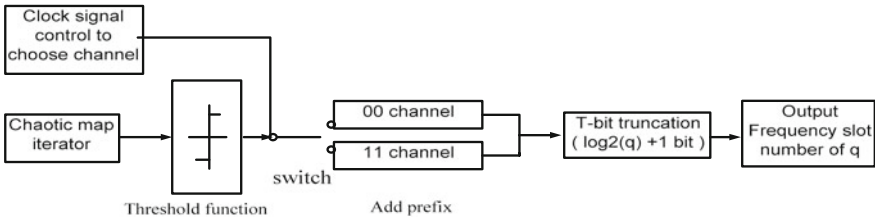


Fig. 131.1 Construction process of the sequence

$$x_n = \begin{cases} 0 & x_n \in [-1, 0] \\ 1 & x_n \in (0, 1] \end{cases} \tag{131.2}$$

Suppose we want to get a chaotic FH sequence with frequency slot number of q and length of N , then we have to iterate logistic map for $N \times (\log_2 q - 1)$ times. Here we transform the binary sequence into a matrix of $N \times (\log_2 q - 1)$.

Then we process the FH gap of sequences as follows. First, we provide two channels: one is to provide two bits of “11”; another channel is to provide two bits of “00”. Under control of the clock signal, each row of the above matrix alternately passes the channels of these two bits prefix, namely, each binary row is added with a “prefix” of two bits, so that the original matrix, after adding the prefix, becomes a matrix with a size of $N \times (\log_2 q + 1)$. After processing the matrix with prefix by T-bit truncation method, we can get a chaotic FH sequence with frequency slot number of q and length of N . While the T-bit truncation method is to truncate at each $(\log_2 q + 1)$ bits by row and execute binary-decimal conversion. Construction process of the sequence is shown as Fig. 131.1. The work of adding “00” and “11” prefix is actually to assign two sequential hopping to two different frequency bands. Assume that the constructed sequence is of frequency slot of number q and bandwidth of q multiplied by bandwidth of each frequency slot. The prefix make each row of the matrix get one more bit, so frequency slot bandwidth after adding prefix is $2q$ times as wide as before. And we only use the front 1/4 and the back 1/4 section of the whole bandwidth, then the frequency slot changes into $[1, q/2]$ and $[3q/2, 2q]$, and the number of frequency slot is still q . In the process of frequency hopping, the hopping is alternatively executed in the first and back section, therefore frequency hopping gap is effectively widened to $Cap \geq (3q/2 - q/2) = q$, compared to theoretical frequency hopping gap $Cap \approx q/3$ of sequence based on [9]. Thus, it can be concluded that the method can result in a considerable increase for frequency hopping interval.

Furthermore, in order to confirm the random-like property and chaos property of the sequences. The concept of exhaustion change rate proposed in paper [11] is used to analyze random-like property of sequences. This method does not require the phase space reconstruction, and has some advantages and convenience in analysis for discrete sequences. The random-like property of sequences is illustrated in Fig. 131.2.

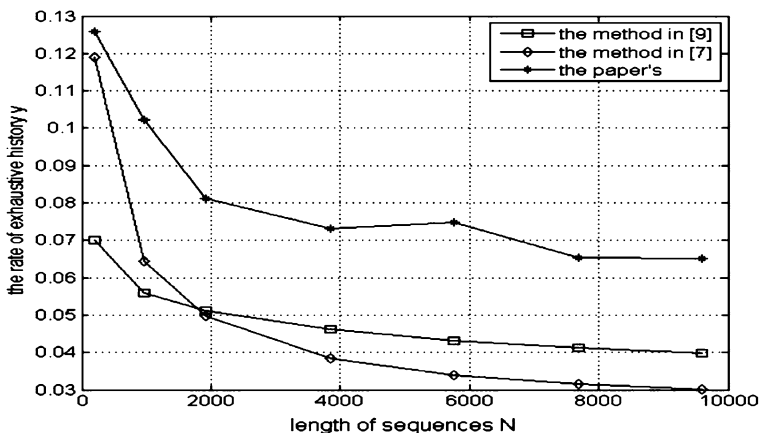


Fig. 131.2 The sequences' random-like property

The simulation is performed in MATLAB7.9, where initial value x_0 of logistic map iteration is a random number between (0, 1) and frequency slot number q is 64. Sequence length is (192, 960, 1,920, 3,840, 5,760, 7,680, 9,600) respectively.

From Fig. 131.2, it can be found that compared to sequences processed by method mentioned in paper [7] and without wide interval process, sequence processed by method in this chapter has larger rate of exhaustive history. Moreover, the proposed method has more steps of exhaustive history and better random-like property in condition of the same length. So, it can be concluded that the sequence constructed by the proposed method is a chaotic sequence with better random-like property and better security performance.

131.3 Performance Analysis and Simulation Experiment

Generally, the performance analysis of FH sequences contains following aspects: given minimum gap, hamming correlation property, balance property and bit error rate. Then, we will analysis the performance of FH sequences generated by the proposed algorithm from above several aspects. The drop algorithm has a lower generating efficiency because it must drop some points, we do not consider it in the following discussion just as there is few researches all the time. The symmetry band algorithm has worse hamming correlation and other properties than shift replace [12]. So, this chapter mainly compares the proposed method with the shift algorithm [7] and the non-treatment interval in [9].

The following simulation is preformed in MATLAB7.9. The parameters involved are: initial value $x_0 = 0.3555$, the number of frequency point $q = 64$, the length of sequences $N = L * q$ (L is the average number of any frequency point

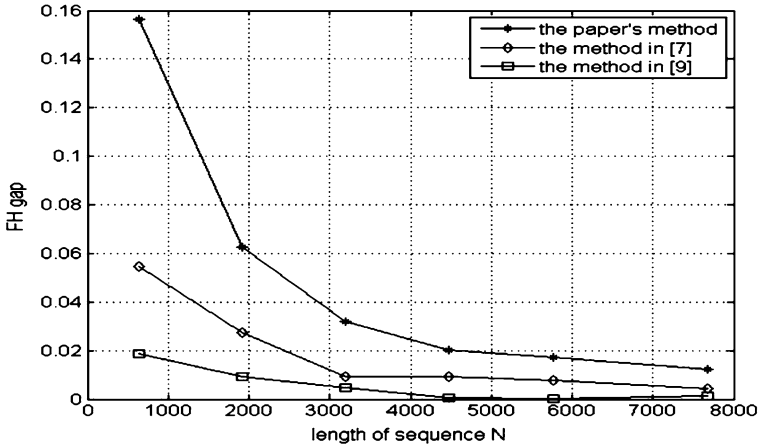


Fig. 131.3 FH gap of different sequences

appeared in the sequences). In the simulation of BER, we choose additive white gauss noise (AWGN) channel and binary phase shift keying (BPSK) modulation.

131.3.1 Frequency Hopping Gap

The interval between two carrier frequencies or between current frequency and last frequency is FH gap. The larger the FH gap is, the better the performance of anti narrowband interference, broadband blocking interference, tracking interference and multipath fading.

In chaos FH communication systems, we define the FH gap as:

$$Cap = \frac{1}{N} \sum_{i=0}^{N-1} |X_{i+1} - X_i| \quad (131.3)$$

We have analyzed that FH gap of FH sequences based on the algorithm in this chapter is more than or equal to q in theory. The FH gap of the three algorithms is compared in Fig. 131.3.

It is obvious that the proposed method is superior to the other methods, about three times of that of sequences base on [9], which is consistent with the theoretical analysis. It can be confirmed that the proposed method can improve communication systems' performance such as anti-jamming and anti-multipath fading. The simulation result indicates that the proposed method can meet the requirements, since our main target is to improve FH gap.

131.3.2 Hamming Correlation Property

In communication systems, user's interference will burst when they use same frequency. It is important to study FH sequences' periodic hamming correlation function. Periodic hamming correlation function contains autocorrelation function and crosscorrelation function, expressed apart as:

$$\begin{aligned}
 H_{xy}(\tau) &= \sum_{i=0}^{N-1} h(X(i), Y(i + \tau)) \\
 Y(i + \tau) &= Y((i + \tau) \bmod N) \\
 h(X(i), Y(i + \tau)) &= \begin{cases} 1 \dots X(i) = Y(i + \tau) \\ 0 \dots \text{other} \end{cases}
 \end{aligned} \tag{131.4}$$

$$\begin{aligned}
 H_{xx} &= \sum_{i=0}^{N-1} h(X(i), X(i + \tau)) \\
 X(i + \tau) &= X((i + \tau) \bmod N) \\
 h(X(i), X(i + \tau)) &= \begin{cases} 1 \dots X(i) = X(i + \tau) \\ 0 \dots \text{other} \end{cases}
 \end{aligned} \tag{131.5}$$

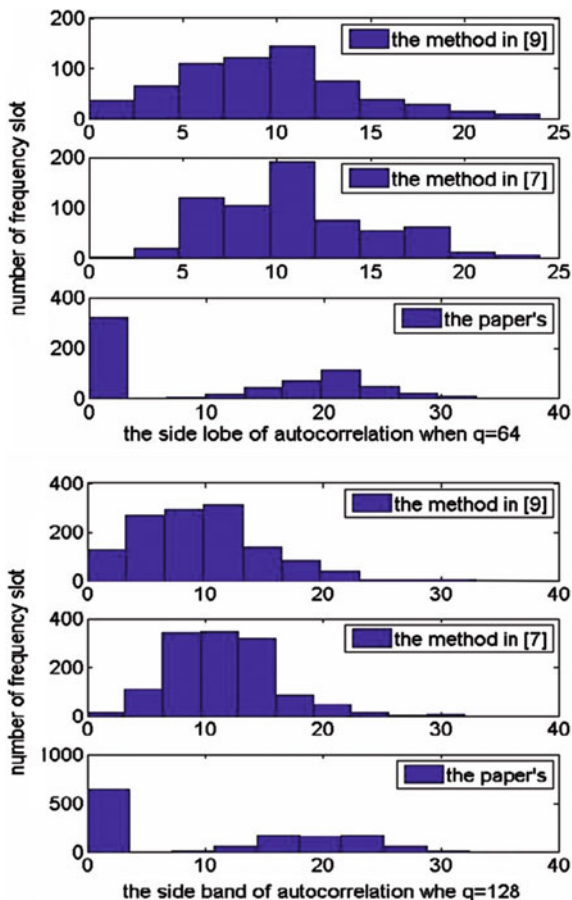
Autocorrelation mainly effects anti-multipath interference of communication systems, and crosscorrelation mainly decides the number of collision between any two sequences in sequences set with any relative time delay. The theoretical average value can be described by [9].

$$\begin{aligned}
 \langle H_{xx} \rangle &= \frac{1}{N-1} \sum_{\tau=1}^{N-1} H_{xx}(\tau) \approx N/q \\
 \langle H_{xy} \rangle &= \frac{1}{N} \sum_{\tau=1}^{N-1} H_{xy}(\tau) \approx N/q
 \end{aligned} \tag{131.6}$$

For FH sequences base on the proposed method, the frequency slot of sequences must be alternately locate in every frequency band, because of current and next frequency slot are in two different frequency bands. It can be observed that when time delay is odd, autocorrelation H_{xx} is 0, and crosscorrelation H_{xy} is also 0. So, half of their side bands of autocorrelation and crosscorrelation will be zero. Figures. 131.4 and 131.5 are the side lobe distributions of autocorrelation and crosscorrelation of FH sequences. The initial values of logistic map are $x_0 = 0.3555$, $y_0 = 0.6111$, frequency slot number of q is separately 64 and 128; length of N is 640 and 1,280 correspondingly.

It can be seen from Figs. 131.4 and 131.5 that half of sequences are $H_{xx} = 0$ and $H_{xy} = 0$ regardless of q , this result is consistent with the theoretical analysis. When q is 64, the sequences constructed by the proposed method have the greater maximum autocorrelation side lobe than others, and maximum crosscorrelation

Fig. 131.4 The autocorrelation side lobe distribution of sequences



side lobe is equal to sequence constructed by the algorithm in [9], slightly smaller than sequence in [7]. When q is 128, the proposed method attains lower maximum autocorrelation side lobe than the scheme in [9], but higher than that of sequence in [7]. The largest crosscorrelaton side lobe is slightly smaller than that in [9], while a little larger than that in [7]. It indicates that the hamming correlation is not very good when the length of sequence and the number of frequency slot is small, but with their increase, the sequences' performance becomes better. In addition, the mean values of hamming correlation are compared in Tables 131.1 and 131.2.

It is shown that mean value of hamming correlation of the sequences designed in the paper is consistent with the theoretical analysis, a little better than that in [7]. On the other hand, to achieve the performance in [9], we need increase the sequence length and frequency slot in the number. So the sequences designed by the proposed method have good hamming correlation properties.

Fig. 131.5 The crosscorrelation side lobe distribution of sequences

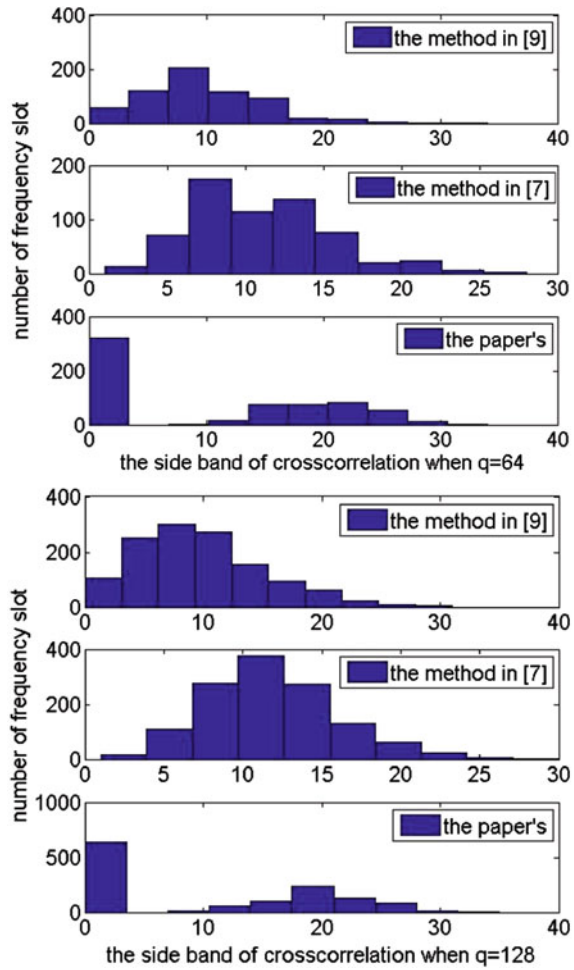


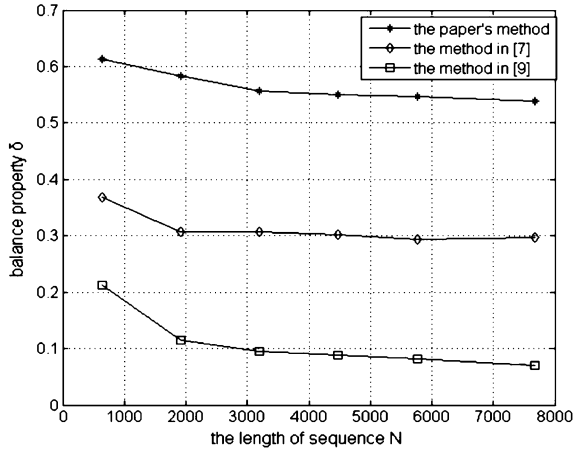
Table 131.1 The mean value of autocorrelation

	$L = 10, q = 64$	$L = 10, q = 128$	$L = 20, q = 128$
The method in [9]	9.5813	9.7312	19.8695
The method in [7]	11.0531	11.5234	23.4617
The paper's method	10.1281	9.7891	19.7738

Table 131.2 The mean value of crosscorrelation

	$L = 10, q = 64$	$L = 10, q = 128$	$L = 20, q = 128$
The method in [9]	9.8797	9.8242	20.0070
The method in [7]	11.0828	11.3406	23.4766
The paper's method	11.1281	10.7891	20.8422

Fig. 131.6 The balance property



131.3.3 Balance Property

Balance property is the performance that mainly determines the uniformity of every frequency slot in a finite length sequence, generally defined as δ , smaller δ means that the sequences' balance property is better. δ is usually expressed as:

$$\delta = \frac{q}{N} \sqrt{\frac{1}{q} \sum_0^{q-1} \left(f_i - \frac{N}{q} \right)^2} \tag{131.7}$$

Sometimes for simplicity, we also define δ as:

$$\delta = |f_i - L|/L \tag{131.8}$$

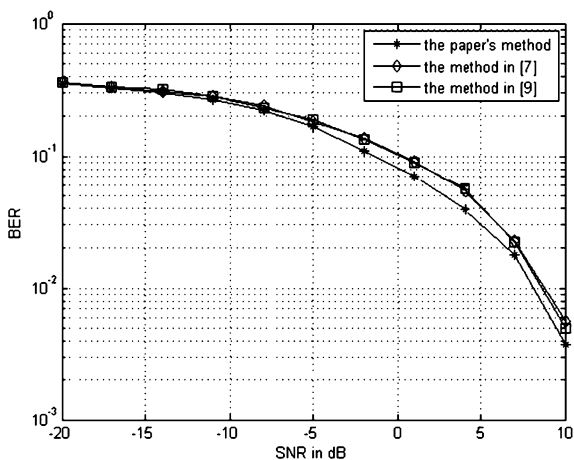
where f_i is the actual times the number of i frequency slot occurs in FH sequences. The simulation results for balance property are shown in Fig. 131.6.

It is shown that the balance property of sequences base on the proposed method decreased relative to other methods. The main reason is that the proposed method makes the original sequence distribute in two bands by adding the prefix. Therefore the uniformity of sequences degrades in whole frequency band, but in their own bands, the sequence will still maintain a certain degree of independent uniform distribution.

131.3.4 Bit Error Rate

In this simulation, we use Monte Carlo method to analyze the BER of sequences, and choose AWGN channel and BPSK modulation. Letting the frequency slot

Fig. 131.7 BER performance of sequences



number $q = 32$, and sequence length $N = 1000$, the BER of sequences with respect to Signal-to-noise ratio (SNR) is shown in Fig. 131.7.

In AWGN channel, BER of this method is lower compared with other methods, at the same SNR. It can be pointed out that the performance of the frequency hopping sequences has been improved in terms of anti-jamming through adding prefix. So that by widening interval method proposed in this chapter, the performance of FH sequences is improved, the anti-jamming and anti-intercept property of communication systems are enhanced.

131.4 Conclusion

On the basis of improved logistic map, a new method for chaotic FH sequences design with given minimum gap has been developed in this chapter. Through adding fixed binary prefix to binary sequences quantified chaos map iteration, the proposed method produces two adjacent frequency slot of frequency hopping sequence naturally distributed in different frequency band and improves the FH gap effectively. The theoretical analysis and simulation results indicate that the sequences FH gap is wider than that of sequences constructed by existed methods, then the system performance, such as anti-jamming and anti-intercept capacities and anti-multipath fading are improved effectively. In addition, as there is no need to judge hopping interval before and after hopping unlike other methods when generating FH sequences, the operation efficiency and efficiency of sequence generation can be improved. The analysis of sequence random-like property indicates that the proposed method has good random-like property, which can improve the system confidential performance effectively. Thus, the chaotic FH

sequences constructed by the proposed method are suitable for chaotic frequency hopping communication and secure communication. Sponsored by the National Natural Science Foundation of China project (No. 60805039).

References

1. Mi L, Zhu Z (2004) Chaotic frequency-hopping sequences based on logistic map. *Chin J Radio Sci* 19(3):333–337
2. Liu H et al (2005) The designing and implementing of high-speed wide-gap chaotic frequency-hopper. *Acta Scientiarum Naturalium Universitatis Nankaiensis* 38(4):104–110
3. Tai N et al (2009) Design of wide-gap Chaos modulation hopping pattern based on TOD. *J Beijing Univ Posts Telecommun* 32(6):77–82
4. Ben N et al (2003) The research for generation the chaotic frequency hopping sequences based on the cosine mapping method and improvement of given gap. *Tianjin Commun Technol* 1:21–25
5. Zhang H (2010) Design and performance analysis of frequency hopping sequences with given minimum gap. *Microwave and millimeter wave technology (ICMMT)*, 2010 international conference, pp 1271–1274
6. Wu G, Wang D (2010) Design and performance investigation on frequency hopping sequence with minimum gap based on RS Code. 2010 international conference on electrical and control engineering, pp 5120–5123
7. Bo Z et al (2010) Realization of wide interval frequency hopping patterns based on chaotic code. *Commun Technol* 43(06):57–59
8. Gan L, Wu Y (1999) FH sequences with given minimum gap based on Chaos. *J China Inst Commun* 20(4):72–76
9. Ling C, Sun S (1997) Frequency-hopping sequences by the logistic map. *Acta Electron Sinica* 25(10):79–81
10. Liu X, Zhang J et al (2010) A generation method for FH sequences with given minimum gap based on chaotic dynamic quantification. *J Circuits Syst* 15(4):96–100
11. Feng M et al (2007) Method to identify the random-like property of chaotic sequences. *Comput Appl* 27(8):1952–1955
12. Hu Y, He Z (2010) Comparisons on two constructions of FH sequences with given minimum gap based on Chaos. *Commun Technol* 43(99):1–3

Chapter 132

Practical Criteria for Generalized Strictly Diagonally Dominant Matrices

Yujing Liu, Jie Xu and Li Guo

Abstract To investigate the generalized strictly diagonally dominant matrix, some properties of generalized strictly diagonally dominant matrix are given and some new sufficient conditions for generalized strictly diagonally dominant matrix are presented. Thus the corresponding results in An-qi and Rong J Acta Math Appl Sin 19(2) 401–406 2006 are generalized and improved.

Keywords Generalized strictly diagonally dominant matrix · Sufficient condition · Nonzero element chain

132.1 Introduction

Generalized strictly diagonally dominant matrices is one of the main subjects studied in computational mathematics, mathematical modelling and matrix theory and it has got widespread use in cybernetics. Seeking the decision conditions of generalized strictly diagonally dominant matrices is of great importance, which has certain study in our country. In this paper, some practical decision conditions are given, and some results in [1] are generalized and improved.

Y. Liu (✉) · L. Guo

Department of Mathematics, Beihua University, Jilin 132013, China
e-mail: liuyujing3512@163.com

J. Xu

College of Sciences, Jilin Institute of Chemical Techno, Jilin 132013, China

132.2 Definition and Lemma

Here $C^{n \times n}$ denotes the set of $n \times n$ complex square matrices, $N = 1, 2, \dots, n$, $A = (a_{ij}) \in C^{n \times n}$, $\Lambda_i(A) = \sum_{j \neq i} |a_{ij}|$, $S_i(A) = \sum_{j \neq i} |a_{ji}|$, $i, j \in N = \{1, 2, \dots, n\}$. If $|a_{ii}| > \Lambda_i(A)$, $\forall i \in N$, then A is called strictly diagonally dominant matrices, denote $A \in D$. If there is Positive diagonal matrix X , such that $AX \in D$, then A is called generalized strictly diagonally dominant matrices, denote $A \in D^*$. If there is $\alpha \in [0, 1]$, such that $|a_{ii}| > \alpha \Lambda_i(A) + (1 - \alpha)S_i(A)$, $\forall i \in N$, then A is called strictly α -diagonally dominant matrix, denote $A \in D_\alpha$. If there is Positive diagonal matrix X such that $AX \in D_\alpha$, then A is called generalized strictly α -diagonally dominant matrix, denote $A \in D_\alpha^*$.

Definition 1 Let $A = (a_{ij}) \in C^{n \times n}$, $\alpha \in [0, 1]$. if $\forall i \in N$, we have

$$|a_{ii}| \geq \alpha \Lambda_i(A) + (1 - \alpha)S_i(A);$$

If

$$\forall i \in I = \{i \in N \mid |a_{ii}| = \alpha \Lambda_i(A) + (1 - \alpha)S_i(A)\}.$$

we have nonzero element chain $a_{ii_1}, a_{i_1 i_2}, \dots, a_{i_{p-1} i_p}$, such that $j \in (N \setminus I) \neq \phi$. Then A is called with nonzero element chain α -diagonally dominant matrix.

This chapter introduces the following notation: when $\alpha \in [0, 1]$, we denote

$$N_1 = \{i \in N \mid 0 < |a_{ii}| \leq \alpha \Lambda_i(A) + (1 - \alpha)S_i(A)\}$$

$$N_2 = \{i \in N \mid |a_{ii}| > \alpha \Lambda_i(A) + (1 - \alpha)S_i(A)\}$$

$$r = \max_{i \in N_2} \left\{ \frac{\alpha \Lambda_i(A) + (1 - \alpha)S_i(A)}{|a_{ii}|} \right\},$$

$$\delta_i = \frac{\alpha \left(\sum_{\substack{j \in N_1 \\ j \neq i}} |a_{ij}| + r \sum_{\substack{j \in N_2 \\ j \neq i}} |a_{ij}| \right) + (1 - \alpha)S_i(A)}{|a_{ii}|}, \quad \forall i \in N,$$

$$N_1^{(1)} = \left\{ i \in N_1 \mid 0 < |a_{ii}| \leq \alpha \left(\sum_{t \in N_1} |a_{it}| + \sum_{t \in N_2} |a_{it}| \delta_t \right) + (1 - \alpha)S_i(A) \right\},$$

$$N_2^{(1)} = \left\{ i \in N_1 \mid |a_{ii}| > \alpha \left(\sum_{t \in N_1} |a_{it}| + \sum_{t \in N_2} |a_{it}| \delta_t \right) + (1 - \alpha)S_i(A) \right\}.$$

If $N_1 = \phi$, then $A \in D_\alpha$, it is easy to see $A \in D_\alpha^*$, furthermore, we get $A \in D^*$; if $N_2 = \phi$, then $A \notin D^*$. So we suppose $N_1, N_2 \neq \phi$, frequently. For simplicity,

denote $\Lambda_i = \Lambda_i(A)$, $S_i = S_i(A)$. we stipulate Λ_i and S_i are not zero at the same time.

Lemma 1 [3] *Let $A = (a_{ij}) \in C^{n \times n}$, If A has nonzero element chain α -diagonally dominant matrix, then $A \in D^*$.*

132.3 Main Result

Theorem 1 *Let $A = (a_{ij}) \in C^{n \times n}$, $\alpha \in [0, 1]$. If $\forall i \in N_1, \forall j \in N_2$, we have*

$$\left[|a_{ii}| - \alpha \sum_{\substack{t \in N_1 \\ t \neq i}} |a_{it}| - (1 - \alpha)S_i \right] \left[|a_{jj}|\delta_j - \alpha \sum_{\substack{t \in N_2 \\ t \neq j}} |a_{jt}|\delta_t - (1 - \alpha)S_j\delta_j \right] > \alpha^2 \left(\sum_{t \in N} |a_{jt}|\delta_t \right) \sum_{k \in N} |a_{jk}| \tag{131.1}$$

then $A \in D^*$.

Proof If there is $i \in N_1$, make $\sum_{t \in N_2} |a_{it}| = 0$, then know, the right-hand side of formula (132.1) equals zero. Thus, we know, the left-hand side of formula (132.1) is greater than zero and have

$$|a_{jj}| - \alpha \sum_{\substack{t \in N \\ t \neq j}} |a_{jt}| - (1 - \alpha)s_j \leq 0, \tag{132.2}$$

Since $\forall j \in N_2$ get

$$|a_{jj}|\delta_j = \alpha \left(\sum_{k \in N} |a_{jk}| + r \sum_{\substack{k \in N \\ k \neq j}} |a_{jk}| \right) + (1 - \alpha)s_j, \tag{132.3}$$

Because $r \geq \delta_k$ ($k \in N_2$), $\delta_j < 1$ ($j \in N_2$), then as formula (132.3) we obtain

$$|a_{jj}|\delta_j - \alpha \sum_{\substack{k \in N_2 \\ k \neq j}} |a_{jk}|\delta_k - (1 - \alpha)s_j\delta_j \geq 0,$$

By the left-hand side of formula (132.1) is greater than zero, we know

$$|a_{ii}| - \alpha \sum_{\substack{t \in N_1 \\ t \neq i}} |a_{it}| - (1 - \alpha)s_i > 0,$$

have a contradiction with formula (132.2).

If

$$|a_{jj}|\delta_j - \alpha \sum_{\substack{k \in N_2 \\ k \neq j}} |a_{jk}|\delta_k - (1 - \alpha)s_j\delta_j = 0,$$

Then know the left-hand side of formula (132.1) equals zero, but it is greater than zero, a contradiction.

To sum up, $\forall i \in N_1$ have $\sum_{t \in N_2} |a_{it}| \neq 0$.

Then formula (132.1) tells $d > 0$, let $\forall i \in N_1, \forall j \in N_2$, we get

$$\frac{|a_{ii}| - \alpha \sum_{\substack{t \in N_1 \\ t \neq i}} |a_{it}| - (1 - \alpha)s_i}{\alpha \sum_{t \in N_2} |a_{it}|\delta_t} > d > \frac{\alpha \sum_{t \in N_1} |a_{jk}|}{|a_{ij}|\delta_j - \alpha \sum_{\substack{k \in N_2 \\ k \neq i}} |a_{it}|\delta_k - (1 - \alpha)s_j\delta_j}, \tag{132.4}$$

Make a plus dominant matrix $D = \text{diag}(x_1, x_2, \dots, x_n)$, in which

$$x_i = \begin{cases} 1, & i \in N_1, \\ d\delta_i & i \in N_2. \end{cases}$$

Let $B = AD = (b_{ij}) \in C^{n \times n}$, and denote $\sum_{j \neq i} |b_{ij}| = \Lambda_i(B), \sum_{j \neq i} |b_{ji}| = S_i(B)$.

$\forall i \in N_1$, according to formula (132.4), we know

$$|a_{ii}| > \alpha \sum_{\substack{t \in N_1 \\ t \neq i}} |a_{it}| + \alpha \sum_{t \in N_2} |a_{it}|d\delta_t + (1 - \alpha)S_i,$$

Then

$$|b_{ii}| > \alpha \left(\sum_{\substack{t \in N_1 \\ t \neq i}} |b_{it}| + \sum_{t \in N_2} |b_{it}| \right) + (1 - \alpha)S_i(B),$$

That is

$$|b_{ij}| > \alpha \Lambda_j(B) + (1 - \alpha)S_j(B).$$

To sum up, $\forall i \in N$, there is $|b_{ii}| > \alpha \Lambda_i(B) + (1 - \alpha)S_i(B)$, from [2] we know, $B \in D^*$, then we get $A \in D^*$.

So it is clear to see Theorem 1 here improves Theorem 1 in [1].

Theorem 2 Let $A = (a_{ij}) \in C^{n \times n}, \alpha \in [0, 1]$. If $\forall i \in N_1, \forall j \in N_2$, we have

$$\begin{aligned} & \left[|a_{ii}| - \alpha \sum_{\substack{t \in N_1 \\ t \neq i}} |a_{it}| - (1 - \alpha)S_i \right] \left[|a_{jj}|\delta_j - \alpha \sum_{\substack{k \in N_2 \\ k \neq j}} |a_{jk}|\delta_k - (1 - \alpha)S_j\delta_j \right] \\ & \geq \alpha^2 \left(\sum_{t \in N_2} |a_{it}|\delta_t \right) \sum_{k \in N_1} |a_{jk}|, \\ J(A) &= \left\{ j \in N_2 \mid \left[|a_{ii}| - \alpha \sum_{t \in N_1, t \neq i} |a_{it}| - (1 - \alpha)S_i \right] \right. \\ & \quad \left[|a_{jj}|\delta_j - \alpha \sum_{k \in N_2, k \neq j} |a_{jk}|\delta_k - (1 - \alpha)S_j\delta_j \right] \\ & \quad \left. > \alpha^2 \left(\sum_{t \in N_2} |a_{it}|\delta_t \right) \sum_{k \in N_1} |a_{jk}|, \forall i \in N_1 \right\} \neq \phi, \end{aligned}$$

And $\forall i \in N_1 \cup (N_2 \setminus J(A))$ have nonzero element chain $a_{i1}, a_{i1i_2}, \dots, a_{i1j}$, such that $j \in J(A)$, then $A \in D^*$.

Proof Because $J(A) \neq \emptyset$, then imitate Theorem 1,

For $\forall i \in N_1$ there is

$$|b_{ii}| \geq \alpha \Lambda_i(B) + (1 - \alpha)S_i(B).$$

For $\forall j \in N_2 \setminus J(A)$ there is

$$|b_{jj}| \geq \alpha \Lambda_j(B) + (1 - \alpha)S_j(B).$$

From the condition set $J(A)$, we know $\forall j \in J(A)$, we get

$$|b_{jj}| > \alpha \Lambda_j(B) + (1 - \alpha)S_j(B).$$

Because multiply matrix A by positive diagonal matrix from right is not change nonzero property of the element, we know B is nonzero element chain diagonally dominant matrix, then from [2] we know $B \in D^*$, furthermore, $A \in D^*$.

So it is clear to see Theorem 2 here improves Theorem 2 in [1].

Inference Let $A = (a_{ij}) \in C^{n \times n}$ is an irreducible matrix, $\alpha \in [0, 1]$, if $\forall i \in N_1, \forall j \in N_2$, there is

$$\begin{aligned} & \left[|a_{ii}| - \alpha \sum_{\substack{t \in N_1 \\ t \neq i}} |a_{it}| - (1 - \alpha)S_i \right] \left[|a_{jj}|\delta_j - \alpha \sum_{\substack{k \in N_2 \\ k \neq j}} |a_{jk}|\delta_k - (1 - \alpha)S_j\delta_j \right] \\ & \geq \alpha^2 \left(\sum_{t \in N_2} |a_{it}|\delta_t \right) \sum_{k \in N_1} |a_{jk}|, \end{aligned}$$

And in formula (132.5) at least one strictly unequal formula holds, then $A \in D^*$.

Theorem 3 Let $A = (a_{ij}) \in C^{n \times n}$, $\alpha \in [0, 1]$. If $\forall i \in N_1^{(1)}$, $\forall j \in N_2^{(1)}$, we have

$$\begin{aligned} & \left[|a_{ii}| - \left(\alpha \sum_{t \in N_1^{(1)}} |a_{it}| + \sum_{t \in N_2} |a_{it}| \delta_t \right) - (1 - \alpha)S_i \right] \left[|a_{jj}| - \alpha \sum_{\substack{k \in N_2^{(1)} \\ k \neq j}} |a_{jk}| - (1 - \alpha)S_j \right] \\ & \geq \alpha^2 \sum_{t \in N_2^{(1)}} |a_{it}| \left(\sum_{k \in N_1^{(1)}} |a_{jk}| + \sum_{k \in N_2^{(1)}} |a_{jk}| \delta_k \right) \end{aligned} \tag{132.6}$$

then $A \in D^*$.

Proof If there is $i \in N_1^{(1)}$, such that $\sum_{t \in N_2^{(1)}} |a_{it}| = 0$, then we know the right-hand of formula (132.6) equals to zero, from this we get the left-side of it is greater than zero as $\forall j \in N_2^{(1)}$, there is

$$|a_{jj}| - \alpha \sum_{k \in N_2^{(1)}} |a_{jk}| - (1 - \alpha)S_j > 0,$$

Then,

$$|a_{ii}| - \alpha \left(\sum_{\substack{t \in N_1^{(1)} \\ t \neq i}} |a_{it}| + \sum_{t \in N_2} |a_{it}| \delta_t \right) - (1 - \alpha)S_i > 0.$$

And as $\sum_{t \in N_2^{(1)}} |a_{it}| = 0$ ($i \in N_1^{(1)}$), then

$$|a_{ii}| - \alpha \left(\sum_{\substack{t \in N_1^{(1)} \\ t \neq i}} |a_{it}| + \sum_{t \in N_2^{(1)}} |a_{it}| + \sum_{t \in N_2} |a_{it}| \delta_t \right) - (1 - \alpha)S_i > 0,$$

That is,

$$|a_{ii}| > \alpha \left(\sum_{t \in N_1} |a_{it}| + \sum_{t \in N_2} |a_{it}| \delta_t \right) + (1 - \alpha)S_i,$$

but $i \in N_1^{(1)}$ a contradiction, so $\forall i \in N_1^{(1)}$ there is $\sum_{t \in N_2^{(1)}} |a_{it}| \neq 0$. Formula (132.8)

exists condition $1 > d > 0$, make $\forall i \in N_1^{(1)}$, $\forall j \in N_2^{(1)}$, we get

$$\frac{|a_{ii}| - \alpha \left(\sum_{t \in N_1^{(1)}, t \neq i} |a_{it}| + \sum_{t \in N_2} |a_{it}| \delta_t \right) - (1 - \alpha)S_i}{\alpha \sum_{t \in N_2^{(1)}} |a_{it}|} > d > \frac{\alpha \left(\sum_{t \in N_1^{(1)}} |a_{jk}| + \sum_{t \in N_2} |a_{jk}| \delta_k \right)}{|a_{jj}| - \alpha \sum_{k \in N_2^{(1)}} |a_{it}| - (1 - \alpha)S_j} \quad (132.7)$$

We make a positive diagonal matrix $D = \text{diag}(x_1, x_2, \dots, x_n)$, where

$$x_i = \begin{cases} 1, & i \in N_1^{(1)}, \\ d, & i \in N_2^{(1)}, \\ \delta_i, & i \in N_2. \end{cases}$$

Let $B = AD = (b_{ij}) \in C^{n \times n}$, then

$\forall i \in N_1^{(1)}$, according to formula (132.7), we get

$$|a_{ii}| > \alpha \left(\sum_{\substack{t \in N_1^{(1)} \\ t \neq i}} |a_{it}| + \sum_{t \in N_2^{(1)}} |a_{it}|d + \sum_{t \in N_2} |a_{it}| \delta_t \right) + (1 - \alpha)S_i$$

That is

$$|b_{ij}| > \alpha \Lambda_i(B) + (1 - \alpha)S_i(B).$$

$\forall j \in N_2^{(1)}$, according to formula (132.7), we get

$$|a_{jj}|d > \alpha \left(\sum_{k \in N_1^{(1)}} |a_{jk}| + \sum_{\substack{k \in N_2^{(1)} \\ k \neq j}} |a_{jk}|d + \sum_{k \in N_2} |a_{jk}| \delta_k \right) + (1 - \alpha)S_jd$$

That is

$$|b_{jj}| > \alpha \Lambda_j(B) + (1 - \alpha)S_j(B).$$

$\forall k \in N_2$, from the condition set δ_k , we know

$$|a_{kk}|\delta_k = \alpha \left(\sum_{t \in N_1} |a_{kt}| + r \sum_{\substack{t \in N_2 \\ t \neq k}} |a_{kt}| \right) + (1 - \alpha)S_k, \tag{132.8}$$

Then as $0 < d < 1$, $r \geq \delta_k$ ($k \in N_2$), $\delta_k < 1$ ($k \in N_2$).

Case 1 If $S_k \neq 0$, then we get $S_k > S_k \delta_k$, so from formula (132.8), we get

$$|a_{kk}|\delta_k > \alpha \left(\sum_{k \in N_1^{(1)}} |a_{kt}| + \sum_{t \in N_2^{(1)}} |a_{kt}| + \sum_{\substack{t \in N_2 \\ t \neq k}} |a_{kt}|\delta_t \right) + (1 - \alpha)S_k \delta_k, \tag{132.9}$$

That is,

$$|b_{kk}| > \alpha \Lambda_k(B) + (1 - \alpha)S_k(B).$$

Case 2 If $S_k = 0$, while $\sum_{t \in N_2} |a_{kt}| \neq 0$. we know $r > \delta_t$ ($t \in N_2$), that is,

$$r \sum_{\substack{t \in N_2 \\ t \neq k}} |a_{kt}| > \sum_{\substack{t \in N_2 \\ t \neq k}} |a_{kt}|\delta_t,$$

then formula (132.9) still holds, so $|b_{kk}| > \alpha \Lambda_k(B) + (1 - \alpha)S_k(B)$.

$$\text{If } S_k = 0, \sum_{t \in N_2} |a_{kt}| = 0.$$

for Λ_i and S_i are not zero at the same time, we get

$$\sum_{t \in N_1} |a_{kt}| \neq 0,$$

that is, $a_{kt} \neq 0$,

let $t \in N_1$ we know $b_{kt} \neq 0$, $t \in N_1$, imitate the proof above again, we obtain $|b_{tt}| > \alpha \Lambda_t(B) + (1 - \alpha)S_t(B)$.

To sum up, B is either strictly Diagonally Dominant Matrices or have nonzero element chain α -Diagonally Dominant Matrices, so from paper [2] and paper [3], we get $B \in D^*$, furthermore, $A \in D^*$.

References

1. An-qi H, Rong H (2006) Some criteria method for Generalized Strictly Diagonally Dominant Matrices. *J Acta Math Appl Sinic* 19(2):401–406
2. Yu-xiang S (1997) Sufficient condition for generalized strictly diagonally dominant matrices, numerical mathematics. *J Chin Univ* 19(3):216–223
3. Neumann M (1979) A note on generalization of strict diagonal dominance for real matrices. *J Lin Alg Appl* 26:3–14

Chapter 133

Research on New Rural Information Service Model of Agricultural Industrial Chain

Hai-han Yang, Xing-xia Shuai, Chun-hua Ju and Tie-zhu Zhang

Abstract The faults of agricultural industry chain are the key problem of our country, and the key to solve this problem is to turn the existing traditional production mode to the new industrial mode of production. Based on the analysis of current situation, it studies the informational service mode which caters to integrating agricultural industrial chain, with a method of starting from modern information technology. By integrating a variety of interactive model, trying to combine farmers' needs with the platform closely, integrate and extend the agricultural industry chain and pour the information-based establishments into produce antenatal, produce medium, postpartum sales of agricultural products. Thereby promote the development of agriculture.

Keywords Information service model · Integrating · Agricultural industry chain · Information technology

133.1 Introduction

Agriculture, as a basic industry to solve people's basic living needs, holds an irreplaceable leading position in the national economy development. The information construction in rural areas abroad start early, and the development is relatively sound and perfect. It is indicated in the Party's agricultural report that the nature of agricultural development is to develop the industrial chain. Agricultural industry chain originated from the United States in the 1850s, and now

H. Yang (✉) · X. Shuai · C. Ju · T. Zhang
College of Computer and Information Engineering, Zhejiang Gongshang University,
Hangzhou 310018, China

America opens up the agricultural chain in the use of information technology, while in China, the model of industry chain in agriculture is still in the initial stage because of management of agricultural households. Jinshan Lin proposed that agricultural chain is a network structure which is made up of industry groups and closely related to primary products, including scientific research which is prepare for agriculture production, prophase industry department such as agricultural means of production etc., middle industry department such as crop planting and livestock breeding etc., later industry department such as processing and storage, transportation and sales etc. whose raw material is agriculture product [1].

The text, from integrating agricultural industrial chain standpoint, dwells on in the use of modern information technology to build agricultural service model, integrate and extend the agricultural industry chain, combine the peasant economy with the market economy to form a strong vitality whole.

133.2 Related Articles

The agriculture takes the weak trend industry in our country, which has obvious Chinese characteristic. Scholars like Han Jun thought that the feeble constitution of agriculture has decided the difficulty of farmers' additional receiving. Especially the natural weather have a big effect on agricultural production as well as the pest problem, but our countryside's infrastructure is also unable to prevent the loss of natural factor bringing by natural factors [2]. Zhang et al. [3] and some other scholars thought that there is double lemon market existing in the agricultural industry chain, and the main question of it is the asymmetry of market's body status. In addition, they also pointed out that our country agriculture industry chain has the market's and government's double malfunction phenomenon; Chinese agricultural industry chain can not achieve the goal of industry concentration and sustainable development of agriculture through normal market competition. At the same time the government regulation and other reasons cause market failure with government failure, which makes the whole industry chain into deeper and deeper trouble [4]. Wang [5] in the "Construction of the agricultural industrial chain " has pointed out the major problem of Chinese agricultural industry chain is for the agricultural industry chain being narrow and short, and the existence of the agricultural industrial chain faults as well as poor connection.

In the final analysis, the biggest problem of Chinese agricultural industry is the industry- strand breaks, and the lack of a guarantee of sustainable development. The problem can be grouped into three areas: (1) Industrial chain's information and market players are asymmetrical. In the agricultural industrial chain, the profit is large in two ends, but small in the middle. Farmers who are in the middle of industrial chain are in the unprofitable position. Occlusion of market information leads to farmers in the weak position in the entire agricultural industry. (2) The feeble constitution of agriculture makes farmers troubled by pests and diseases. The quality of farmers can not solve the increasingly complex problems of pests

and natural disasters, which leads to a lot of loss. (3) Agricultural industrial chain is narrow, having no close contact. The situation of China's rural industry is single on the whole and does not exist cycle, and almost all the industries are not complete, so there is a serious fault phenomenon in the rural.

Therefore, the direction of our research is how to integrate the industrial chain, then realize the effective communication between farmers, markets, business and other main body in the end of industrial chain. Meanwhile, the extension of agricultural industry chain will promote the communication between rural and urban areas and Link tertiary industry, which will become one of important ways to make contributions to the construction of new socialist countryside [6].

133.3 Agricultural Industrial Chain for the New Service Model

133.3.1 Come Up with Model

The integration of the agricultural industrial chain is the core of the agricultural industrial chain management which is realized by strengthening the communication and coordination among the information, values and logistics of upstream and downstream of agricultural products, stretching or strengthening the industrial chain to increase the added value and competitiveness of agricultural products. China is a traditional peasant society—small-scale agricultural production, weak capital in Human and material and agriculture sporadic. Because the problem of lacking appropriate method to process information leads to the phenomenon of poor information, which may restrict the development of agriculture in China. So how to combine household small peasant economy with modernization market economy by integrating modernized industrial chain, how to integrate production factors of modern agriculture into Chinese traditional small peasant economy and how to build agricultural organization form which suits Chinese characteristics modernization agricultural industrial chain, which are key issues in Chinese nowadays agricultural process. It is a pressing need to form around a central circle of agricultural development by putting the scattered farmers together.

In order to make all aspects of the industry chain form a whole and improve the efficiency of conformity and coordination, it is required to achieve the sharing of information and resources. The development of the Internet provides the opportunity to achieve information without gap. Domestic scholars generally believe it is the comprehensive information service platforms that will be the development mainstream of rural information services to provide a variety of information services [7] and meet the needs of various types of rural information by creating various forms of rural information platforms. Based on this, we come up with an information service model on the idea of integrating industrial chain, which integrates a wide range of interactive mode. Through the functional services of

information, it can satisfy the needs of information and services at all stages, thus supporting the agricultural industry chain.

Agricultural chain integration, which is through the strengthening of agricultural production, supply or tissue between upper and lower reaches of agricultural products, information, values and communication and coordination of logistics, stretches or strengthens the industrial chain and increase the added value and competitiveness of agricultural products. Agricultural industrial chain integration is the core of the agricultural industrial chain management. China is a traditional peasant society—small-scale agricultural production, weak in the farmers' physical and human capital, scattered state of agriculture. So how to combine family-run small peasant economy with the modern market economy by means of integrating the industrial chain, how to implant production factors of modern agriculture into Chinese traditional small peasant economy by chain, and how to build organization forms which is suitable for modern agriculture with Chinese characteristics, which are key questions of agriculture at present in China. It is imperious demands to get the dispersed household together and then formed around a central circle of agricultural development

133.3.2 Model Structure

In order to meet the needs of the farmers' daily production and sales, we come up with a countryside oriented model. This model has a variety of means of communication and conforms to the actual characteristics and habits of Chinese farmers. It creates a simple and convenient communication channel that can transmit the information to the target group effectively. At the same time, the integrative service function can solve the problems of farmers in agricultural activities completely [8]. Its purpose is to integrate industry chain, to raise farmers' production efficiency and expand the interests of farmers. At last, it can raise the farmers' position in the industry chain and make the entire agriculture developing orderly, organized and vitality (Fig. 133.1).

133.4 The Key Points of the New Information Service Mode

133.4.1 The Interaction of Industry Chain

First, only by use various kinds of interactive way can we transmit and receive information conveniently and accurately. So we can use information platform connecting farmers and market.

In China, the penetration rate of rural areas is far behind cities and towns, it is still at the primary stage. In contrast to the internet, the rate of telephone is relatively high, especially mobile phone. According to China's Internet report

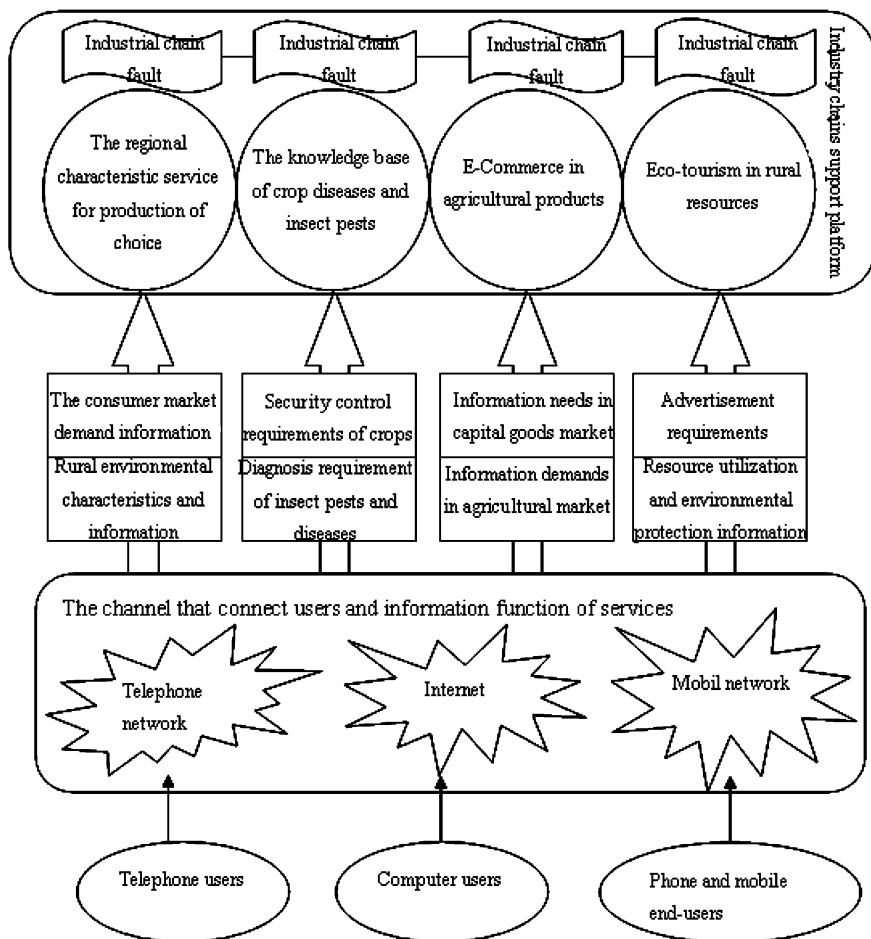


Fig. 133.1 We can see that interactive model of mobile network, telephone network and internet is the channel which connecting users and information services. We can connect faulted industry chain by building four service function of informationization and make it as entirety

shows, in 2009, the user scale between mobile phone and fixed phone have already pass more than 400 million, every 100 people has 56.3 mobile phones, while has 23.6 fixed phone.

Mobile phone and telephone have become the mainstream media, especially mobile phone. With the development of technology, more and more cheaper and practical mobile phone step into the country, so the development of 3G and integration of mobile phone’s function is the trend for future. This model is based on the triple network and focus on the development of mobile phone network; it set the short MMS send, cell phone video conversations and wireless Internet voice functions as a entirety. Implementation of mobile phones closely combined with the platform.

(1) MMS sending. We use RSS Subscription function to send information which farmers are interested into users by binding the cell phones and information platform. When the crop productions meet insect or disease problems, we can take pictures by using mobile phone and send to the platform. In this way, the staff can make preliminary judgment and get the key words according to certain characteristics. Through those keywords, we can search plant diseases in insect database or consult to experts. At last solve those problems.

(2) Voice and video dialogue. With the development of 3G networks, video conversation using a mobile phone will become possible. Joining the video conversation let experts and farmers across the space and the distance to face to face communication. The advantage of this communication is that they can directly transfer it to the experts to view the screen when farmers meet the actual situation of the scene. It just like that the experts are at the scene. This can greatly improve the efficiency of solving the problem.

(3) Wireless Internet access. Although Internet penetration rate in rural China is far below the town, the mobile Internet make up for this shortcoming. Farmers keep abreast of the “three rural” dynamic and make the information absolute impartiality and openness via mobile phone to access the wireless WAP network platform.

Phone is one of the commonly used medium in rural areas. As Some users without mobile phone or with mobile phone just to call someone, building the telephone network by the platform meets their access to information and communication needs. The function of phone, as well as the telephone’s speech function, consists of sound call and staff service. The sound call is a treatment regimen which auto-calls expert system knowledge base by IVR process from Call Center, and it combines CTI Servicer from Call Center with traditional knowledge base Servicer and Web Servicer, to realize the function of diagnosing the agricultural pest distantly with the help of telephone network. Staff service is for the experts to give out problems solutions.

For the part farmers who have strong economy ability and high quality, they can log the Internet to visit the platform. The significance of the Internet platform is to provide the most comprehensive information and services. The Internet is the foundation and premise of the platform services.

133.4.2 The Featured Services Function of the Industrial Chain Service Center

133.4.2.1 The Application Capabilities Features of Service Function

The information service function under the mode has some features, that are targeted (for the actual needs of farmers) and composite high (integration of multiple functions of service). The specific functions are as follows:

(1) Regional special services for farmers selective product.

The main function of the service is to create an expert system which can support a guide to Pre-production of agricultural production and the pre-selection of producer. In different geographical, climatic conditions and soil conditions and so on are not the same. So they are suitable for the cultivation of different agricultural products. The expert system is based on the characteristics of rural areas and combined with the demand of future market. Its aim is to provide farmers with the guidance of professional choice of production.

(2) Diagnostic knowledge base for crop diseases pests.

This service is combined by two blocks. One part is a strong knowledge base for diseases pests. The knowledge base contains various pest problems which a wide variety of farmers in the production process encountered. It provides users with a powerful search database and is the base of pest and disease system. The other one is pest and disease expert system based on remote technology. Expert system for user interaction with the platform provides various services, including human services, intelligent voice services and so on. Through combination remote technology with knowledge base, rural households can communicate with the platform via a variety of ways. This can solve all kinds of production problems.

(3) E-commerce for agricultural products

This is a center of commercial farmers' information published. It also has rural areas as one of business services, such as consignment purchasing, network pawn, secondary trading, and convenient mobile payment services and other convenient services. It can provide farmers with agricultural inputs, agricultural products, household items and practical information on various types of rural finance. Meanwhile, the feature also has the role of the media as the transaction to facilitate the farmers with the market transaction

(4) Eco-tourism for natural resources in rural areas.

This function is to serve the rural development and utilization of tourism resources. Promoting the rational development and utilization is to provide a wide range of social activities in rural areas, so urban residents can experience country life and green landscape. Advocacy is promotion for rural image and various local agricultural industries. The aim is to promote the development of various industries in rural areas.

133.4.2.2 The Integrated Effect of Service Function on the Industrial Chains

Information service function contains many sections in the agriculture industry chain, such as before production, in production and the consumption after production. It plays a different role in different period of the industry chain.

Information service function runs through the prenatal prediction of agricultural products and the market demand information, the diagnostic render of the insect pests and diseases and postpartum sales, transportation and reprocess. It makes every section sharing the same information; what's more, it makes every section much more omnibearing and zero distance. Through information service function, informations are brought into the public, every section connected with each other much better and the cost estimated before was declined sharply. It also shortens the time of agricultural products from production to consumption. All in all, the information service function has strong scientificity. Each section of the agriculture industry chain not only mutual independence, it is also interactional. The key point of the conformity and extension of the agriculture industry chain is how to deal with the relationship of the three.

During the prenatal stage in agriculture, farmers most in need of information is what the market needs, what is suitable for planting in the local environment how to choose the kind of varieties and how to obtain means of production. Village special services and e-commerce of the agricultural service platform can resolve these issues one by one. According to the village special service model, they have access to appropriate breeds of each village, climate characteristics, demand forecasts and other information, guiding the proceed of agricultural prenatal stage and providing comprehensive resource information for the prepare for the prenatal stage. Meanwhile, to access to the latest market information by the combination of e-commerce platform and by comparing the lowest price and quality of means production provided by each company. The information which is provided by the Internet platform has a very high transparency, thus it can ensure the purchase price Reasonable for farmers. It solves the problem of the blind demand of farmers on the market in prenatal stage and reducing the production cost, and reducing overhead of its market research.

In the stage of agriculture produce medium, farmers can solve various production problems in the production process through pest and disease knowledge base, particularly the difficulty miscellaneous disease related with the pest and disease and natural disaster. Due to the current deterioration of the environment, pest and disease has become increasingly prominent and developed more complex, and the reasons of the formation is not so clear as before, sometimes it needs the authoritative experts to help to solve. Pest and disease knowledge provides online diagnosis and services. Introducing the voice services into the system and having experts, it helps farmers to solve the problem in the shortest possible time, to reduce the damage to the lowest. Meanwhile, popularizing crops health knowledge in the stage of produce medium, it helps farmers improve the basic quality of daily production and improve the situation of farmers living with weather (Fig. 133.2).

In the stage of agricultural postpartum, farmers can connect to the e-commerce platform, publish information of their own supply and access to market buy information. While start to advertise through e-commerce platform, perform the online trading services, expand the largest market for farmers, resolve the information gap and increase the transparency of agricultural markets. Network pawn of e-commerce platform can provide micro-finance for farmers, so they do the

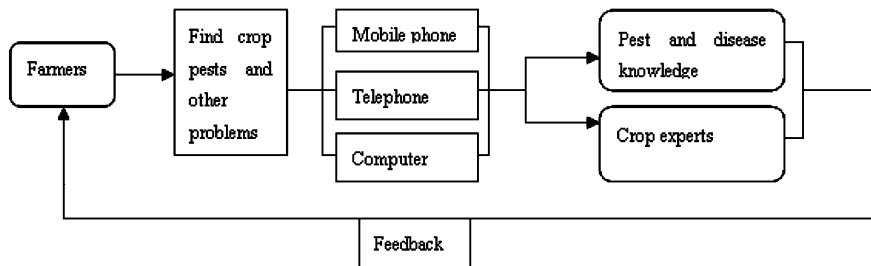


Fig. 133.2 It can be easy to know how the integrated effect of service function on the stage of agriculture produce medium through this figure

financial preparations for further expanding the sales results in the postpartum period and it helps farmers to invest effectively. Eco-tourism platform will make industrial chain root in all aspects of village. The development of eco-tourism increase the amount of tourists of village, not just a use of natural resources, but also greatly promote the development of other industries and advertise the image of the countryside, meanwhile, the rise of eco-tourism will further promote village infrastructure building and improve the traffic environment, and play a promoting role in the development of e-commerce in village areas.

133.5 Conclusion

In summary, the use of modern information technology to build a model for information services in village can effectively solve most of the problems that exist in the agricultural industrial chain. The across time and space, low cost, high-efficiency features of information services are suitable to act as the integration platform of agriculture under the current conditions. Of course, this model is not a panacea, but it can eliminate the information gap between farmers and markets as possible as it can, enhance the linkages of farmers and market, to a certain extent, provide a more accurate and forward-looking ideas for resolving the agricultural industrial chain problems.

References

1. Jinshan L (2002) Coordination market of the agricultural industrial chain: a quest [J]. Shanghai Econ Rev 3:45-55
2. Han J (2006) Western development and three rural problems [J]. Love Pap
3. Zhang L (2007) Industrial Organization and industry integration: research of China industry sustainable development—Based on the feed industry “hundred thousand “ survey engineering and cases research [J]. Manag world 4

4. Zhang L, Zhang X, Xie S (2007) Research on integration of modern agriculture industrial chain of China [J]. Teach Res 10
5. Wang J (2009) Analyses of structure of the agricultural industrial chain [J]. Econ Work 4
6. Xiangyang Q, Chunjiang Z, Baozhu Y (2006) Information technology support the main research building new countryside and applications [J]. Agric Inf Netw 12:7-8
7. Wang F Research on the model of the modern agricultural services in the perspective of Industrial chain [D]. Doctoral Dissertation, 1 June 2007
8. Wu J, Yang B, Zhang J, Zhou Z, Qu Y (2010) Research on the information service platform faced to agriculture industrial chain [J]. Comput Eng Des 31(7)

Chapter 134

Reliability Evaluation Algorithm for Distribution Power System

Jianghua Li and Juan Hu

Abstract. This paper presents a new method for evaluating distribution system. It is based on the properties of elements in fault delivering. The algorithm is suitable for evaluating reasonably complex EDNs with multiple sub feeders. It applies a forward-search-method to identifying the section controlled. First, the distribution system is simply defined into two parts called major-network and sub-network. Second, the upstream fault indices of all nodes in sub-network can be legibly calculated by utilizing the properties of elements in upstream fault delivering. Third, load comprehensive reliable indices on the major-network and downstream nodal reliable indices on the branch-appending-points are determined according to the failure mode and effect analysis (FMEA) method, and by using the properties of elements in downstream fault delivering, the upstream fault indices of all nodes in sub-network also can be calculated. At last, by synthesizing the upstream and downstream reliable indices, the reliable indices of all nodes in the sub-network can be obtained. Compared to the FMEA method, the proposal method not only has the same effect but also save computing time. The extended test system presented in this paper and the concepts presented assist in satisfying this requirement.

Keywords: Distribution system · Reliability evaluation · Elements · Fault deliver in

J. Li (✉)

Hunan Yueyang Electric Power Bureau, Yueyang, China
e-mail: ljh_01976@163.com

J. Hu

Institute of computer Hunan Institute of Science and Technology,
Yueyang, 414000 Hunan, China

134.1 Introduction

Reliability evaluation of a complete electric power system including generation, transmission, station and distribution facilities is an important ability in overall power system planning and operation [1–5]. Due to the enormity of the problem, reliability analysis is not usually conducted on a complete power systems and reliability evaluations of generating facilities, transmission systems, station configurations, and of distribution system segments are usually performed independently. Distribution system is an important part of the power system. In contrast with transmission system, it has much more kilometers, covers a broad area and makes great impact on the reliability of electricity [6–8]. Utility statistics show that distribution system failure accounts for 80% of the average customer interruptions approximately. Thus, an accurate and fast knowledge of reliability evaluation for arbitrarily structured distribution systems is crucial.

Combining with the properties of the elements in fault delivering, the unidirectional feature of the power flow in sub-network and the sharing of the search algorithm, a rapid method for the distribution network reliability assessment is presented in this paper which takes both the features of network structures and the algorithm analysis into consideration.

134.2 New Network Model for Distribution System

Figure 134.1 shows a typical radial distribution system consisting of lines, transformer and an alternative supply. Under normal operation state, the alternative supply is disconnected. Each load point is usually connected to the power system via a transformer, fuse and lateral transmission line.

The joint that connects two elements is named node. Consequently, for any radial distribution system there is an element-node network. As shown in Fig. 134.1, between nodes 0–1 is a breaker, nodes 1–2 is a transmission line, nodes 2–3 is a disconnected element, nodes 2–16 is a fuse, and nodes 16–17 is a transformer. The element-node network built above is the model for us to evaluate the reliability of distribution network in this paper.

The elements in distribution system are mainly classified into six types: transmission line, transformer, fuse, disconnect, breaker and tie switch. Under normal operation state, the main supply is connected while the alternative supply is disconnected. Suppose i which side the power flows in is the sending node of the element, at the same time, j which side the power flows out is the ending node. Under normal operation state, the structure of the distribution network is radial and the direction of power flow is single from source to loads, thus an element i can be uniquely named by its ending node [9].

Definition e_j is the general description model of element j which contains seven parameters

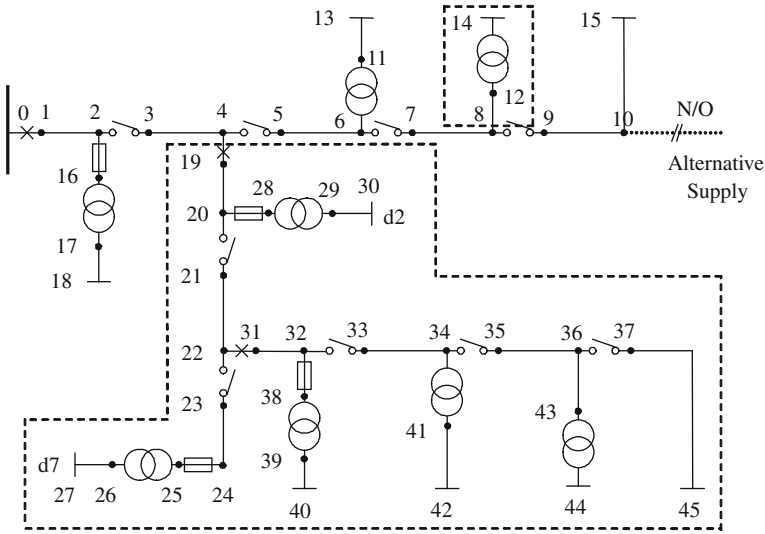
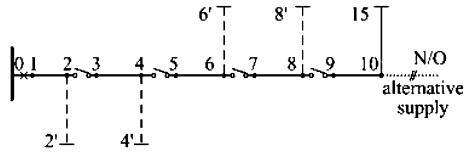


Fig. 134.1 Typical distribution system

Fig. 134.2 Major-network



$$e_j = \{i_j, j, a_j, \lambda_j^e, r_j^e, p_j, t_j\} \tag{134.1}$$

where i_j and j are the sending node and the ending node [10], respectively, λ_j^e and r_j^e are the failure rate and repair time, respectively, a_j and p_j represent the category and the availability, respectively, and t_j is the operate time of the element if the element is a disconnect or a tie switch.

In order to consider the maintenance of the equipment, the failure rate can be replaced by the result that the failure rate itself plus the maintenance outage rate. In a similar manner, the repair time can be replaced by the result that the repair time plus the maintenance outage duration.

The network [11] organized by the series elements and the single shunt elements which links the main supply and the alternative supply is named major-network. As shown in Fig. 134.2, the solid network is a major-network. Any conjoint network out of the major-network is defined as sub-network. The joint connected the major-network with the sub-network is named branch node.

Only one major-network is in existence in a distribution system, however, many sub-networks may be in existence. The direction of the power flow through the sub-network is invariable whether the alternative supply is in operation or not.

The definition of element-node network, the general description model of element, and the models of major-network and sub-network collectively constitute the basic model for evaluating the reliability proposed in this paper.

First, for sub-network, the element delivers the impact caused by the downriver failure to the upriver node in the upstream direction, the upstream reliability indices of the node point are obtained via systematizing those impacts; second, every sub-network is replaced by an equivalent element connected with the branch node whose reliability indices are equal to the upstream reliability indices, then, the systemized reliability indices of the major-network and the downstream reliability indices of the branch nodes can be calculated using FMEA method; third, as to the element which connects with the branch node in the sub-network, the downstream reliability indices of its sending node are obtained by systemizing the reliability indices of the element itself and the downstream reliability indices of its sending node, in a same manner, the reliability indices of every element can be gained one by one in a downstream processing [12, 13]. At last, systemizing the upstream reliability indices and the downstream reliability indices, the reliability indices of the sub-network node are achieved.

The principle proposed above considers the impact of every element fault on the reliability indices and its effect is equivalent with FMEA method. However, it greatly reduces the computation and the time.

134.3 New Algorithm of Distribution System Reliability Evaluation

For any distribution system [13], one or several sub-networks are obtained when removing the major-networks. Assuming the upstream fault indices of the sending node k of each sub-network is zero. That is: $\lambda_{k0}^u = 0$, $r_{k0}^u = 0$.

- (i) Select an ending element e_k arbitrarily, if the element is a switch or a fuse, remove it from the sub-network and then select another element arbitrarily till the element is neither a switch nor a fuse. Assuming the upstream fault indices of the sending node i_k is equal to the fault indices of the element itself. That is: $\lambda_{i_k k}^u = \lambda_k^e$, $r_{i_k k}^u = r_k^e$. And the element e_k is regarded as the element which has just been processed.
- (ii) Let j represent the sending node of the element which has just been processed. Then element e_j is named the upstream close neighbor element which has just been processed, and let i_j represent the sending node of element e_j .
- (iii) Element e_j delivers the upstream fault indices $\lambda_{j k}^u$ and $r_{j k}^u$ of its sending node j to its sending node whose upstream fault indices are $\lambda_{i_j k}^u$ and $r_{i_j k}^u$. And the e_j is regarded as the element which has just been processed, go to step (ii).

Table 134.1 Upstream reliable indices: case 1

Node point	Failure rate (occ/year)	Outage duration (h)	Node point	Failure rate (occ/year)	Outage duration (h)
36	3.4619	2.8806	46	1.8343	5.0000
37	3.2799	2.7630	47	1.6263	5.0000
38	3.1174	2.6464	48	1.4443	5.0000
39	3.0134	2.5651	50	1.2168	5.0000
40	2.9549	2.5169	51	1.1128	5.0000
41	2.8509	2.4264	53	0.8385	5.0000
43	2.6884	2.2708	64	0.6565	5.0000
45	2.5844	2.1610	65	0.4940	5.0000
46	0.8645	5.0000	67	0.3900	5.0000
57	0.6565	5.0000	68	0.1820	5.0000
58	0.5525	5.0000	52	0.5525	5.0000
60	0.3705	5.0000	53	0.3705	5.0000
61	0.2080	5.0000	–	–	–

- (iv) Carry through the above process (ii) and (iii) repeatedly until the element which has just been processed has no upstream close neighbor element. At that time, we complete the process of upstream delivering of the element e_k which then can be removed from the sub-network.
- (v) Carry through the above process (i–iv) repeatedly until all the elements of the sub-network are removed.

The processes (i–v) above accomplishes the upstream delivering and traversal of the element’s fault indices of all the sub-networks. Using the systemization Eqs. 134.2–134.3, the upstream reliability indices which consider the fault of the downriver elements can be obtained. The fault rate λ_i^u and repair time r_i^u are:

$$\lambda_i^u = \sum_{k \in \Omega_i} \lambda_{ik}^u \tag{134.2}$$

$$r_i^u = (\sum_{k \in \Omega_i} \lambda_{ik}^u r_{ik}^u) / \lambda_i^u \tag{134.3}$$

where Ω_i is the set of all serial number of the downriver nodes of node i .

134.4 Experiment

Case 1 To illustrate the approach presented in a general sense, breakers 6, 7 and 8 are assumed to be 80% reliable with no alternative supply to main feeder 4. The detailed analysis is as follows. There are three subfeeders in this system. The first

Table 134.2 Downstream reliable indices: case 1

Node point	Failure rate (occ/year)	Outage duration(h)	Node point	Failure rate\ (occ/year)	Outage duration(h)
36	0.0000	5.0000	45	1.6254	5.0000
37	0.1720	5.0000	46	1.8334	5.0000
38	0.3236	5.0000	47	2.0272	5.0000
39	0.4456	5.0000	51	2.2351	5.0000
40	0.5034	5.0000	51	2.3491	5.0000
41	0.6133	5.0000	53	2.7911	5.0000
43	0.7723	5.0000	64	2.9731	5.0000
45	0.8756	5.0000	65	3.1356	5.0000
46	2.7721	2.3345	67	3.2396	5.0000
57	2.9767	2.4323	68	3.444	5.0000
58	3.0832	2.5323	52	3.0198	5.0000
60	3.2621	2.7511	53	3.2049	5.0000
61	3.4243	2.8523	54	3.4056	5.0000

Table 134.3 Reliable indices of the sub-network: case 1

Node point	Failure rate (occ/year)	Outage duration (h)	Node point	Failure rate (occ/year)	Outage duration (h)
36	3.4619	2.8806	47	3.4613	5.0000
37	3.4619	2.8806	48	3.4632	5.0000
38	3.4619	2.8806	50	3.4634	5.0000
39	3.4619	2.8806	64	3.6245	5.0000
40	3.4619	2.8806	65	3.6256	5.0000
41	3.4619	2.8806	67	3.6267	5.0000
43	3.4619	2.8806	68	3.6287	5.0000
57	3.4619	2.8806	69	3.6245	5.0000
58	3.4619	2.8806	53	3.5736	5.0000
60	3.4619	2.8806	54	3.5725	5.0000
61	3.4619	2.8806	55	3.5721	5.0000
62	3.4619	2.8806	–	–	–

step is to calculate the upstream indices of each node. Table 134.1 shows a representative sample of the upstream reliable indices of case 1.

After the upstream reliable indices of the sub-network have been found, the next step is to calculate the downstream reliable indices of the sub-network. Table 134.2 shows a representative sample of the downstream reliable indices of case 1.

By synthesizing the downstream and upstream reliable indices, we can obtain the reliable indices of any node in the sub-network. Table 134.3 shows a representative sample of reliable indices of the sub-network of case 1.

After the reliable indices both the sub-network and the major-network have been found, we can obtain the load-point reliable indices. Table 134.4 shows a representative sample of load-point reliable indices of case 1.

Table 134.4 Load-point reliable indices: case 1

Node point	Failure rate (occ/year)	Outage duration (h)	Node point	Failure rate (occ/year)	Outage duration (h)
LP1	0.3336	2.4717	LP21	3.4769	2.9113
LP2	0.3433	2.4533	LP22	3.4769	2.9112
LP3	0.3400	2.5449	LP23	3.4769	2.9115
LP4	0.3331	2.4718	LP24	3.4769	2.9116
LP5	0.3423	2.4295	LP25	3.4769	5.0211
LP6	0.3356	2.3630	LP26	3.4769	5.0211
LP7	0.3693	2.3157	LP27	3.4769	5.0216
LP8	0.3725	2.4443	LP28	3.5874	5.0209
LP9	0.3725	2.3396	LP29	3.5874	5.0209
LP10	0.3595	2.2434	LP30	3.5874	5.0209
LP11	0.3693	2.4568	LP31	3.6498	3.0103
LP12	0.3595	2.3519	LP32	3.6498	3.0103
LP13	0.3693	2.3160	LP33	3.6498	3.0103
LP14	0.2275	2.5429	LP34	3.6498	3.0103
LP15	0.2372	3.5211	LP35	3.6498	3.0103
LP16	0.2405	4.1214	LP36	3.6446	5.0206
LP17	0.2275	5.2145	LP37	3.6446	5.0206
LP18	3.4769	2.9113	LP38	3.6446	5.0206
LP19	3.4769	2.9113	LP39	3.6446	5.0206
LP20	3.4769	2.9113	LP40	3.6446	5.0206

134.5 Conclude

By utilizing the structure features of the distribution system adequately, especially there is no supply in the sub-network, the approach in this paper can substantially accelerate the reliability assessment speed. The order of upstream traversal can simply use the opposite order of the downstream traversal which commendably overcomes the disadvantage of the FMEA method which needs to build two traversal tables, and the researches of the table are expressly complicated. The proposed method can uniquely determine the sub-network which conquers the deficiency of the equivalent approach which needs to consider the multistage branches, the hierarchical search and the complex calculation. The method avoids the overlap influence of the fault of the downriver elements traversing through the multi-fuse which makes the fault's diffuse distortion when dealing with the multistage branches using the equivalent approach. This method is not only a fast and reliable evaluation method but also has the same effect as the FMEA method. It has taken into account the impact on reliability evaluation caused by the fault of the switch element which overcomes the limitation of the existent methods. Through carrying on the combination between the main power source and each alternative supply, the proposed method can also be applied for multistage arbitrary branches and more alternative supplies distribution system.

References

1. Billinton R, Billinton JE (1989) Distribution system reliability indices. *IEEE Trans Power Deliv* 4(1):561–568
2. Billinton R, Wang P (1999) Teaching distribution system reliability evaluation using Monte Carlo simulation. *IEEE Trans Power Syst* 14(2):397–403
3. Ming D, Jing Z, Shenghu L (2004) A sequential Monte-Carlo simulation based reliability evaluation model for distribution network. *Power Syst Technol* 28(3):38–42
4. Allan RN, Billinton R, Sjarief I (1991) A reliability test system for educational purpose-basic distribution system data and results. *IEEE Trans Power Syst* 6(2):813–830
5. Billinton R, Jonnavithula S (1996) A test system for teaching overall power system reliability assessment. *IEEE Trans Power Syst* 11(4):1670–1676
6. Zhaohong B, Xiuli W, Xifan W (2000) Reliability evaluation of complicated distribution systems. *J Xi'an Jiao Tong Univ* 34(8):9–13
7. Kaigui X, jiaqi Z, Billinton R (2003) Reliability evaluation algorithm for complex medium voltage electric distribution networks based on the shortest path. *IEE Proc Gener Transm Distrib* 150(6):686–690
8. Weixing L, Zhimin L, Yingchun L (2003) Evaluation of complex radial distribution system reliability. *Proc CSEE* 23(3):69–73
9. Zhenxia X, Jiangxin Z (2005) An improved fault traversal algorithm for complex distribution system reliability evaluation. *Power Syst Technol* 29(14):64–67
10. Peng Z, Shouxiang W (2004) A Novel interval method for reliability evaluation of large scale distribution system. *Proc CSEE* 24(3):77–84
11. Billinton R, Wang P (1998) Reliability network equivalent approach to distribution system reliability evaluation. *IEE Proc Gener* 145(2):149–153
12. Yang G, Xu X, Gang Y, Zhang J (2010) Semi-supervised classification by local coordination, lecture notes in computer science. *Neural Inf Process Model Appl* 6444:517–524
13. Yang G, Xu X, Yang G, Zhang J (2010) Research of local approximation in semi-supervised manifold learning. *J Inf Comput Sci* 7(13):2681–2688

Chapter 135

State-of-the-Art Line Drawing Techniques

Lei Li, Yongqiang Zhou, Chunxiao Liu, Yingying Xu and Jie Fu

Abstract As a simple yet effective means of visual communication, line drawings express meaningful visual information with sparse lines while ignoring relatively unimportant details. We cover some fundamental algorithms in image and 3D model based line drawings, and separately talk about their advantages and disadvantages. We hope these can help motivate and guide readers who are interested in line drawing techniques.

Keywords Line drawings · Image space · 3D model · Edge detection · Feature lines

135.1 Introduction

Line drawing is a drawing composed of just sparse lines, representing the main information of an image or object. It often plays crucial roles in preserving or emphasizing salient features, making higher level of abstraction possible,

L. Li · Y. Zhou · C. Liu (✉) · Y. Xu · J. Fu
College of Computer Science and Information Engineering, Zhejiang Gongshang
University, Hangzhou 310018, China
e-mail: cxliu@mail.zjgsu.edu.cn

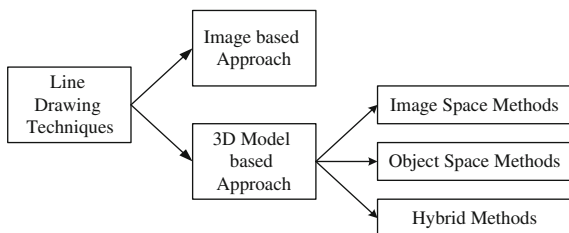
L. Li
e-mail: anguslilei@163.com

Y. Zhou
e-mail: zyqhi2010@163.com

Y. Xu
e-mail: 195884240@qq.com

J. Fu
e-mail: 375424632@qq.com

Fig. 135.1 Overview of the state-of-the-art line drawing techniques



providing more effective communication and so on. This enables us to quickly identify and appreciate the subject without distraction from relatively unimportant contents. Also, this kind of representation may give us significant benefits. For example, we could spend more less time and storage space in the subsequent processing of the data.

In recent years, there emerge more and more algorithms dealing with the image based line drawing and the 3D model based line drawing problems. Traditionally, various problems in image based line drawing have been addressed by a low-level image analysis technique called edge detection and we will see some methods later. However, edge detectors typically have some limitations, such as that the resulting edge map is less meaningful to the viewers and so on. Furthermore, we will also see some methods that do not have these limitations.

In 3D model based line drawing, a variety of methods have been reported. We have known the shape of a 3D object is normally explicitly represented by low-level geometric elements (such as points or vertices on the surface) whose coordinates usually we have known, and thus the problem of 3D model based line drawing is reduced to identification of important contours formed by these elements. We will describe some methods for finding lines efficiently, including image- and object space algorithms.

135.2 Technical Overview of Line Drawing Methods

Although there emerged lots of line drawing methods, most of them can be classified as image or 3D model based approaches according to their starting point is an image or 3D model. In terms of the line extraction spaces, 3D model based approach can be further sorted into three categories, i.e., image space methods, object space methods and hybrid methods, as shown in Fig. 135.1. We describe them in detail at the following sections.

135.3 Image Based Line Drawings

For the image based line drawing, traditional edge detectors are often employed and adapted, such as Canny's, DoG filtering, and so on. There are also many other methods dealing with the Image based line drawing problem.

Canny's edge detector is an often used choice for many applications because it typically produces cleaner results [1]. DeCarlo and Santella use the robust variant of the Canny detector to produce line drawing [2]. In their work, images are transformed into a line drawing with bold edges and large regions of constant color. Kang et al. show that it is also possible to produce artistic lines ranging from region-based to outline-based by using a modified Canny edge detector [3]. Orzan et al. develop a multiscale image abstraction system based on the Canny edge detector and the gradient reconstruction method [4]. Their method allows to remove details without necessarily simplifying shapes. Fischer et al. employ the Canny edge detector and morphological operator to obtain the edges in their stylized augmented reality work [5]. And the edges appear to be thicker and more continuous. But, because Canny's method strongly depends on the gradient magnitudes, it is often impossible to exclude a set of strong but unimportant edges without losing other important features

Also one may choose other edge detection method for line drawing. Gooch et al. [6] present a facial illustration system based on the difference-of-Gaussians (DoG) filter, similar to Marr. Hildreth edge detector [7]. Winnemöller et al. extend this technique to general color images and video by employing the DoG filter for line drawing [8]. Compared with Canny's method, their DoG edge model has proven to be more effective: It captures interesting structures better automatically produces stylistic lines, and creates an impressive look reminiscent of pen-and-ink line drawing done by real artists. Due to the nature of isotropic filter kernel, the aggregate of edge pixels may not clearly reveal the sense of 'directedness' (and thus may look less like lines). Also, the thresholded edge map may exhibit some sets of isolated, scattered edge components that clutter the output, especially in an area with image noise or weak contrast.

Due to the weakness of isotropic filter kernel, Kang et al. present a technique that automatically generates a high-quality line drawing from a photograph by introducing a novel flow-driven anisotropic filtering framework [9]. The lines are extracted by applying a Difference-of-Gaussians (DoG) filter followed by thresholding. First, they constructed the edge flow field which they called edge tangent flow (ETF) from the input image. And then, they apply a flow-guided anisotropic DoG filter using the kernel whose shape is defined by the local flow recorded in the ETF and they named it flow-based Difference-of-Gaussians (FDoG) filter. Their contribution lies in reducing fragmentation and extracting highly coherent lines by adaptively deciding the shape and orientation of the filter kernel from the characteristics of local image. So their modified line extraction filter dramatically enhances the spatial coherence of lines and also suppresses noise. But the FDoG filter may have some limitations since it builds on the DoG filter. For example, a

high-contrast background will be filled with a dense set of lines although this area may be perceptually unimportant.

Moreover, Rosin et al. apply some well known techniques from the computer vision literature to improve the output of Kang et al.'s method [10]. They apply hysteresis and got further improvement, for example, many small disconnected edges are removed, and in some places lines are extended by retaining edges that were previously removed by the constant threshold.

Son et al. present a approach that is based on local line fitting, mimicking the human line drawing process [11]. The main idea is to extract lines that locally have the biggest 'likelihood' of being genuine lines that will be of interest to the viewers. Their overall framework consists of two parts and can be decomposed into six steps: feature point sampling, likelihood function computation, feature scale and blurriness computation, linking, curve fitting, and texture mapping. Compared to the FDoG filtering approach, this algorithm exclusively provides control over feature selection, level-of-focus, and line style, each of which could lead to more effective visual communication. In this framework, the image gradient still plays a vital role in determining the level of pixel salience. As a result, it may not be easy to entirely discard some strongly textured but unimportant background.

135.4 3D Model Based Line Drawings

Here, we focus on the direction of research in 3D mesh: the algorithms for extracting lines. A variety of different algorithms used to compute outlines for 3D models already exist. Reference to the researches of Isenberg et al. [12] and Rusinkiewicz et al. [13], we divide these algorithms into two major classes: image space algorithms, and object space algorithms. Finally, there are hybrid algorithms, which perform a bit of processing in object space, but the lines ultimately show up only in the frame buffer. Most of our sight will be on outlines such as silhouettes, creases, and suggestive contours. Now we analyze them separately.

135.4.1 Image Space Methods

Image space methods, render something such as a depth map or cosine-shaded model, then extract lines only operate on image buffers and provide a silhouette represented as features in a pixel matrix.

Saito and Takahashi suggest using the z-buffer, and applying an edge detector such as the Sobel operator which has the advantage of only finding object-relevant edges such as silhouette lines and contours [14]. Hertzmann extends this method by using a normal buffer instead [15].

Some approaches extract curves from the silhouette pixels such as presented by Loviscach, who fits Bezier curves to the pixels [16]. For each pixel, their solution stores the depth (more precisely, the ray length) and the direction of the normal vector of the underlying geometry. However, the process of fitting curves to the pixel silhouettes might introduce new artifacts and inaccuracies.

Raskar et al. present a novel NPR camera which detects depth edges using multi-flash images [17]. They use the location of the shadows abutting depth discontinuities as a robust cue to create a depth edge map in both static and dynamic scenes which can highlight the detected features, suppress unnecessary details or combine features from multiple images. The resulting images more clearly convey the 3D structure of the imaged scenes, and the method is both surprisingly simple and computationally efficient.

Lee et al. propose to automatically extract lines at appropriate scales from abstract shading [18]. Then they render lines along thin dark areas or tone boundaries in the shaded image. The resulting line drawings generated by this algorithm effectively convey both shape and material cues.

This kind of algorithm is relatively simple and fast to implement, and provides some notions of view dependent level of detail. However, it is difficult to control the appearance, and the silhouette is not available in the form of an analytic line description result in while silhouettes can typically not be stylized or used for further processing.

135.4.2 Object Space Methods

Object space methods, perform all calculations in object space, and provide the resulting lines represented by an analytic description which meet the requirement of applying further stylization to the lines.

Buchanan and Sousa suggest using a data structure called an edge buffer to support the process of silhouette edge detection [19]. However, this algorithm computes the silhouettes with the brute-force, and checks each edge whether has one adjacent front facing and one adjacent back-facing.

McGuire describes GPU hardware methods for extracting silhouettes from 3D meshes, and shows how to use hardware to create thick screen-space contours with end-caps that join adjacent thick line segments [20].

There is a large body of work on acceleration techniques that try to reduce running time. Sander et al. propose a popular method for silhouette edge detection which constructs a hierarchical search tree to store the mesh's edges [21]. Another interesting acceleration technique involves the Gauss map which is used by Gooch et al. [22], Benichou and Elber [23] and Hertzmann [24]. Here, every mesh edge corresponds to an arc on the Gaussian sphere, which connects the normal's projections of its two adjacent polygons. A very different sort of acceleration technique, most suited to interactive systems, is suggested by Markosian et al. who

suggest a stochastic algorithm to gain silhouettes [25]. They observe that only a few edges in a polygonal model are actually silhouette edges.

In the past decade, there has been a large amount of different definitions on object space lines, including suggestive contours by DeCarlo et al. [26], ridge-valley lines by Ohtake et al. [27], apparent ridges by Judd et al. [28], suggestive highlights by DeCarlo and Rusinkiewicz [29], photic extremum lines (PELs) by Xie et al. [30], demarcating curves by Kolomenkin et al. [31], etc. Cole et al. have shown that different families of lines should be combined to effectively depict a large variety of shapes [32, 33]. According to this, Grabli and Turquin define a new data structure, the view map, which combines geometry, topology, and the rest of the properties found on feature lines, makes it easy to query scene property values and manipulate feature lines in the style module [34]. Flexible, their implementation is not restricted to a given set of lines and can easily be extended to include new types of lines.

Inspired by the Laplacian-of-Gaussian (LoG) edge detector in image processing, Zhang et al. present a novel object-space line drawing algorithm which can depict shapes with view dependent feature lines in real-time [35]. They define Laplacian lines as a set of points where the Laplacian of the diffuse illumination vanishes and the gradient magnitude is greater than certain user-specified threshold. Laplacian lines inherit the advantages of the LoG operator by employing a smoothing preprocessing to reduce the high frequency noise prior to the differentiation step. They provide considerably abstractive visual information and suppress distracting details. In this way, most expensive computations can be preprocessed, and their Laplacian line extraction algorithm is robust to noise and irregular mesh tessellation.

These algorithms tend to be a little more complex which need longer to compute, and it is more difficult to adapt them to take advantage of graphics hardware. On the other hand, they provide good control over stylization and typically yield highly precise visibility information.

135.4.3 Hybrid Methods

Hybrid methods perform manipulations in object space but yield the lines in an image buffer, which are followed by rendering the modified polygons in a more or less ordinary way using the z-buffer. This usually requires two or more rendering passes. The result is similar to image space algorithms in that the silhouette is represented in a pixel matrix.

Hybrid algorithms are much less general than the above kinds of algorithms, and are specialized for e.g., contours. Rustagi presents a simple algorithm using the stencil buffer that delivers the contour only, without the complete silhouette of an object [36]. Then Rossignac and Emmerik present a method based on z-buffer rendering that not only contours, but also silhouettes [37].

Raskar describes a hardware-accelerated method of rendering silhouette edges [38]. They use traditional z-buffering along with front- or back-face culling to automatically determine front- and back-facing polygons, respectively. Further, Raskar proposes a one-pass hardware implementation that basically adds borders around each triangle [39]. And he uses environment mapping in addition to regular shading that achieves a stylistic effect.

Due to the lack of connectivity information, most existing polygon-based silhouette generation algorithms cannot be applied to point-based models, Xu propose a new method to generate silhouettes not only bypasses this connectivity requirement, but also accommodates point-based models with sparse non-uniform sampling and inaccurate/no normal information [40]. They render the points as enlarged opaque disks in the first pass to obtain a visibility mask, while as regular size splats/disks in the second pass. In this way, edges are automatically depicted at depth discontinuities, usually at the silhouette boundaries.

These algorithms are the typically higher degree of control over the outcome. But the silhouette lines have pixel resolution and don't facilitate further stylization, and the algorithms might have some numerical problems due to limited z-buffer resolution.

135.5 Future Work

The fundamental algorithms presented here can guide developers who need to choose one or more from these algorithms for their applications. Possible future research directions could include, for example, (1) for image based line drawings, the acceleration schemes for the proposed filters, and (2) for 3D model based line drawings, Laplacian lines are an effective view dependent feature and promising to convey large scale models for interactive graphics applications.

Acknowledgments This project was supported by the National Natural Science Foundation of China under Grant No. 61003188, Technology Plan Program of Zhejiang Province under Grant No. 2009C11034, and Zhejiang Provincial Natural Science Foundation of China under Grant No. Y1111159, No. Z1101243 and No. Z1111051.

References

1. Canny J (1986) A computational approach to edge detection. *IEEE Trans Pattern Anal Mach Intell* 8(6):679–698
2. Decarlo D, Santella A (2002) Stylization and abstraction of photographs. *ACM Trans Graph* 21(3):769–776
3. Kang H, Chui C, Chakraborty U (2006) A unified scheme for adaptive stroke-based rendering. *Vis Comput* 22(9):814–824
4. Orzan A, Bousseau A, Barla P, Thollot J (2007) Structure-preserving manipulation of photographs. In: *Proceedings of NPAR*, pp 103–110

5. Fischer J, Bartz D, Strasser W (2005) Stylized augmented reality for improved immersion. In: Proceedings of the IEEE VR, pp 195–202
6. Gooch B, Reinhard E, Gooch A (2004) Human facial illustrations: creation and psychophysical evaluation. *ACM Trans Graph* 23(1):27–44
7. Marr D, Hildreth EC (1980) Theory of edge detection. In: Proceedings of the royal society of London, pp 187–217
8. Winnemöller H, Olsen SC, Gooch B (2006) Real-time video abstraction. *ACM Trans Graph* 25(3):1221–1226
9. Kang H, Lee S, Chui CK (2007) Coherent line drawing. In: Proceedings of NPAR, pp 43–50
10. Rosin PL, Lai YK (2010) Towards artistic minimal rendering. In: Proceedings of the NPAR, pp 119–127
11. Son MJ, Kang H, Lee YJ, Lee SY (2007) Abstract line drawings from 2D images. In: Proceedings of the 15th pacific conference on computer graphics and applications, pp 333–342
12. Isenberg T, Freudenberg B, Halper N, Schlechtweg S, Strothotte T (2003) A developer's guide to silhouette algorithms for polygonal models. *IEEE Comput Graph Appl* 23(4):28–37
13. Rusinkiewicz S, Cole F, Decarlo D, Finkelstein A (2008) Line drawings from 3D models. In: Proceedings of ACM SIGGRAPH, class notes
14. Saito T, Takahashi T (1990) Comprehensible rendering of 3-D shapes. In: Proceedings of ACM SIGGRAPH, ACM SIGGRAPH, *Comput Graph* 24(4):197–206
15. Hertzmann A (1999) Introduction to 3D non-photorealistic rendering: silhouettes and outlines. In: Proceedings of the ACM SIGGRAPH, course notes
16. Loviscach J (2002) Rendering artistic line drawings using off-the-shelf 3-D software. In: Proceedings of the eurographics: short paper, pp 125–130
17. Raskar R, Tan KH, Feris R, Yu J, Turk M (2004) Non-photorealistic camera: depth edge detection and stylized rendering using multi-flash imaging. *ACM Trans Graph* 23(3):679–688
18. Lee Y, Markosian L, Lee S, Hughes JF (2007) Line drawings via abstracted shading. *ACM Trans Graph* 26(3):18:1–18:5
19. Buchanan JW, Sousa MC (2000) The edge buffer: a data structure for easy silhouette rendering. In: Proceedings of the NPAR, pp 39–42
20. McGuire M, Hughes JF (2004) Hardware-determined feature edges. In: Proceedings of the NPAR, pp 35–47
21. Sander PV, Gu XF, Gortler SJ, Hoppe H, Snyder J (2000) Silhouette clipping. In: Proceedings of the ACM SIGGRAPH, pp 327–334
22. Gooch B, Sloan PJ, Gooch A, Shirley P, Riesenfeld R (1999) Interactive technical illustration. In: Proceedings of the I3D, pp 31–38
23. Benichou F, Eiber G (1999) Output sensitive extraction of silhouettes from polygonal geometry. In: Proceedings of the PG, pp 60–69
24. Hertzmann A, Zorin D (2000) Illustrating smooth surfaces. In: Proceedings of ACM SIGGRAPH, pp 517–526
25. Markosian L, Kowalski MA, Trychin SJ, Bourdev LD, Goldstein D, Hughes JF (1997) Real-time nonphotorealistic rendering. In: Proceedings of the ACM SIGGRASP, pp 415–420
26. Decarlo D, Finkelstein A, Rusinkiewicz S, Santella A (2003) Suggestive contours for conveying shape. *ACM Trans Graph* 22(3):848–855
27. Ohtake Y, Belyaev A, Seidel HP (2004) Ridge-valley lines on meshes via implicit surface fitting. *ACM Trans Graph* 23(3):609–612
28. Judd T, Durand F, Adelson E (2007) Apparent ridges for line drawing. *ACM Trans Graph* 26(3):19:1–19:7
29. Decarlo D, Rusinkiewicz S (2007) Highlight lines for conveying shape. In: Proceedings of the NPAR, pp 63–70
30. Xie XX, He Y, Tian F, Seah HS, Gu XF, Qin H (2007) An effective illustrative visualization framework based on photic extremum lines (PELs). *IEEE Trans Vis Comput Graph* 13(6):1328–1335

31. Kolomenkin M, Shimshoni I, Tal A (2008) Demarcating curves for shape illustration. *ACM Trans Graph* 27(5):157:1–157:9
32. Cole F, Golovinskiy A, Limpaecher A, Barros HS, Finkelstein A, Funkhouser T, Rusinkiewicz S (2008) Where do people draw lines? *ACM Trans Graph* 27(3):88:1–88:11
33. Cole F, Sanik K, Decarlo D, Finkelstein A, Funkhouser T, Rusinkiewicz S, Singh M (2009) How well do line drawings depict shape? *ACM Trans Graph* 28(3):28:1–28:9
34. Grabli S, Turquin E, Durand F, Sillion FX (2010) Programmable rendering of line drawing from 3D scenes. *ACM Trans Graph* 29(2):18:1–18:20
35. Zhang L, He Y, Xie XX, Chen W (2009) Laplacian lines for real-time shape illustration. In: *Proceedings of the I3D*, pp 129–136
36. Rustagi P (1989) Silhouette line display from shaded models. *IRIS Univ* 9:42–44
37. Rossignac JR, Emmerik M (1992) Hidden contours on a frame-buffer. In: *Proceedings of the 7th eurographics workshop on computer graphics hardware*, pp 188–204
38. Raskar R, Cohen M (1999) Image precision silhouette edges. In: *Proceedings of the I3D*, pp 135–140
39. Raskar R (2001) Hardware support for non-photorealistic rendering. In: *Proceedings of the ACM SIGGRAPH/EUROGRAPHICS workshop on graphics hardware*, pp 41–47
40. Xu H, Nguyen M, Yuan X, Chen B (2004) Interactive silhouette rendering for point-based models. In: *Proceedings of the eurographics symposium on point-based graphics*, pp 13–18

Chapter 136

The Cooperation and Competition Mechanism of Supply Chain Based on Evolutionary Game Theory

Peng Yang and Fu Pei-Hua

Abstract The paper establishes the evolutionary game model of supply chain partnership based on view of evolutionary game theory, and analyzes the cooperative and competition mechanism in supply chain, and its dynamic evolutionary procedure. The results showed that the evolution of the system the direction of the payoff matrix with two game-related, but also by the initial state of the system. And propose the key factors in co-competition supply chain, put forward some related research proposals.

Keywords Supply Chain · Co-competition · Evolutionary Game

136.1 Introduction

Supply chain consists of many organizations acting together, The competition and cooperation issues are jointly addressed in a supply chain. Arunachalam et al. (2003) [1] observes the importance of corporations forming alliances in order to best exploit market conditions and improve competitiveness, The collaboration between firms of different groups can gain resource complementarily and synergy on technique, products and finance, which can increase their own and their partners' profitability together.

In understanding the bounded rationality of channel members of a supply chain, we think that the evolutionary game theory can provide another way to study the channel coordination mechanism. Evolutionary game theories originate as an

P. Yang (✉) · F. Pei-Hua
School of Computer and Information Engineering, Zhejiang Gongshang University,
Hangzhou 310035, China

application of the mathematical theory of games to biological problems based on the realization that frequency dependent fitness introduces a strategic aspect to evolution. Recently, there has been an increased interest in using this theory by economists, sociologists, and anthropologists that are social scientists in general as well as philosophers (e.g. [2–4]). The evolutionary game theory consists of two main approaches. The first approach derives from the work of Maynard-Smith et al. (1973) [6] and employs the concept of an evolutionarily stable strategy as the principal tool of analysis. The second approach constructs an explicit model of the evolution process in which the frequency of strategies changes with the population sizes, and studies the properties of the evolutionary dynamics based on that model. Although there exists one study by Wang et al. (2004)[5] that develops an evolutionary game model for supply chain partnerships from an evolutionary game’s viewpoint, only a simple theoretical framework is presented.

The remainder of the paper is organized as follows: Section 136.2 provides a evolutionary game model of supply chain. In Sect. 136.3, we describe the details of Analysis of the model, study the factors that affect the supply chain strategy. Finally, we conclude the paper with a summary and discussion in Sect. 136.4.

136.2 The evolutionary Game Model of Supply Chain

136.2.1 Background and Assumptions

For asymmetric game between two groups, each game is carried out between members of one group and members of another group, the analysis framework is that a member of two groups can be repeatedly chosen at random to be compared in a game way. Their study and imitation can be limited to their own group. Strategy adjustment mechanism is still similar to replication dynamics. Where two members are in the state of symmetric.

If a group of suppliers S carried out strategy game with manufacture M, in supply chain system, both part’s strategy sets are(collaboration, competition),Moreover, no organization devise or arrange such supply chain system. The supply chain system evolves spontaneously from the principle of survival of the fittest, in which supplier and manufacturers select and adjust their own strategies by considering their relative adaptability in the group according to other members’ strategies selection. Assumptions of this model as follows:

(1) U_s : Suppliers’ normal benefits when adopt competitive strategy; (2) U_m : Manufacturers’ normal benefits when adopt competitive strategy; (3) C_s : Suppliers’ input initial cost when selecting cooperation; (4) C_m : Manufacturers’ input initial cost when selecting cooperation; (5) ΔV_s : Suppliers’ excess income under cooperative strategy. (6) ΔV_m : Manufacturers’ excess income under cooperative strategy; (7) ΔV : the excess income sum, and $\Delta V, \Delta V_s, \Delta V_m \geq 0$; (8) δ_s : the discount factor of suppliers; (9) δ_m : the discount factor of manufacturers;

Table 136.1 The payoff matrix of game in supply chain

		Suppliers	
		Cooperation	Competition
Manufacturers	Cooperation	$U_m + \Delta V_m, U_s + \Delta V_s$	$U_m - C_m, U_s$
	Competition	$U_m, U_s - C_s$	U_m, U_s

136.2.2 Payoff Matrix and Replication Dynamics Equations

As such, the payoff matrix of cooperation and competition game in supply chain is as Table 136.1.

Among it, the payoff matrix of manufacturers is $A, A = \begin{bmatrix} U_m + \Delta V_m & U_m - C_m \\ U_m & U_s \end{bmatrix}$ And the payoff matrix of suppliers is $B, B = \begin{bmatrix} U_s + \Delta V_s & U_s - C_s \\ U_s & U_s \end{bmatrix}$

If the proportion of manufacturer’s collaboration strategy is x ($0 \leq x \leq 1$), then the proportion of completion strategy is $1 - x$; and if the proportion of suppliers’ collaboration strategy is y ($0 \leq y \leq 1$), the proportion of completion strategy is $1 - y$, the replication dynamics equations is:

$$\begin{aligned} \hat{x} &= x[e.A(y, 1 - y) - (x, 1 - x).A(y, 1 - y)] = x(1 - x)[(\Delta V_m + C_m)y - C_m] \\ \hat{y} &= y[e.B(x, 1 - x) - (y, 1 - y).B(x, 1 - x)] = y(1 - y)[(\Delta V_s + C_s).x - C_s] \end{aligned}$$

136.2.3 System Balance Points

Judging from the replication dynamics equations, the system has 5 balance points in the area: $M = \{(x, y) \mid 0 \leq x, y \leq 1\}$, which are respectively as:

$O(0, 0), A(0, 1), B(1, 0), C(1, 1)$ and $D(\frac{C_s}{\Delta V_s + C_s}, \frac{C_m}{\Delta V_m + C_m})$, Cressman [7] put forward that, for the two matrix evolutionary game of two populations two strategies, the balance point is ESS determined by it is local asymptotic stability, and by the method proposed by Friedman (1991) [8], the stability of balance point can be get by analyzing the local stability of its’ Jacobin matrix.

According to the analysis of Jacobin matrix local stability in these balance points, the system’s Jacobin matrix is:

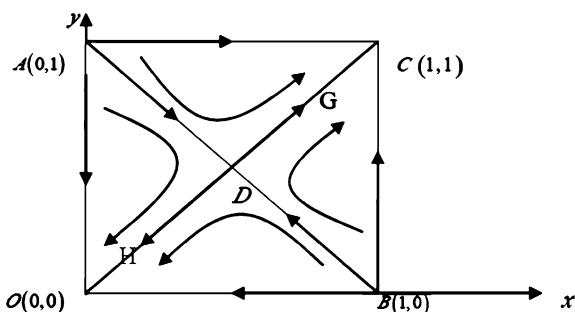
$$J = \begin{bmatrix} [(\Delta V_m + C_m)y - C_m](1 - 2x) & (\Delta V_m + C_m)x(1 - x) \\ (\Delta V_s + C_s)y(1 - y) & [(\Delta V_s + C_s)x - C_s](1 - 2y) \end{bmatrix}$$

The Determinant of Jacobin matrix is:

Table 136.2 Local stability of balance

Balance point	det	trJ	Local stability
$x = 0, y = 0$	$C_m \cdot C_s$	$+ \quad -C_m - C_s$	$- \quad$ ESS
$x = 0, y = 1$	$\Delta V_m \cdot C_s$	$+ \quad \Delta V_m + C_s$	$+ \quad$ Unstable
$x = 1, y = 0$	$C_m \cdot \Delta V_s$	$+ \quad C_m + \Delta V_s$	$+ \quad$ Unstable
$x = 1, y = 1$	$\Delta V_m \cdot \Delta V_s$	$+ \quad -\Delta V_m - \Delta V_s$	$- \quad$ ESS
$x = \frac{C_s}{\Delta V_s + C_s},$ $y = \frac{C_m}{\Delta V_m + C_m}$	$-C_s \cdot C_m \cdot (1 - \frac{C_s}{\Delta V_s + C_s}) (1 - \frac{C_m}{\Delta V_m + C_m})$	$- \quad 0$	Saddle point

Fig. 136.1 Phase diagram



$$\det J = [(\Delta V_m + C_m)y - C_m](1 - 2x)[(\Delta V_s + C_s)x - C_s](1 - 2y) - (\Delta V_m + C_m)x(1 - x)(\Delta V_s + C_s)y(1 - y)$$

The Jacobin matrix trace is:

$$\text{tr}J = [(\Delta V_m + C_m)y - C_m](1 - 2x) + [(\Delta V_s + C_s)(1 - 2y)]$$

Then, the determinant of matrix J in 5 balances and its trace value or symbol are obtained, and its local stability are judged accordingly.

Table 136.2 shows that O(0,0) and C(1,1) are evolutionary stable strategy; A(0,1), B(1,0) are unstable points, and $D(\frac{C_s}{\Delta V_s + C_s}, \frac{C_m}{\Delta V_m + C_m})$ is saddle point.

The about conclusion can be brought out in system phase diagram as Fig. 136.1.

In Fig. 136.1, the connection between point D and any vertex is not generally a straight line, because “pass through” two rail are not known accurately, straight line here replace.

Analysis shows that on the phase diagram, when initial state is in the region of G, which is the upper right area of the connection of point A(0,1), B(1,0) and D, system will converge to C(1,1), that is, all manufactures and suppliers employ “collaboration” strategy; and when initial state is in the region H, which is the left below area of the connection of point A(0,1), D and B(1,0), system will converge to O(0,0), which shows all manufactures and suppliers adopt “competition” strategy.

136.3 Analysis of the model

The long-term equilibrium of the system evolution may be full cooperation, it may be perfectly competitive, how to reach what path along which the state of the game with payoff matrix are closely related. In certain information to guide the mechanism, the system will be the convergence point of a balanced game where the initial state of occurrence. Therefore, in the course of the game, two game payoff functions constitute some of the parameters and change the initial value of the evolution of the system will lead to different equilibrium convergence.

Use δ_s, δ_m to represent the supplier and the manufacturer's discount factor, $\Delta V_s = \frac{\delta_s(1-\delta_m)}{1-\delta_m\delta_s} \cdot \Delta V$, $\Delta V_m = \frac{1-\delta_s}{1-\delta_m\delta_s} \cdot \Delta V$

Can be learned from the above equation:

$$X_D = \frac{1}{\left(\frac{1-\delta_s}{1-\delta_m\delta_s}\right) \cdot \frac{\Delta V}{C_m} + 1}, Y_D = \frac{1}{\left(\frac{\delta_s(1-\delta_m)}{1-\delta_s\delta_m}\right) \cdot \frac{\Delta V}{C_s} + 1}$$

From the above analysis shows that the evolution of the system parameters are: co-generated excess profits ΔV , the two sides co-pay for the initial cost C_s, C_m , and the two sides of the discount factor δ_s, δ_m , Respectively discuss as follows.

136.3.1 The Excess Profits ΔV by Co-Produced

When the excess profits generated by increased co-operation, the coordinates of point D will be smaller and that point will move to the lower left of M, the area G will be expanded. While the H area will be reduced on the other hand, when cooperation Generated excess profits decrease, the point coordinates and D will become larger, the point will move to the top right of M, the area G will be reduced, but the H will expand. In practice, supply chain requires both partners to focus on whether the complementarity of resources, technology, products and financial effect of such coordination, to achieve excess profits generated by maximizing cooperation to ensure that suppliers and manufacturers the establishment of strategic partnerships and stability.

136.3.2 The Initial Cooperation Costs of Suppliers and Manufacturers

Initial cost is the major parties choose to assess the pre-paid partners and other costs. When the initial cost C_s, C_m increased, and the coordinates of the point D will be larger, the point will move to the top right of M, the area G will be reduced, while the H area will be expanded. The initial cost are related with the cooperation

environment, in a good environment for cooperation, manufacturers or suppliers attach importance the benefits of collaborative, the desire for cooperative are strong and sincerity, easy to find and evaluate partners, will lower the cost paid. In addition, the initial cost but also with the business of technology, equipment and other factors, enterprises only continue to improve technology, updated equipment in order to reduce initial costs and to ensure supply chain companies to establish strategic partnership and stability.

136.3.3 The Discount Factor of Suppliers and Manufacturers

The discount factor δ_s, δ_m may be understood as the degree of dependence or attention of suppliers and manufacturers par for cooperates on future excess profits generated. The greater the discount factor, indicating that future earnings on the game the greater the effectiveness of both parties, and when the discount factor decreases, indicating that more attention on the immediate interests of both parties. Learn from the phase diagram, when the C_s, C_m and ΔV is constant, the greater the value, the more attention paid by two sides on the future cooperation, the greater the area of G above the line, and the system s will have more probability to convergent to point C. On the contrary, the interests of both partners take immediate attention to opportunistic behavior, will not be conducive for the system evolution to full cooperation.

136.4 Conclusions and Discussions

The paper uses evolutionary game theory to study the supply chain evolution of cooperative and competitive mechanisms, it was found that the direction of the evolution of the system and game payoff matrix related parties, and subject to the initial state of the system. In addition, the co-generated excess profits, the two sides due to the initial cost of investment cooperation, as well as its discount factor affecting the evolution of supply chain cooperation and competition between the important parameters. Interest only to comply with the maximum of cooperation, establish a good working environment and adhere to long-term perspective, supply chain enterprises establish and maintain a healthy partnership, to achieve “win-win” state.

References

1. Arunachalam R, Sadeh N, Eriksson J, Finne N, Janson S (2003) A supply chain management game for the trading agent competition. *Electronic Markets*
2. Christina P (2007) Finite populations choose an optimal language. *J Theor Biol* 349:606–616

3. Michael A (2008) Asymmetric evolutionary games with nonlinear pure strategy payoffs. *Games and Economic Behavior* 63:77–90
4. Mizuho S, Toshio Y (2007) Punishing free riders: Direct and indirect promotion of cooperation. *Evol Human Behav* 28:330–339
5. Wang H, Guo M, Efstathiou J (2004) A game theoretical cooperative mechanism design for a two-echelon decentralized supply chain. *Eur J Oper Res* 157:372–388
6. Maynard SJ, Price GR (1973) The logic of animal conflicts. *Nature* 246:15–18
7. Cressman R (1992) *The stability concept of evolutionary game theory*. Springer-Verlag, Berlin Heidelberg
8. Friedman D (1991) Evolutionary games in economics. *Econometric* 59:637–666

Chapter 137

A Method of Function Modeling Based on Extenics

Jing Zhang, Ping Jiang, Jiaqi Wang, Runhua Tan and Yang Yang

Abstract Because it is always very hard to choose the best solution in conceptual design, this chapter will combine extenics with functional design. First of all, the overall function is decomposed and solved, for a number of solutions. Then, the contain system method which is included in expand analysis method is used to describe the functional relationship, establishing the functional method based on extenics. Finally, the optimal degree evaluation method is used to determine the merits of the system, thereby improving the function design of system. And this chapter takes spring coiling machine for example to prove the feasibility of this process.

Keywords Function modeling · Extension theory · Spring coiling machine

137.1 Introduction

The conceptual design process of a product includes a number of complex reasoning and decision process [1]. Functional design is a very important part during conceptual design. In the design, the functional decomposition is the key step.

J. Zhang (✉) · P. Jiang · J. Wang · R. Tan · Y. Yang
Hebei University of Technology, Tianjin, China
e-mail: zjsuper111@126.com

P. Jiang
e-mail: jiangping@hebut.edu.cn

J. Wang
e-mail: wangjiaqi041@163.com

R. Tan
e-mail: rhtan@hebut.edu.cn

Y. Yang
e-mail: eryang053516@163.com

However, the decomposition process finished mainly depends on the experience of the designer [2]. This will cause many uncertainty and unstructured factors, therefore, only if we consider the uncertainty of mechanical design, can we design a more rational, economic and applicable products.

At present, during the function designing, choosing the best solution needs designers' experience. This is the most important reason which causes the uncertainty factors. Based on this problem, this chapter proposes a method of functional modeling based on extenics.

137.2 A Method of Function Designing Based on Extenics

Goals of functional design are to provide the reliable modeling method and application tool through mapping between functional specification of the products and geometric shape. Those carry out automatic deducing process from the functional requirement input of the products to the design structure output of the products. Function modeling is based on the denoting, contact and restriction of the various elements from conceptual functional analysis to geometry design. Function modeling is the base of the functional design. Based on those characteristic, the Extenics is combined with function designing, obtaining many solutions. Then, the Priority Degree Evaluation Method is used to choose the best solution, convenient for designers to obtain the best solution and make consumers know about the product's functions easily. The functional design process based on extenics is seen in Fig. 137.1.

137.2.1 Establish Functional Model

137.2.1.1 Function Decomposing

Function is the conversion between input streams and output streams, whose flow include material flow, energy flow and information flow. Function decomposing began in the description of the total function of production. Then, separate the total function into the sub-function that the complexity is low and easy, reducing the complexity of the problem and facilitating the solution of product's functional feature.

There are many methods of function decomposing, whether is function tree method or task list, can be used to establish the function model. Here use the function tree method to decompose the overall function. It contains two ways: one is the systematic top-down function analysis system technique, the other is the bottom-up [3].

137.2.1.2 Function Solving

After establishing a functional model, you need to find the physical structures to realize the sub-functions. This is the process of function solving.

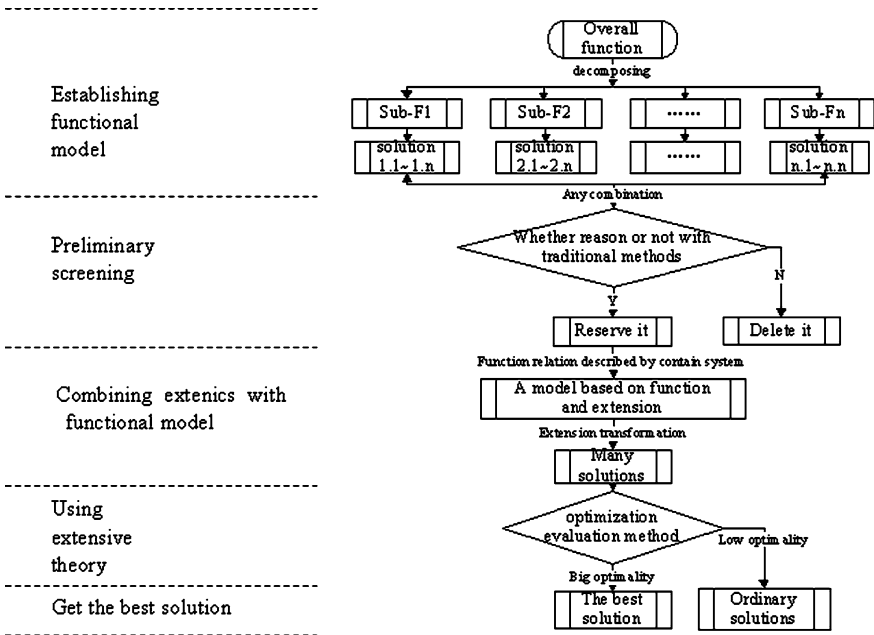


Fig. 137.1 The functional design process based on extenics

Function solving is realized through conventional methods. They contain information gathering, analysis of natural system, analysis of existing technical systems, analogies and measurements and model tests [3]. For a given function, according to the correlation of function and effect, effect and structure to do the functional solving.

137.2.1.3 Combining Solutions

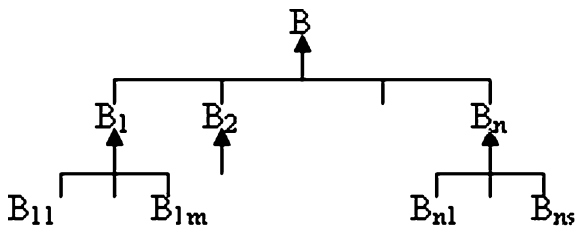
When the solutions for sub-function are available, they have to be combined in order to arrive at an overall solution. The methods for combining solutions are systematic combination and combining with the help of mathematical methods. The following example will use mathematical methods to combining solutions.

137.2.2 Theoretical Basis of Extension Engineering—Extension Theory

137.2.2.1 Introduction of Extension Theory

Matter-element is the basic logic unit for the extension theory to understand and analyze things. The given name of thing is N , and its value about feature c is v . Orderly triad $R = (N, c, v)$ is the primitive to describe things, for short

Fig. 137.2 The general form contain system



Matter-element. One thing has many features, if use features c_1, c_2, \dots, c_n and corresponding value v_1, v_2, \dots, v_n to describe the thing, then it's expressed as Eq.137.1

$$R = \begin{bmatrix} N & C_1 & V_1 \\ & C_2 & V_2 \\ & \vdots & \vdots \\ & C_n & V_n \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_n \end{bmatrix} \tag{137.1}$$

Then, R is called n-dimension matter-element, whose brief note is $R = (N, C, V)$.

Extension methods include the expanded analysis method, the conjugate analysis method, the extension transformation method, extension set method, excellent degree evaluation methods and the extension thinking mode [4–6]. This article mainly uses the contain system method in expand analysis method.

137.2.2.2 Contains Analysis Theory

Definition 2.1 Set B_1 and B_2 are primitives, if B_1 can achieve, then B_2 would achieve too. So the primitive B_1 contains B_2 , written for $B_1 \rightarrow B_2$. “ B_i achieving” is written for “ $B_i@$ ” ($i = 1, 2$).

Principle 2.1 If $B_1 \rightarrow B_2, B_2 \rightarrow B_3$, then $B_1 \rightarrow B_3$, then written for $B_1 \rightarrow B_2 \rightarrow B_3$

Corollary 2.1 (1) If $B11 \wedge B12 \rightarrow B1, B21 \wedge B22 \rightarrow B2$, and $B1 \wedge B2 \rightarrow B$, then $B11 \wedge B12 \wedge B21 \wedge B22 \rightarrow B$;

(2) If $B11 \vee B12 \rightarrow B1, B21 \vee B22 \rightarrow B2$, and $B1 \vee B2 \rightarrow B$, then $B11 \vee B12 \vee B21 \vee B22 \rightarrow B$.

Corollary 2.1 shows that the system established in this corollary is called primitive contain system. Its general form is shown in Fig. 137.2 [7].

137.2.3 The Relations of the Contain System Method and Functional Modeling

Both are used to solve contradictions problem, although aimed at different goals, but solution similar, in the end, they all obtain many solutions. Functional modeling method is only to function, while the contain system method to the

matter-element, thing-element and relationship-element, can solve the problem more widely, In view of the advantages of both, make them incorporation combine, use the contain system method to solve function problem.

The key idea of this chapter is finding the similar features of these two models. Only this, can they be combined, and can the following optimization evaluation method be used.

137.2.4 Optimization Evaluation Method

The basic steps of optimization evaluation method:

- 1) Determine Measure Indicators. The selection of Measure index must pay attention to the following principles: the purpose, all sidedness, feasibility and stability of evaluation.
- 2) Confirm Weight Coefficient. The weight coefficient shows the importance of each measure indicator. The indicators what must be satisfied use “ Λ ” to be said, and other measures are assigned between values 0 and 1, denoted by $\alpha_1, \dots, \alpha_n$. If $\alpha_{i0} = \Lambda$, then weight coefficient of the measure is 1.
- 3) First Evaluation. The index “ Λ ” is used to evaluate the objects, removing the objects unsatisfied with the index. The objects satisfied with the index “ Λ ” are made as following.
- 4) Establish correlation functions. Refer to the establishing process of the engineering example.
- 5) Calculate qualified degrees $K_i(x)$. The correlation functions is used to get the qualified degrees.
- 6) Regulate qualified degrees k_i . The biggest number of K_{ij} is divided by each number, getting k_i .
- 7) Calculate optimization. In the condition of each measure “ N_i ”, the standard qualified degrees is $K(N_i)$. So the optimization is $C(N_i)$ (see Eq. 137.2). $C(N_i) = \alpha_i K(N_i)$ (137.2).
- 8) Choose the best solution. Through the comparison of “ N_i ” optimality, one or several high optimality of the scheme [8] is selected as the best solution.

137.3 Case Study

The design of spring coiling machine is a system design, which make a wire-steel that the diameter is 0.3 mm manufacture into a spring that the diameter and pitch are both 2 mm. It can achieve the alignment, guided materials, orientation, rolling, confirming the diameter and pitch for the wire.

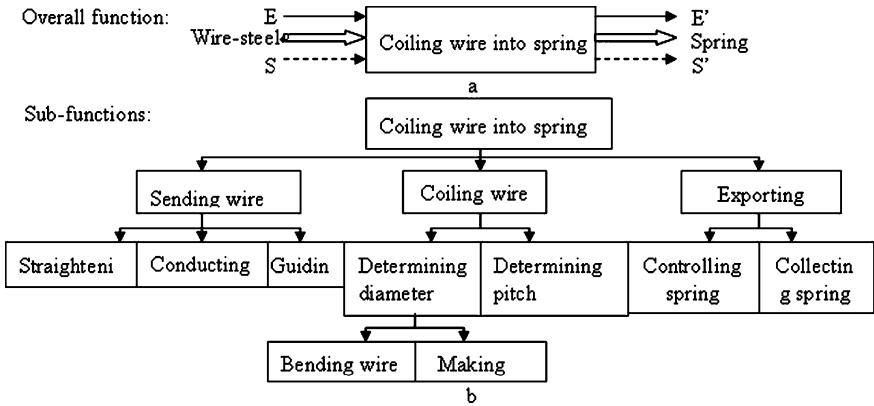


Fig. 137.3 The overall function a and the major sub-functions b about the spring coiling machine

137.3.1 Function Decomposing of the Spring Coiling Machine

The main structure of the spring coiling machine consists of the transmission section, the work portion and the support parts [9]. The transmission section of the spring coiling is made up of a motor, a belt and gears. The belt and gears can achieve gearing-down, so as to drive the rotation of the conducting wheel. The work port consists of the alignment bodies, conducting rollers, guide wire tubes, double-mandrel, a pitch block and a slitter knife for cutting the wire. These parts are all fixed on the stent. In work process, the wire-steel is driven by the rollers, and coiled by double-mandrel and a pitch block. Finally, we get a spring. The diameter of the spring is adjusted by the double-mandrel, and the pitch size of the spring is determined by the pitch blocks. Establish a overall function and a sub-function (see Fig. 137.3).

137.3.2 Solving Function

137.3.2.1 Principle Introduction

The initial materials what is made by wire-drawing, is changed into a diameter of 0.2 mm wire. What structure to be designed is achieving the function that change wire to spring. During the sending wire process, make sure wire is in line (horizontal or vertical), and the degree of tightness should be appropriate. During the coiling wire process, we should make the error of diameter and pitch as small as possible. Thus, the coiling wire process is the key to the design of this mold. There are two ways to coil wire: no core mold and core mold.

Fig. 137.4 Conical mandrel mold

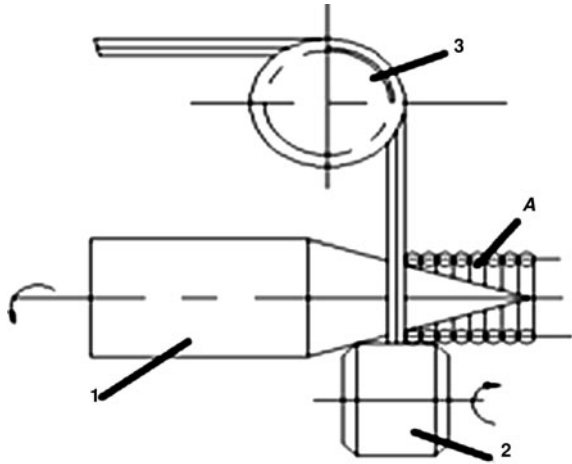


Fig. 137.5 Cylindrical spindle mold

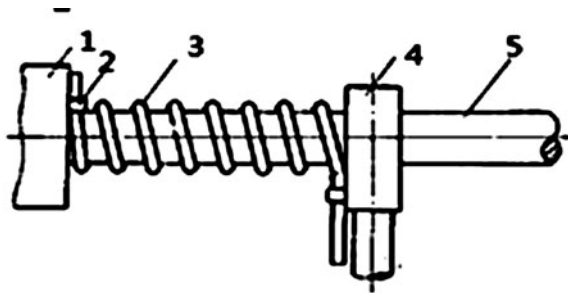
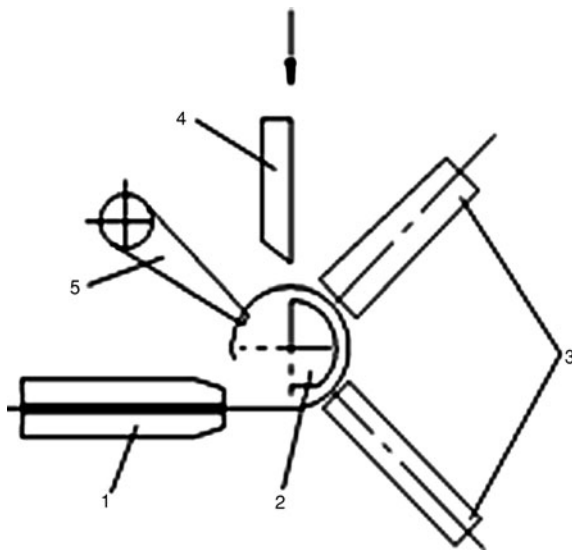
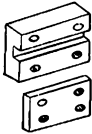

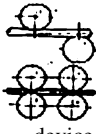
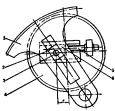


Fig. 137.6 Non-spindle mold



137.3.2.2 The Solutions of Sub-Functions (see Table 137.1)

Table 137.1 The solutions of sub-functions

Solution	Sub-function	1	2	3
1	Straightening F11	Straightening in level and vertical P111		
2	Conducting F12	 <p>Channel type device P121</p>	 <p>Patterns device P122</p>	 <p>Wheel device P123</p>
3	Guiding F13	Guiding slab P131 Guiding tube P132		
4	Bending wire F211	Screw (see Fig. 137.4-2) P2111	Twist bar (see Fig. 137.5-4) P2112	Mandrel (see Fig. 137.6-4) P2113
5	Determining diameter F212	Conical mandrel (see Fig. 137.4-1) P2121	Cylindrical mandrel (see Fig. 137.5-5) P2122	Non-spindle (see Fig. 137.6-2) P2123
6	Determining pitch F22	Screw (see Fig. 137.4-2) P221	Twist bar (see Fig. 137.5-4) P222	Pitch block (see Fig. 137.6-6) P223
7	Exporting spring F31	Smooth-plate orbit P311	Smooth semi-circular orbit P312	
8	Determining the length of the spring F32	 <p>Control mechanism P321</p>		

137.3.3 Establish a Model Based on Function and Extension

From Fig. 137.7, through calculation, there are 324 kinds of combinations to design spring coiling machine.

137.3.4 Excellent Degree Evaluation

- Measure indicators. Suppose Measure conditions is $V = (V1, V2, V3, V4)$.
 V1-Energy dissipation rate (EDR) of spring coiling machine: friction-free 0, rolling 0.1%, sliding 10%. 0–0.1% is the best, 10%–0.1% Is available, less than 10% can not satisfy the requirement; V2-Stability (S) of structure: 1 level is the most stable, 2 level is more stable, 3 level can't satisfy the

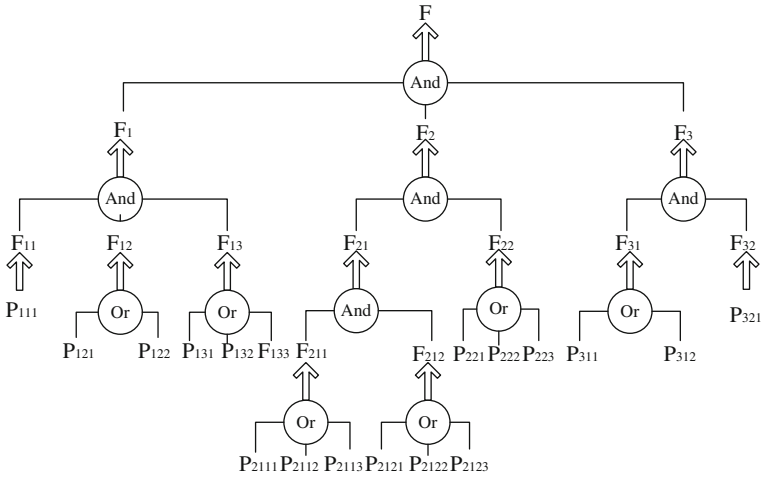


Fig. 137.7 A model based on function and extension

requirements; V3-Service Life (SL): more than 5 years; V4-Feasibility of spring coiling machine.

$$\begin{matrix} & V_1 & V_2 & V_3 \\ V_1 & \begin{bmatrix} 1 & 1/5 & 3 \end{bmatrix} \\ V_2 & \begin{bmatrix} 5 & 1 & 5 \end{bmatrix} \\ V_3 & \begin{bmatrix} 1/3 & 1/5 & 1 \end{bmatrix} \end{matrix} \tag{137.2}$$

- (2) Determine weight coefficient. According to the actual demand, V₄ must be satisfied in design, so its weight coefficient notes for \wedge . According to the design personnel expectations and needs of the customer, we can get judgment matrix of V₁, V₂, V₃, as follow:
Through AHP [10], we get weight coefficient vector is $W = (0.2021, 0.7007, 0.0992)$, so Weight coefficient vector of measure conditions V is $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0.2021, 0.7007, 0.0992, \wedge)$.
- (3) First evaluation. Doing the first evaluation with V₄, what don't satisfy this condition is undesirable solution. Because the functions of F₂₁₁, F₂₁₂ and F₂₂ only correspond to three structures that are conical mandrel mold, cylindrical spindle mold and non-spindle mold. And parts of one structure can't be combined with others'. There are 288 kinds of programs can't meet the feasibility, remaining 36 kinds of programs. According to measure conditions, we can make the 36 kinds of programs boil down to 10 kinds of systems (see Table 137.2).
- (4) Establish correlation functions.
The correlation function of V₁ is K₁(x) (see Eq. 137.3). The one of V₂ is K₂(x) (see Eq. 137.4). The one of V₃ is K₃(x) (see Eq. 137.5).

Table 137.2 The conversion from programs to systems

Function	Systems			R
	Part	EDR	S SL	
Conducting	a. Channel type	10%	(a.d.f.l)(a.d.f.o)(a.e.f.l)(a.d.g.l)(b.d.g.o)(b.e.g.l)	systemC ₁ EDR 15% S 2L SL 5Y
	b. Patterns	5%	(a.e.f.o)(b.e.g.o)	systemC ₂ EDR 15% S 1L SL 6Y
	c. Wheel	0.1%	(a.d.g.l)(a.d.g.o)(a.e.g.l)	systemC ₃ EDR 20% S 2L SL 5Y
	d. Guiding slab	2L	(a.e.g.o)	systemC ₄ EDR 20% S 1L SL 6Y
Guiding	e. Guiding tube	1L	cdgl, cdgo, cegl, adhl, adho, aehl, bdfi, bdflo, befl	systemC ₅ EDR 10% S 2L SL 6Y
	f. conical mandrel mold diameter	5%	(a.e.h.o)(b.e.f.o)(c.e.h.o)	systemC ₆ EDR 10% S 1L SL 7Y
Coiling wire and determining diameter	g cylindrical spindle mold	10%	(b.d.h.l)(b.d.h.o)(b.e.h.l)(c.d.f.l)(c.d.f.o)(c.e.f.l)	systemC ₇ EDR 5% S 2L SL 7Y
	h. non-spindle mold	0	(b.e.h.o)(c.e.f.o)	systemC ₈ EDR 5% S 1L SL 8Y

(continued)

Table 137.2 (continued)

Programs		Systems		
Function	Part	EDR	S	SL
Exporting spring	i. Smooth-plate orbit	2L	(c.d.h.l)(c.d.h.o)(c.e.h.l)	R
	o. Smooth semicircular orbit	1L	(c.e.h.o)	R

[systemC ₉	EDR	0.1%
S	SL	2L
SL	SL	7Y

[systemC ₁₀	EDR	0.1%
S	SL	1L
SL	SL	7Y

$$K_1(x) = \begin{cases} \frac{x-0.001}{0.001}, & x \leq 0.001 \\ x - 0.1, & 0.001 < x \leq 0.1 \\ \frac{0.1-x}{1-0.1}, & x > 0.1 \end{cases} \tag{137.3}$$

$$K_2(y) = \begin{cases} 1, & y = 1L \\ 0.5, & y = 2L \\ -1, & y = 3L \end{cases} \tag{137.4}$$

$$K_3(z) = z - 5, \quad z > 0 \tag{137.5}$$

(5) Calculate qualified degrees.

$$K1(x) = (K1(C1), \dots, K1(C10)) = (-0.055, -0.055, -0.11, -0.11, 0, 0, -0.05, -0.05, 0, 0)$$

$$K2(y) = (K2(C1), \dots, K2(C10)) = (0.5, 1, 0.5, 1, 0.5, 1, 0.5, 1, 0.5, 1)$$

$$K3(z) = (K3(C1), \dots, K3(C10)) = (0, 1, 0, 1, 1, 2, 2, 3, 2, 2)$$

(6) Regulate qualified degrees. Regulating $K1(x)$, $K2(y)$ and $K3(z)$, can get the qualified degrees:

$$k1 = (-0.5, -0.5, -1, -1, 0, 0, -0.45, -0.45, 0, 0) = (k11, k12, k13, k14, k15, k16, k17, k18, k19, k10)$$

$$k2 = (0.5, 1, 0.5, 1, 0.5, 1, 0.5, 1, 0.5, 1) = (k21, k22, k23, k24, k25, k26, k27, k28, k29, k20)$$

$$k3 = (0, 0.33, 0, 0.33, 0.33, 0.67, 0.67, 1, 0.67, 0.67) = (k31, k32, k33, k34, k35, k36, k37, k38, k39, k30)$$

(7) Calculate optimization. The regulate qualified degree of $C1$ is $K(C1)$. The optimization is $C(C1)$.

$$K(C_1) = [k_{11} \quad k_{12} \quad k_{13}]^T = [-0.5 \quad 0.5 \quad 0]^T,$$

$$C(C_1) = \alpha K(C_1) = \alpha K(C_1) = (0.2021, 0.7007, 0.0992)[-0.5 \quad 0.5 \quad 0]^T = 0.2493$$

In the same way, the optimizations of others' are as follow. $C(C_2) = 0.6324$, $C(C_3) = 0.1483$, $C(C_4) = 0.5313$, $C(C_5) = 0.3830$, $C(C_6) = 0.7334$, $C(C_7) = 0.3259$, $C(C_8) = 0.7090$, $C(C_9) = 0.4168$, $C(C_{10}) = 0.7672$.

(8) Choose the best solution. Comparing these optimizations, we get: $C(C_{10}) > C(C_6) > C(C_8) > C(C_2) > C(C_4) > C(C_9) > C(C_5) > C(C_7) > C(C_1) > C(C_3)$. Therefore, $C(C_{10})$ is the best solution.

137.4 Summery

(1) Extenics is combined with functional design. Conclude a method of function modeling based on extenics. (2) Spring coiling machine is taken as an example to prove the feasibility.

Acknowledgments Authors are grateful to the Special Fund for Innovation From Ministry of Science and Technology of P.R.C. (2009IM020700), the National Natural Science Fund of P.R.C (No. 70972050), the Key Technologies R&D Program of Hebei Province (No.09212102D),the Key Technologies R&D Program of Hebei Province (No.10242116D) for funding this work.

References

1. Yan LI (2003) Creative thinking and computer aided product innovation. *Comput Integr Manuf Syst* 12:1092–1096
2. Stone R, Wood K, Crawford R (2000) A heuristic method for identifying modules for product architectures. *Des Stud* 21:5–31
3. Pahl G, Beitz W (1988) *Engineering design: a systematic approach*. Springer, New York, pp 29–185
4. Yanlin JIA (1993) *Modularization design*. Machinery Industry Publisher, Beijing
5. Wen C (1994) *Element model and application*. Science and Technology Literature Publishing House, Beijing
6. Haijun W (2004) Methods supporting product modularization process design for mass customization. *Comput Integr Manuf Sys* 10:1171–1175
7. Chunyan Y, Wen C (2007) *Extensive engineering method*. Science Academy Publisher, Beijing
8. Chunyan Y, Bin H (1999) In: *The application of extensive method in new product conception, system engineering theory and practice*, pp 122–123
9. Hui L (1987) *Mechanical manufacturing technology for spring*. Machinery Industry Publishing, Beijing, pp 100–208
10. Kai L (2009) In: *Application of Analytical Hierarchy Process to Integrate Evaluation of Eco-environment*. Environmental Science and Technology, p 184

Chapter 138

Application of MCR-ALS Computational Method for the Analysis of Interactions Between Copper Ion and Bovine Serum Albumin

Zhu Xin-feng, Wang Jian-dong and Li Bin

Abstract The multivariate curve resolution—alternating least squares (MCR-ALS) computational method has been applied to spectra data obtained in the study of the interaction between Cu(II) ions and bovine serum albumin (BSA). Fluorescence and UV-vis spectra of complex mixtures of individual BSA and the binary complexes with BSA resulted in significantly overlapping spectral profiles. Quantitative information about the various complex ligand–protein species formed was obtained with the resolution of the spectral two-way data matrices by MCR-ALS method. Individual spectra of the BSA, their binary complexes with BSA were extracted, and quantitative concentration profiles for each species in a particular interaction were constructed. Such analyses made it possible to interpret the role and behaviour of each reaction component. It was found that Cu formed a 6:1 Cu-BSA complex.

Keywords MCR · MCR-ALS · Cu · BSA

Z. Xin-feng
College of Information Technology, Yangzhou University, Yangzhou, China
e-mail: zxfeng168@163.com

W. Jian-dong (✉)
College of Computer Science and Technology, Nanjing University of Aeronautics and
Astronautics, Nanjing, China
e-mail: aics@nuaa.edu.cn

L. Bin
College of Information Technology, Yangzhou University, Yangzhou, China
e-mail: lb@yzu.edu.cn

138.1 Introduction

Binding of metal ions to plasma proteins is an important pharmacological parameter, because very strong binding with albumin reduces the active concentration of the metal ions in the plasma, understanding of interactions of such metal ions with key biopolymers such as proteins and DNA, are of considerable significance, and appropriately methods of qualitative and quantitative analysis are required for this purpose.

Most metal ions are transported as a complex with Serum Albumin. Bovine Serum Albumin (BSA) or Human Serum Albumin (HSA), have similar tertiary structures and they both play very important roles in the binding of the metal ions. The amount of metal ion bound to the protein depends on the total ion concentration and its affinity for the protein; there is equilibrium between bound and free molecules.

The application of direct spectrofluorometric methods for the analysis of compounds in biological samples has been sparse because the measured fluorescence spectra are often complex and contain overlapping bands arising from the individual analytes. This precludes their identification and quantification of the analytes by direct spectrofluorometric methods. Recently, it has been demonstrated that such overlapping band problems can be resolved with the use of chemometrics [1], e.g. the multivariate curve resolution-alternating least squares (MCR-ALS) was used to resolve the overlapping spectra of the three components in the alpinetin-BSA system [2].

The aim of this paper was to combine the analytical capabilities of fluorescence and UV-vis spectroscopies supported by chemometrics for data interpretation to provide some quantitative and qualitative information about their modes of binding to BSA. This information is of significance for various applications, for example, in pharmaceutical, medical, veterinary, forensic, sports and body-building domains.

138.2 Materials and Methods

138.2.1 Apparatus

All fluorescence spectra were measured on a Hitachi F4500 Spectrofluorometer equipped with a thermostatic bath (Model ZC-10, Ningbo Tianheng Instruments Factory, China) and a 1.0 cm quartz cuvette. The excitation and emission slits were set at 10 nm, while the scanning rate was 1500 nm min⁻¹. The UV-vis spectra were measured on an SHIMADZU UV 2501 spectrophotometer. All the measurements were carried out at room temperature (25 ± 0.5°C) unless stated otherwise.

138.2.2 Materials

A stock solution of 1.0×10^{-3} mol/L Copper Ion was prepared by dissolving $\text{CuSO}_4 \cdot 5\text{H}_2\text{O}$ crystals (0.0169 g) in distilled water. Bovine Serum Albumin (BSA) (2.0×10^{-3} mol/L) was prepared by dissolving the purified protein ($M = 68,000$ Da; Bomei Biological Co. Ltd., Hefei, China) in 10 mL of 5.0×10^{-2} mol/L sodium chloride solution and stored at 4°C and dissolved to the needed concentration by using buffer solution. All experimental solutions were adjusted with the Tris-HCl ((hydroxy methyl) amino methane-hydrogen chloride) buffer to pH 7.4. Other chemicals were analytical grade reagents, and doubly distilled water was used throughout.

138.2.3 Procedures

The solutions used in the following experiments were prepared with 3.0 mL of pH 7.4 Tris-HCl buffer containing appropriate amounts of BSA and Cu. The total added volume (BSA and Cu) was less than 0.1 mL. Titrations were performed manually using suitable micropipettes. A mixed solution was shaken thoroughly and equilibrated for 10 min at 298 K. Fluorescence spectra were then measured in the range of 300–500 nm at the excitation wavelength of 280 nm.

The molecular fluorescence (300–500 nm) and UV-vis (220–400 nm) spectra were recorded every 0.5 nm and 1 nm, respectively. Thus, two sets of data matrices D_F^{CU} and $D_{\text{UV}}^{\text{CU}}$ were obtained. The row-wise spectral data matrices were combined, and two expanded data matrices, $[D_F^{\text{CU}} D_{\text{UV}}^{\text{CU}}]$, was thus, obtained. It was subsequently resolved by the MCR-ALS approach.

138.2.4 Chemometrics Methods (MCR-ALS)

The obtained data matrices (Sect. 138.2.3) were resolved by MCR-ALS, which is a multi-variate self-modeling curve resolution method developed by Tauler.²⁷ It involves the bilinear decomposition of the experimental data set, D , in order to obtain matrices, C and ST , which have a real chemical significance, according to Eq. (138.1):

$$D = CST + E \quad (138.1)$$

where D is the matrix of the experimental data, with dimensions of M (spectral objects) $\times N$ (wavelengths); C ($M \times F$) is the matrix of concentration profiles of the different F analytes in the samples; ST ($F \times N$) is the transposed matrix of emission spectra, whose F rows contain the pure spectra associated with the F

species in the samples; E ($M \times N$) is the matrix of residuals. The experimental spectral data matrix, D , is resolved as follows [3]: when the same chemical system is monitored using more than one spectroscopic technique (e.g. fluorescence and UV-vis spectra), a row-wise augmented data matrix can be built up. The individual data matrices corresponding to the two types of spectra are placed side by side. The related bilinear model for MCR-ALS analysis is shown in Eq. (138.2):

$$DFDUV = C[S_F^T S_{UV}^T] + [EFEUV] \quad (138.2)$$

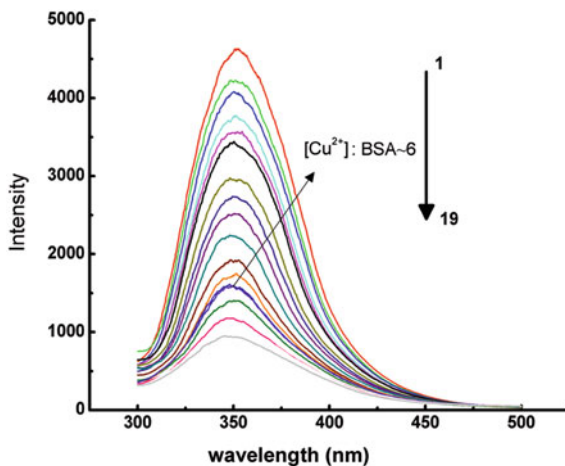
If DF and DUV are the measurements for the same experiment obtained with the two techniques, there is a single matrix of concentration profiles, C , valid for the two sets of raw measurements and a row-wise augmented matrix of spectra, where S_F^T and S_{UV}^T contain the pure spectra for the techniques used to obtain DF and DUV , respectively. Solving Eq. (138.2) for C and $[S_F^T S_{UV}^T]$ can help to extract the related spectral objects of all the species in the system [4].

When the matrix $[DF \ DUV]$ of the two data matrices is being built, the number of analytic species, F , can be obtained with the aid of EFA (Evolving Factor Analysis) [5] or PCA (Principal Component Analysis). EFA provides an estimation of the regions or windows where the concentration of different components is changing or evolving, and it also provides an initial estimation of how these concentration profiles change during the experiment. The EFA method is based on the evaluation of the magnitude of the singular values (or of the eigenvalues) associated with all the submatrices of a matrix built up by adding successively all the rows of the original data matrix. The calculation is performed in two directions: forward (in the same direction of the experiment) and backward (in the opposite direction of the experiment). The concentration profiles from EFA were used as the initial estimates for the concentration matrix input in the constrained ALS optimization [6]. PCA is a common chemometrics method of data analysis, which attempts to determine the number and direction of the relevant sources of variation in a bilinear data set or singular value decomposition (SVD) [7].

Then, to initiate the iterative ALS procedure, an initial estimate is needed for the spectral or concentration profiles for each species [3]. An initial estimate of the pure spectra can be obtained by the application of SIMPLISMA [8] and an initial estimate of the concentration profiles is obtained from the EFA plot. In this work, such initial estimates were made with the use of the ‘most pure concentrations’ method [9].

This kind of simultaneous data analysis is more powerful compared to that described by Eq. (138.1) and allows for improved resolution of very complex data structures [4, 10]. In general, the method produces more reliable solutions because it eventually removes rotational ambiguities and rank-deficiency problems as described in the literature [3].

Fig. 138.1 Spectra obtained from fluorescence experiment D_F^{Cu}



138.3 Results and Discussion

Figures 138.1 and 138.2 shows the fluorescence and UV-visible data of Sect. 138.2.3 After being excited at 280 nm, BSA showed a emission peak at 350 nm. This peak was intrinsic protein fluorescence which was caused by the amino acid side chains residues (tryptophan, tyrosine and phenylalanine). When small molecules bind with BSA, this fluorescence can be quenched. Copper ions (Cu^{2+}) is one kind of such ions, thus its interaction with BSA was studied by monitoring the intrinsic fluorescence intensity change of BSA after the addition of Cu^{2+} . Figure (138.2)a shows such spectra quench. With the increase of the amount of the added Cu^{2+} , the BSA fluorescence intensity decreased. The indole groups of the tryptophan residues are the dominant source of UV-vis absorbance and emission in proteins. The interaction of BSA with some small molecules sometimes would cause the exposure of the indole group, thus cause the increasing of the UV absorbance. Figure (138.2)b shows that the absorbance of BSA increased gradually as a function of increasing Cu^{2+} concentration. However, the obtained spectral profiles were complex and difficult to interpret, especially to estimate the amount of each component in the mixture. So the MCR-ALS method was applied to extract further information. Fluorescence and UV-vis spectroscopic data (Fig. 138.1, experiments 1 and 2 in Sect. 138.2.3) obtained from the interaction between each Cu^{2+} and the BSA, were used to build the expanded matrices.

Simultaneous methods of analysis for the complex ligand–protein species formed during the interactions of the copper ion and BSA, were developed with the use of fluorescence and UV-vis spectroscopic techniques and chemometrics methods MCR-ALS and PARAFAC. The binary complexes with BSA and their ternary complexes were extracted from complex mixtures of these substances in

Fig. 138.2 Spectra obtained from UV-vis experiment D_{UV}^{Cu}

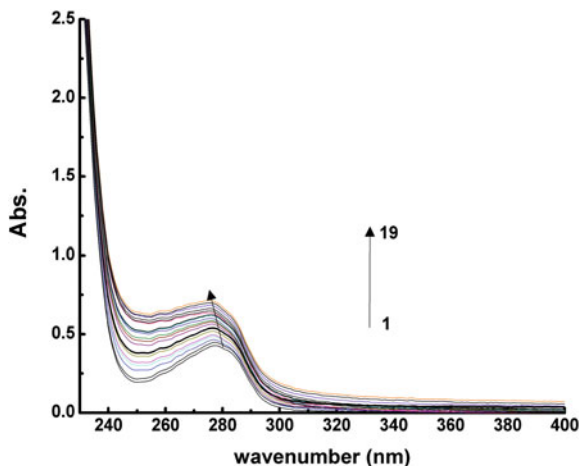
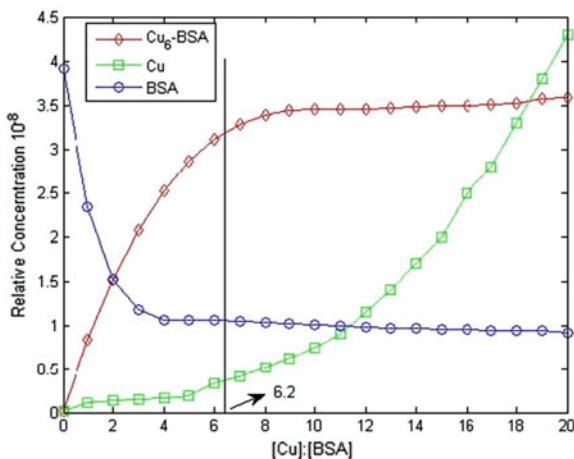


Fig. 138.3 Recovered concentration profile results of the simultaneous analysis of the UV-vis and fluorescence data matrices



solution. Fluorescence and UV-vis spectral data matrices were simultaneously analysed with the use of the MCR-ALS method.

The concentration profiles recovered by the MCR-ALS method (Fig. 138.3) with the mole ratio method and keeping concentration of BSA constant, showed that for copper ion, with the mole-ratio method, the concentration of the intercalation complex increased sharply and reached the equilibrium when $rCu:BSA \approx 6$ (Fig. 138.2). From above result it would appear that for the $Cu \leftrightarrow BSA$ interaction, the equilibrium position suggested an (Cu) 6-BSA complex formation. This research work is meaningful to the judgement of the interactions and their cohesion ratio in the mixture of metal ions and BSA.

Further more, acceleration of the alternating least squares algorithm has been proposed Masahiro Kuroda [11], and many regularized methods are also introduced, such as MCR-LASSO [12], regularized alternating least squares[13]. In the future, we will deeply research on the MCR-ALS with some improvement skills mainly depend on regularization methods.

Acknowledgements This work is supported by the Jiangsu Key Laboratory of Environmental Material and Engineering (K08021), National Natural Science Foundation of China (No.61070133), the Natural Science and Technology Foundation of Jiangsu Province of China (No.BK210311) and the Natural Science Foundation of Jiangsu Province of China (No. BK2009697).

References

1. Escandar GM, Damiani PC, Goicoechea HC, Olivieri AC (2006) A review of multivariate calibration methods applied to biomedical analysis. *J Microchem* 82:29–42
2. Ni Y, Wang SK, Serge (2010) Spectrometric study of the interaction between Alpinetin and bovine serum albumin using chemometrics approaches. *Anal Chimica Acta* 663:139–146
3. Vives M, Gargallo R, Tauler R (2000) Multivariate extension of the continuous variation and mole-ratio methods for the study of the interaction of intercalators with polynucleotides. *Anal Chim Acta* 424:105–114
4. Jaumot J, Gargallo R, de Juan A, Tauler R (2005) A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB. *Chemomet Intell Lab Syst* 76:101–110
5. Gampp H, Maeder M, Meyer CJ, Zuberbühler AD (1985) Calculation of equilibrium constants from multiwavelength spectroscopic data-III Model-free analysis of spectrophotometric and ESR titrations. *Talanta* 32:1133–1139
6. Abdollahi H, Mahdavi V (2007) Tautomerization equilibria in aqueous micellar solutions: A spectrophotometric and factor-analytical study. *Langmuir* 23:2362–2368
7. Sanchez Perez Isidro, Culzoni Maria J, Siano Gabriel G (2009) Detection of unintended stress effects based on a metabonomic study in tomato fruits after treatment with carbofuran pesticide. Capabilities of MCR-ALS applied to LC-MS three-way data arrays. *Anal Chem* 81:8335–8346
8. Windig W, Guilment J (1991) Interactive self-modeling mixture analysis. *Anal Chem* 63:1425–1432
9. Azzouz T, Tauler R (2008) Application of multivariate curve resolution alternating least squares (MCR-ALS) to the quantitative analysis of pharmaceutical and agricultural samples. *Talanta* 74:1201–1210
10. Ni YN, Su SJ, Kokot S (2008) Synchronous fluorescence and UV-vis spectroscopic studies of interactions between the tetracycline antibiotic, aluminium ions and DNA with the aid of the Methylene Blue dye probe. *Anal Chim Acta* 626:130–146
11. Kuroda Masahiro, Mori Yuichi, Iizuka Masaya, Sakakihara Michio (2011) Acceleration of the alternating least squares algorithm for principal components analysis. *Comput Stat Data Anal* 55:143–153
12. Pomareda Víctor, Calvo Daniel, Pardo Antonio, Marco Santiago (2010) Hard modeling multivariate curve resolution using LASSO: application to ion mobility spectra. *Chemomet Intell Lab Syst* 104:318–332
13. Andrzej Cichocki, Rafal Zdunek (2007) Regularized alternating least squares algorithms for non-negative matrix/tensor factorization. *Advances in Neural Networks*. Springer

Chapter 139

Self-Learning Algorithm for Visual Recognition and Object Categorization for Autonomous Mobile Robots

Anna Gorbenko and Vladimir Popov

Abstract In order to execute tasks and to navigate in an environment, an autonomous mobile robot needs a complex visual system to cope with detection, characterization and recognition of places and objects. We are interested here in the development of detection and characterization functions, integrated on a robot. In this paper we consider an approach to the development of categorization systems based on building by a robot of its own semantics, which used only by the robot and is not designed for human perception.

Keywords Neural network · Genetic algorithm · Visual recognition and object categorization

139.1 Introduction

In order to execute tasks and to navigate in an environment, an autonomous mobile robot needs a complex visual system to cope with detection, characterization and recognition of places and objects. We are interested here in the development of detection and characterization functions, integrated on a robot. Visual recognition problems can be divided into two main categories, object recognition and object class recognition. The former involves the efficient detection of a specific object with certain visual attributes. In the latter category we are not looking for a specific

A. Gorbenko (✉) · V. Popov
Ural State University, 620083, Ekaterinburg, Russia
e-mail: gorbenko.aa@gmail.com

V. Popov
e-mail: Vladimir.Popov@usu.ru

instance within a class, but we rather try to learn an internal model that corresponds to all objects of a class. The concept that labels an object class can be as wide or restricted as needed. Currently, the categorization of general objects or scenes is an active research area in computer vision society to realize helper robots and human assisting vision systems (see [1]). The categorization problem in object recognition is the assignment of semantic categories to objects or parts of objects. Categorizing visual elements is fundamentally important for autonomous mobile robots to get intelligence such as novel object learning and topological place recognition. The main difficulties of visual categorization are two folds: large internal and external variations caused by surface markings and background clutters, respectively.

A common approach involves the construction of a visual vocabulary which is used to express the content of the image, and the application of text retrieval techniques for inference purposes. In [2] developed a prototypic categorization system which classifies unknown objects based on their visual properties to a corresponding category of predefined domestic object categories. This system uses the bag of features approach which does not rely on global geometric object information. A major contribution of this work is the enhancement of the categorization accuracy and robustness through a combination of a set of supervised machine learners which are trained with visual information from example objects. This system is integrated on a mobile service robot to enhance its perceptual capabilities, hence computational cost and robot dependent properties are considered as essential design criteria. Note that the bag of features approach is often used (see [3]).

Many learning functions required on a robot, have been studied in the robotics community, taking advantage of the Bayesian framework to deal with uncertainties and noisy measurements, e.g. the automatic construction of stochastic maps to represent landmarks or features detected for robot localization from visual or laser data [4], the autonomous learning of topological models from panoramic images, including the place categorization [5], the automatic construction of a geometrical model for a 3D object to be grasped by the robot, the active learning of local structures from attentive and multi-resolution vision, etc.

Self-learning is a commonly used technique for semi-supervised learning. In self-learning a classifier is first trained with the small amount of labeled data. The classifier is then used to classify the unlabeled data. Typically the most confident unlabeled points, together with their predicted labels, are added to the training set. The classifier is re-trained and the procedure repeated. Note the classifier uses its own predictions to teach itself. In [6] self-learning applied to object detection systems from images. Perhaps, the best categorization performance provide mammalian brain functions that motivate to partly mimic such a formation for the application to machine learning and automation. Curiosity and experimental willingness support human recognizing and detecting in relation to still unknown objects. In order to learn an understanding of the topology of such objects, objects are observed from series of different viewpoints. These results in a collection of object views that afterward can be used by cognitive processes. Intelligent mobile

robots should have visual perception capability akin to that provided by human eyes. Currently, many researchers have tried to develop human-like visual perception capabilities such as self-localization and object recognition for intelligent mobile robots (see [7]).

In [8] proposed a model of the development of visual object recognition, based on the combination of two different artificial neural architectures, both supporting self-organization.

In [9] studied the one-shot and zero-shot learning problems, where each object category has only one training example or has no training example at all. An approach of [9] to this problem is based on transferring knowledge from known categories to new categories via object attributes. In [9] noted that with the increasing number of real training examples, the improvement on classification due to the prior knowledge decreases accordingly. This suggests that attributes do not contain all the information in target categories. Furthermore, some attributes may be difficult to learn and some are less informative to the categories. Thus when we have sufficient number of real training examples, the prior knowledge behaves more and more like noise and inevitably degrade the classification performance.

Note that the problem of stimulating of self-learning process is very common (see [10]). This is due to the methodology based on human semantics. In most cases, categorization systems are strictly dependent on human semantics. For example, it is interesting to note that in [10] the fact that the same human-semantics can be assigned to a set of images, does not exclusively imply that these images belong to the same visual object class. The final output of the process, is a number of clusters (unknown in the beginning) that correspond to the coherent Visual Object Classes present in our Data. However, in [11] supposed that human semantics can then be applied to these clusters.

In this paper we consider an approach to the development of categorization systems based on building by a robot of its own semantics, which used only by the robot and is not designed for human perception.

139.2 Principles of Learning and Self-Learning

First of all, our algorithm must be able to recognize basic properties of the object. Among these properties, in case of simple object we can note the following: 2D-shape of the object; color of the object; dimensions of the object. The algorithm obtains knowledge about these properties by supervised learning of neural networks from the training set. We use separate neural networks for each value of each of the three properties. It is obvious that we can not provide training sets for all values for these properties. Therefore, initially we train neural networks only for a limited set of values. After that, the robot itself obtains new images. From these images the robot creates new training sets. New neural networks are trained on the basis of unrecognized areas of images. New neural network define new

values of properties. These values do not necessarily correspond to a human semantics. For example, if {red, blue, green, yellow} is a color set of some image and {red, blue} is a set of known colors then the new color will correspond to some set of green shades and yellow shades. Note that values of properties are usually studied only partially. Therefore, in general, the new color will also contain some shades of red and blue. Some other properties are determined only by conditions on outputs of neural networks and methods that the robot gets new images for new training sets. In particular, we can set the number of images, trajectory, time of capture, frequency, etc. Examples of such properties are distribution of points of the object in the neighborhood, texture, luminosity, brilliance, brightness, number of holes, connectivity, etc. After the generation of neural networks for values of just discussed properties robot begins to self-generate new properties. We use knowledge bases to store information about studied objects. Information about each object is represented as a list of its properties. This information can be incomplete. In particular, information about one object can contain only its color, and for another are all possible properties. At first, knowledge bases are some predefined initial data. Then the robot itself adds new features and objects. In addition the robot can create a new database for some new property or to optimize the use of data. Using the knowledge base allows to analyze existing data, to aggregate features of objects, or to construct new objects from several known features. In particular, as a mechanism to define a new property, we use a genetic algorithm that generates a new neural network. This neural network should be divided into two halves a sufficiently large set of images such that none of known properties can not divide it. As a result, we obtain two new properties that have no relation to human semantics but of considerable interest to optimize the robot perception. Another incentive for the creation of new properties is the presence of a small set of images that can not be categorized on the basis of known properties.

Now we consider methods of allocation of new objects. We suppose that some object is new if it is continuously presented on images for a certain period of time. The system detects some connectivity area which is distinctive from known objects. Process of exploration of new area we can divide on a number of stages. It can be described as the operation of sequential addition of properties. Consider an example of this process in detail. At first we try to recognize a new area according to its color. For this we can use existing instruments. In particular, neural networks which we have previously trained to recognize such color. Once we recognize the color of the object we are trying to recognize its shape. As in the previous case, we can apply a neural system. Note that in this process we can use others algorithms of image processing. In particular, we can apply one of deterministic algorithms. Algorithms can work also in spatial and possibly frequency domains. Once we have investigated another property of the object, the algorithm proceeds to the next property. Then the process continues. With each step the algorithm adds a new property to information about this object. After this, properties of new object are compares with properties of existing objects. Depending on this comparison there are two types of action. In some cases new object may be labeled as known object

but having some new property. In other cases it is absolutely new and do not met so far. In both cases the algorithm updates knowledge bases relatively to new properties. In addition in both cases the algorithm updates its knowledge base. In the first case it appends new properties to the existing objects. In the second case it adds a new object with a set of studied properties. Finally, system creates a new recognizer that is capable to recognize areas that contain new features.

It should be noted that objects in the image can be complex. This means that a detected object in the image consist of two or more parts. For example, two skittles standing near can be considered as one object painted by two different colors. Similarly the object can be bicolor. Obviously the self-learning algorithm should be able to separate objects on images or to construct them from the already known. Separation of one object from many other can be implemented by intelligent algorithms. Possibly it let to find some fast solution. But these algorithms have a heuristic nature and achievement of solution is not ensured. Because of it such type of algorithms can be useful only optionally. Performing in the background, the algorithm provides ability to partition or merging objects into one if it is necessary. This optional knowledge can also be used for the invention of new shapes of objects and an expanding of knowledge bases. If on images these objects appear again, the information about it will already be written in knowledge bases. The use of such knowledge is of an emergency nature. We use them only when we do not have enough information to make a decision.

139.3 Generation of New Neural Networks

In order to use knowledge about new studied objects and for better recognition and detection of them, we should to customize existing algorithm and programs. Usually operators of such systems should do it manually. Self-learning process based on the generation of new neural networks. For these goals we use the algorithm of automatic generation of artificial neural networks. General approach can be described as follows. As known, main parameters of learning of neural systems are a number of layers and numbers of neurons in each layer. Also it is known that if the number of neurons or layers is too large, then it can contribute to long-term learning of the neural network and to effect of learning unnecessary dependencies. Conversely, if the number of neurons is too small, then it can lead to a slow convergence of a learning process of the neural network and a poor performance of the pattern recognition. It is necessary to select the optimal structure of neural network. This provides a sufficiently high level of recognition and a reasonably fast learning process. In our approach to achieve this goal we use a genetic algorithm. We create and run a population of neural networks. A structure of each network is selected from a predefined domain. For example, our algorithm operates with neural networks with a number of layers from 3 to 20. In each layer there are from 5 to 30 neurons. The approach to the generation of new networks is

based on the usage of previously trained networks as teachers for the new network. For example, we have two neural networks. The first is able to recognize red squares. The second is able to recognize blue circles. Using these two networks as teachers we can generate a new neural network that can recognize red circles. The performance of the genetic algorithm for automatic generation of neural networks essentially depends from the number of layers, numbers of neurons in each layer, the structure of interneuronal connections, types of nonlinear functions, methods of the generation of initial weights, and the training step of the neural network. This genetic algorithm use values of these parameters as the external data generated by other genetic algorithm. Note that the genetic algorithm for the creation of neural networks starts anew each time. Auxiliary genetic algorithm for generating parameters works permanently. With each new run of the algorithm for neural networks it gives current best chromosomes. On the other hand performance for the generation of a neural network is used to adjust the fitness function of the algorithm for the generation of parameters.

139.4 Control of the Overall Performance

Our algorithm for visual recognition and object categorization uses a large number of neural networks. The consequence of this fact is its relatively low computational performance. In particular, when we tested the algorithm on a mobile robot with an onboard computer ASUS Eee PC 1,000 HE, processing of one image by all neural networks took time equal to approximately 6 min. Therefore, to increase computing performance, we replace neural networks by deterministic algorithms. In particular, we use the Fourier transform, wavelet transform and threshold filters. Generation of these algorithms is done automatically. Note that deterministic algorithms are used only when the neural network guarantees a high level of recognition and categorization. This allows the use of neural networks for training and verification of deterministic algorithms. It should be noted that it is of interest not only the automatic generation of new deterministic algorithms but also the automatic modification of existing ones. For example, consider the following case. On the first stage of work we have a wavelet transform algorithm. We use it for compression of patterns before it will be recorded to the knowledge base. Obviously different patterns can have different degrees of compression. In particular, a two dimensional pattern of skittle can be compressed in ten times for example, whereas a circle in twenty times. It is important to choose the optimal settings for each algorithm to process every studied object. Clearly, this improves system performance. To create or reconfigure a deterministic algorithm we use genetic algorithms. Existing neural networks are used to determine values of the fitness function of the genetic algorithm. Another genetic algorithm permanently operates in the background. It determines the start time of the first algorithm and algorithms need to be replaced or modified.

139.5 The Overall Mechanism of Functioning of the Algorithm

The functioning of the algorithm is based on an implementation of a market economy [12]. The algorithm used by a mobile robot having a wireless connection to a supercomputer. This allows a significant part of calculations are performed on a supercomputer.

139.6 Summary

In this paper we have presented the basic concepts of the self-learning algorithm for visual recognition and object categorization for autonomous mobile robots. The main advantage of this algorithm is based on a lack of compliance with human semantics that allows the algorithm to avoid the effect of damping of incentives.

References

1. Zhang J, Marszałek M, Lazebnik S, Schmid C (2007) *Int J Comput Vis* 73:213
2. Information on <http://www.sim.informatik.tu-darmstadt.de/simpar2010/ws/sites/DSR2010/03-DSR.pdf>
3. Zhang W, Surve A, Fern X, Dietterich T (2009) In: Danyluk AP, Bottou L, Littman ML (eds) *Proceedings of the 26th annual international conference on machine learning*. ACM press, New York, p 1241
4. Hayet JB, Lerasle F, Devy M (2003) In: *IEEE computer society conference on computer vision and pattern recognition*. IEEE computer society press, New York, p 313
5. Zivkovic Z, Bakker B, Kröse B (2005) In: *Proceedings of IEEE/RSJ international conference on intelligent robots and systems*. IEEE computer society press, New York, p 7
6. Rosenberg C, Hebert M, Schneiderman H (2005) In: *Proceedings of 7th IEEE workshop on applications of computer vision*. IEEE computer society press, New York, p 29
7. Goebel PM, Vincze M, Favre-Bulle B (2008) In: *Proceedings of IEEE international conference on emerging technologies and factory automation*. IEEE computer society press, New York, p 1370
8. Plebe A, Domenella RG (2005) In: *Proceedings of 5th workshop on self-organizing maps*, p 489
9. Yu X, Aloimonos Y (2010) In: Daniilidis K, Maragos P, Paragios N (eds) *Proceedings of 11th European conference on computer vision, part V*. Springer, Berlin, p 127
10. Melekhin VB (1984) *Cybernetics and systems analysis*, vol 20, p 600
11. Noulas A, Kröse BJA (2006) In: *Advanced school of computing and imaging*, p 354
12. Dias MB, Stentz A (2000) In: *Proceedings of the 6th international conference on intelligent autonomous systems*, p 115

Chapter 140

Time-Complexity of the Algorithm for Physical Ability Test

Feng Qiming

Abstract This paper introduces the time-complexity of the algorithm for the arrangement of physical ability test. The time arrangement is a bin-packing problem, a kind of heuristic hybrid genetic algorithm composed by the best-fit (BF) and GA is used to solve this problem, by this way, we can get the reliable result, which is relatively stable while the evolution generations is more than 45. In the other hand, the paper also makes a comparison between the genetic algorithm with some other approximation Algorithms from the point of time complexity.

Keywords Time-complexity · Time arrangement · Genetic algorithm · Comparison · Bin-packing problem · The best-fit

140.1 Introduction

In order to understand physical condition of each student, one school arranges the students of each class to take physical ability test according to its teaching plan. The test includes five items: height and weight, standing long jump, vital capacity, grip strength and step test. All of their information is automatically measured, recorded and kept by electronic instruments. The school gets three sets of instruments for measuring height and weight, one for standing long jump, one for vital capacity, two for grip strength and two for step test.

In the four items of height and weight, standing long jump, vital capacity and grip strength, average test time for each student on one instrument is 10_s, 20_s, 20_s,

F. Qiming (✉)

Department of Mathematics, Wuxi Institute of Commerce, Wuxi 214153, China
e-mail: feng_qiming@163.com

and 15 s respectively. In the step test, one instrument can test five students at a time, which needs $3_m 3_s$.

Before one item of each student is tested, he or she must enter personal information of student No. Averagely it needs 5 s. After finishing one student's test, the instrument will automatically enter the next student No. So if student No. is contiguous, it can omit entering time. Student No. of one class is always sequential.

Every day the school arranges two periods of test time: from 8:00 to 12:10 and from 13:30 to 16:45. The five test items are processed in a small-sized location where can admit at most 150 students. The test items do not have fixed sequencing. There are totally 56 classes to take the test. The specific number of students of each class is taken from the problem D of 2007 CUMCM.

The school requires all the students of one class to finish all the test items in the same time period. It also tries to save the students' waiting time as much as possible under the condition of minimum time periods needed in the whole test. The purpose of this paper is to illustrate the problem of arrangement of test time for each class with mathematical symbol and language, to work out a computation method for this mathematical problem, and to explain to school staffs and students of each class the plan of test time as clearly and intuitively as possible with a diagram.

140.2 Analyzing and Modeling

Analysis of Problem. The five test items (height and weight, standing long jump, vital capacity, grip strength and step test) are separately recorded as $i = 1, 2, 3, 4, 5$ while the 56 classes that attend the test are recorded as $j = 1, 2, \dots, 56$ respectively. The student number of Class j is x_j , the student number that can take one test item at a time is $n_i = 3, 1, 1, 2, 10$. The time needed for testing each item at a time is $\tau_i = 10, 20, 20, 15, 210$ (seconds). The average time for one student to pass each test is

$$\tau_i/n_i = 3.3, 20, 20, 7.5, 21(\text{seconds}).$$

The time (seconds) that Class j spends in testing item i is $t_{ij} = x_j \times \frac{\tau_i}{n_i} + 5$ (where the 5 s are the entering time for each class to start the test). According to the test time arranged by the school (8:00–12:10 and 13:30–16:45), we set the length of period of test time as

$$T_k = 11700 + \frac{1 - (-1)^k}{2} \times 3300, \quad k = 1, 2, \dots$$

Suppose only when finishing all the test items can the class leave the test place, and the transfer time among the test items is neglected. We can make a definition that the waiting time c_j of the students in class j is the difference between their

time of leaving and entering the testing place, and all the students' waiting time is to sum j of $x_j c_j$. In order to save the students' waiting time as much as possible, the following aspects should be considered:

Firstly, from the above discussion we know that average time for one student to pass various test items is different, while time of the step test is the longest. In order to save their waiting time, the instrument's spare time should be decreased as much as possible. This problem can be solved through proper sequencing of the test items. The item which costs the longest time should be arranged first, the others should be kept in order one by one. After finishing the first test item, the student can complete the following test items quickly, and the students followed by do not need to wait. Conversely, students may accumulate and their waiting time will increase. Therefore, students should first take step test, which costs the most time, and then take the test of standing long jump, vital capacity, grip strength and height and weight in order.

Secondly, in order to save students' waiting time, student No. should be put as successively as possible, so as to reduce time for entering student information. Since student No. of one class is continuous, it can save five seconds of entering time for one person if the test is taken according to student number. So we suppose one class take the test as a whole.

Thirdly, according to the above analysis, the test time (seconds) of class j is

$$t_j = x_j \times \frac{\tau_5}{n_5} + 5 + \sum_{i=1}^4 \frac{\tau_i}{n_i}$$

(where the first item is the time that the class passes the step test which costs the most time, the third item is the time that the student spends for passing the remaining four testing items. If there is totally s classes of students who take the tests in the same period, the total testing time is

$$t = \frac{\tau_5}{n_5} \times \sum_j x_j + 5s + \sum_{i=1}^4 \frac{\tau_i}{n_i}$$

(where $\sum_j x_j$ is the total students of s classes, the second item is the sum of entering time for s classes). Time for one class to enter test place is the start time of this time period (the start time in the morning is 8:00) plus the sum of time that previous classes spend for passing the step test (the first and second item in t are the accumulative time that the s classes pass the step test). If the classes take the tests one by one in the order discussed above, students of each class may only wait before taking the step test. Without any other waiting, the whole waiting time is minimized. Meanwhile, the test site can only contain at most two classes, the capacity limitation of 150 students in the test place is sufficient. In order to ensure that the test is continuous and in order, the following class is required to enter the test site slightly ahead of time, so as to call the roll, line up according to the student No. and reaffirmation of attentions, etc. In brief, here we suppose this period of

time will not be calculated into students' waiting time and each class's entering time. Thus, $c_j = t_j$, and the waiting time of all the students is $T = \sum_{j=1}^{56} x_j t_j$.

Establishment of Model. Based on the above analysis, if taking one class as a unit, the students' waiting time will be the most economical when test is in order of step test, standing long jump, vital capacity, grip strength and height and weight. Then the problem can be concluded as: putting the time t_{5j} ($j = 1, 2, \dots, 56$) that each class used for passing the step test and the time that the last student used for passing the remained four test items into T_1, T_2, \dots, T_k , making k and the students' waiting time minimum. This is a bin-packing problem, whose mathematical model is as follows:

$$\begin{aligned} \min \quad & z = \sum_{i=1}^n y_i \\ \text{s.t.} \quad & \sum_{j=1}^n p_{ij} t_{5j} \leq y_i \left(T_i - \sum_{i=1}^4 \frac{\tau_i}{n_i} \right) \quad (i = 1, 2, \dots, n) \\ & \sum_{i=1}^n p_{ij} = 1 \quad (j = 1, 2, \dots, n) \end{aligned}$$

where, $n = 56$,

$$y_i = \begin{cases} 1 & \text{if some time is packed in to } T_i \\ 0 & \text{otherwise} \end{cases}$$

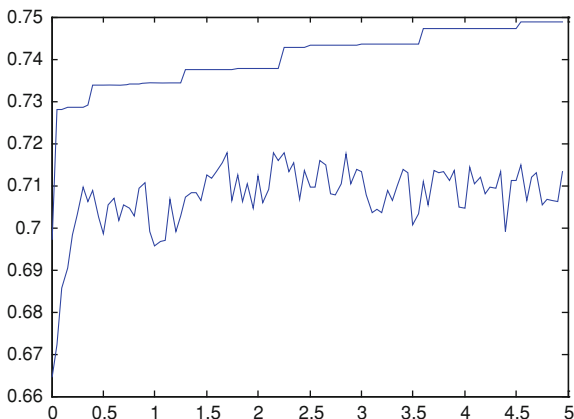
$$p_{ij} = \begin{cases} 1 & \text{if the time } t_{5j} \text{ is packed in to } T_i \\ 0 & \text{otherwise} \end{cases}$$

140.3 Design of Algorithm

The bin-packing problem is a difficult one of NP [1]. At present, only several heuristic algorithms aiming at this kind of problem are brought out. These algorithms are based on greedy algorithm, which is constituted by certain simple rules, such as next-fit (NF), first-fit (FF) and best-fit (BF), etc. Garey and Johnson indicate that the result of simple heuristic algorithm can not be worse (neither better) than the optimum solution multiplied by a littlish coefficient [2]. They have strong partial searching capability. Usually these algorithms are embedded in a genetic algorithm to strengthen the overall search capability, so as to find out the answer nearer to the optimum solution of bin-packing problem.

The genetic algorithm is classified as simple genetic algorithm and hybrid genetic algorithm. The shortcomings for simple genetic algorithm to solve the bin-

Fig. 1 Some Results



packing problem are that it may result several ineffective chromosome (that is the bin-packing program). This will make the volume sum of the contained goods in several boxes exceeds the regulated volume of boxes, thereby the calculation efficiency is reduced and the calculation result is not very good [3]. Here, we combine the best-fit, which is used to solve the bin-packing problem, with GA, to form a heuristic hybrid genetic algorithm. This kind of bin-packing program is in accordance with the idea of best-fit approximating algorithm; meanwhile the volume sum of contained goods in all the boxes will not exceed the regulated volume.

140.4 Computed Result

For the above problem, the MATLAB program could be used to achieve the single genetic algorithm. Its operation parameters are: the probability of transposition operator $p_e = 0.25$, the probability of shift operator $p_s = 0.25$, the probability of inversion operator $p_i = 0.25$. Based on the fitness, 75% of the filial generation is chosen to reinsert, the number of filial generation is one, the evolution generations are 100, and every population size is 50. The evolution process can be seen from “Fig. 140.1”.

In “Fig. 140.1”, the solid line indicates the variation of the largest target value of each generation. The dotted line indicates the variation of the mean target value of each generation. If the evolution generations are more than 45, the result will be more stable and will go toward the optimal value. The result shows that the physical ability test would take four testing periods; the optimal target value is 0.7489. At this time the corresponding programs are : classes 1, 9, 13, 14, 16, 24, 29, 34, 35, 36, 38, 39, 42, 44, 45, 46, 48, 51, 52, 56 and classes 2, 3, 5, 6, 10, 15, 18, 19, 23, 26, 27, 28, 32, 40, 47, 49, 53, 54, 55 are chosen to test in two groups in

Table 140.1 Time-Complexity of Approximation Algorithms

Algorithm	Time complexity
Next-fit (NF)	O(n)
First-fit (FF)	O(nlogn)
Best-fit (BF)	O(nlogn)
First-fit decreasing(FFD)	O(nlogn)

the morning. Classes 4, 7, 8, 11, 12, 17, 20, 21, 25, 31, 33, 37, 41, 43, 50 and classes 22, 30 are chosen to test in two groups in the afternoon. According to time for each class to take step test, we can get time for each class to enter test site.

140.5 Time Complexity of Some Approximation Algorithms

We have obtained the time complexity of first-fit heuristic algorithm. According to the steps of genetic algorithm, we can write a program in MATLAB for this algorithm. Suppose the time complexity of the genetic algorithm is denoted by $T(n)$, then

$$T(n) = O(N * NIND) + O(MAXGEN * NIND)$$

$$= \begin{cases} O(N * NIND) & N \geq NIND \\ O(MAXGEN * NIND) & N < NIND \end{cases}$$

where, the NIND is the number of individuals., and the MAXGEN is the maximum number of generations.

Similarly, we can also obtain the time complexity of the other approximation algorithms, Table 140.1 shows the time complexity of approximation algorithms.

Comparison of NF, FF and FFD heuristic algorithm, their approximate level of priority than a good one, but it does not mean that NF, FF lost value of its existence. FF, FFD heuristic algorithms require that all boxes may be shipped together until all goods are loaded, but the NF heuristic algorithm has no such restriction. FFD heuristic algorithm also requires that all goods arrive before starting to pack. However, if adopting NF, FF heuristic algorithm, we can not know how the length of the next goods, that is, goods can arrive one by one, without the need to know its length in advance before their arrival [4].

140.6 Summary

Common method of terminating the genetic algorithm is to preset maximum number of generation. If there is no acceptable answer, the genetic algorithm to restart or initial conditions with a new search. [5] Thus, in general, the MAXGEN is a constant.

When NIND and MAXGEN are much less than n , the time complexity of the genetic algorithm is $T(n) = O(n)$, where, $N = n$. Otherwise, $T(n) = O(n^2)$.

Therefore, we understand that, this genetic algorithm has the following merits:

Firstly, the number of boxes is not needed in advance. But the objects should be put into the boxes in turn. Only when the boxes are overloaded will another box be added. In this way each box will be filled as fully as possible.

Secondly, the convergence speed is faster, because evolution generation of 45 roughly can achieve a relatively stable result. The coded method proposed in this paper take objects as chromosome. Even at the beginning, the problem of overload and discontent will not appear. Therefore the efficiency is greatly enhanced.

Lastly, the greatest advantage of the chromosome structure representation based on the objects will never generate infeasible solution.

In order to make the algorithm simple, we assume that each student passes testing instrument at average time period. But in practice this is not the case. This is the inadequacy of this algorithm. We hope it can be solved in future research.

References

1. Garey M, Johnson D (1979) Computers and intractability: a guide to the theory of NP-completeness. W.H. Freeman, New York, pp 13–23
2. Garey M, Johnson D (1979) Computers and intractability: a guide to the theory of NP-completeness. W.H. Freeman, New York, pp 13–23
3. Mitsuo G, Runwei C (2004) Genetic algorithms and engineering optimization. Tsinghua University Press, Beijing pp 51–53 100-103
4. Eengyu Y, He Y, Shiping C (2001) Mathematical programming and combinatorial optimization. Zhejiang University Press, Hangzhou, pp 106–109
5. Yinjie L, Shanwen Z, Xuwu L, Chuangming Z (2006) MATLAB genetic algorithm toolbox and its application. Xi'an University Press UESTC, Xi'an, p 102

Part XIII
Data Management and Database System

Chapter 141

A Distributed Keyword Search Algorithm in XML Databases Using MapReduce

Yun Ling and Guangjian Xu

Abstract With the extensive use of XML technology, the using of XML databases which consist of mass structured and semi-structured XML documents has become increasingly popular. How to acquire the data to meet users' needs in XML databases quickly and efficiently has become an urgent problem. Current researches mainly focus on XML streams and XML documents. On the contrary, the keyword search algorithm in XML databases gets little attention. In this chapter, we combine the concept of keyword search algorithm-SLCA in XML documents and the characteristics of MapReduce to propose a distributed keyword search algorithm in XML databases, and implement it by open-source framework-Hadoop. Finally, sufficient experiments show that our method is efficient in practice in various aspects.

Keywords Mapreduce · Xml database · Keyword information retrieval · Hadoop · Slca · Algorithm

141.1 Introduction

With the extensive use of XML technology, how to obtain the information to meet users' needs quickly and efficiently in the XML databases which consist of mass structured and semi-structured XML documents has become a hot research.

Y. Ling (✉) · G. Xu

College of Computer Science and Information Engineering, Zhejiang Gongshang University, 310035 Hangzhou, China
e-mail: yling@mail.zjgsu.edu.cn

G. Xu

e-mail: xugg09@gmail.com

Current researches mainly focus on XML streams and XML documents: (1) using XPath/XQuery technology [1–4], (2) keyword-based queries [5–8]. However, there're few studies related to the XML databases which using XML documents as their basic logical storage unit. Such XML databases are often large, growing increasingly, and sometimes one business logic data is stored by many joint XML documents with various structures.

XPath/XQuery is a fully functional XML query language, which uses pre-acquired hierarchy of documents and path expression to search specified content in documents. When it is applied to the large-scale XML databases, how to get the complete hierarchies of all the documents has become a big challenge, which might lead to the unrealization of some complex queries. Even if we can get hierarchies of all documents in advance, the complicated hierarchies will lead to writing queries difficultly, optimizing queries hard, unfriendly user interfaces and other issues. In addition, the limited server resources can only deal with limited data queries, and can't satisfy the demand of fast queries over massive data.

Keyword-based search algorithms such as SLCA attempt to get the most compact fragment keyword tree through the operations upon XML stream's or document's directed tag tree. It does not require users to know the hierarchy of XML document in advance, it has a friendly user interface, and has made a good use in XML streams and XML documents. When this keyword search algorithm is used in large-scale XML databases, the algorithm will create a huge tag-tree and will take up a lot of system resources, which might result in inefficient or even unavailable query.

In this chapter, we combine the concept of keyword search algorithm for XML documents-SLCA [4] and the characteristics of MapReduce [9], put forward an efficient, fast and scalable distributed MapReduce-based keyword search algorithm and implements the algorithm through open source framework Hadoop [10, 11]. The sufficient experiments show that our method is efficient in practice in various aspects.

The rest of the chapter is structured as follows: [Sect. 141.2](#) introduces existing algorithms includes XPath/XQuery and keyword-based search algorithms. [Section 141.3](#) provides the problem model and the related concepts. [Section 141.4](#) describes the system structures and algorithm models. Experimental results and evaluations are reported in [Sect. 141.5](#). And finally [Sect. 141.6](#) concludes the chapter.

141.2 Related Works

XPath/XQuery is a XML data query language developed by W3C. It requires users to master complex query language, and know the hierarchy of XML document prior. The literature [1–3] focus on how to query in XML streams or XML documents effectively. The main methods include automatic machines based and stack-based. The goal for using XPath express query is quite clear, which is to find

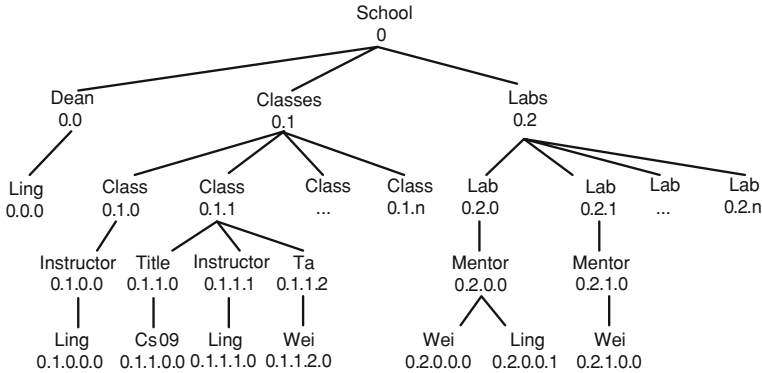


Fig. 141.1 School.xml (each node is associated with its Dewey number)

out a series of nodes that satisfy the query conditions to the user. The literature [4] focus on executing XQuery operations in XML data streams or XML documents, since the syntax of XQuery is more complex than XPath’s, it’s more difficult to deal with.

Relative to the XPath/XQuery, keyword-based queries in XML documents can return the XML documents that contain k according to the keyword sequence $k = \{k_1, k_2, \dots, k_n\}$ which is provided by users. Currently, most XML keyword searches are based on LCA technology which returns the most compact segment tree. The related studies also include ELCA, SLCA and so on. The representative literatures are [5] and [6, 7], [5] mainly discusses how to quickly find out the LCA node set that contains keyword in XML tree, and presents a method to divide ranks. Literatures [6, 7] mainly address how to find out the SLCA node set quickly in XML documents. The chapter considers that such a result is more useful to users. The concerns of the keyword search on XML documents is to find out the node set of LCA or SLCA. As shown in Fig. 141.1, we take a keywords search $Q = \{\text{Ling}, \text{Wei}\}$ on a tag directed tree School.xml which is with the Dewey encoding [12]. It can be seen from Definition 1 that LCA-based query will return the node set: $\{\text{Class}(0.1.1), \text{Mentor}(0.2.0.0), \text{Classes}(0.1), \text{Labs}(0.2), \text{School}(0)\}$. And it can be seen from Definition 2 that SLCA-based query will return the node set: $\{\text{Class}(0.1.1), \text{Mentor}(0.2.0.0)\}$.

141.3 Problem Model

MapReduce is a programming model for distributed and parallel computing on large-scale data sets on clusters of computers. Concepts Map and Reduce, and their main thoughts are borrowed from functional programming language and also borrowed from the features of vector programming language. In short, a mapping

function is to do specified operations on each element in the list which consists of some independent elements. In fact, each element is to be operated independently, which means that, Map operation is highly parallel, which is very useful in parallel computing domain and the applications which require high performance. Reduce operation is an appropriate merger over a list. It is not as parallel as Map function, but it always outputs a simple answer and the large-scale computations are relatively independent, so the Reduce function is also very useful in highly parallel environments.

Keyword-based search algorithm on XML documents has the following key concepts. Assume expression $v <_a v'$ represents node v is an ancestor node of node v' , and expression $v \leq_a v'$ represents node v is an ancestor node of v' or is node v' .

Definition 1 LCA (Lowest Common Ancestor) Given m nodes v_1, v_2, \dots, v_m , if the node w for each v_i ($1 \leq i \leq m$) satisfies $w \leq_a v_i$, while another node u does not exist for each v_i satisfies $u <_a v_i$, and $w <_a u$, then w is called an LCA, expressed as $w = \text{LCA}(v_1, v_2, \dots, v_m)$.

Definition 2 SLCA (Smallest Lowest Common Ancestor) Given m nodes v_1, v_2, \dots, v_m , if the node $w \in \text{SLCA}(v_1, v_2, \dots, v_m)$, then there is no node $u \in \text{LCA}(v_1, v_2, \dots, v_m)$ makes $w <_a u$.

Literatures [13–15] describe an approximate keyword research algorithm. The chapters point out that the smaller the sub-tree which contains all the keywords is, the closer the link among the keywords is. The results of the LCAs can be simplified based on this assumption. SLCA is based on this idea. In short, SLCA is to find out the node set which at least contains one set of query keywords, and its descendants can't contain the instance which has all the query keywords. Namely the returned results do not have the ancestral or descendant relationships.

From the Definition 1 and Definition 2 we can see that in order to implement a distributed LCA and SLCA operations on tree T , T must be split and the structural integrity of the tree fragment must be ensured. When we perform LCA and SLCA operations on the fragments, the node executed by SLCA must be one of the final result sets, and it will be easy to find SLCA nodes from the node sets executed by LCA. Through the methods above, we can get the final SLCA node set quickly and efficiently. Related definitions are as follows:

Definition 3 SCS (Structurally Complete Sub-tree) Suppose tree T_1 was the sub-tree of the tree T , and both T_1 and T have the same root node, then we claim that T_1 is a SCS of T , denoted as $T_1 = \text{SCS}(T)$.

Lemma 1 Suppose tree T_1 was the sub-tree of the tree T , then the $\text{SLCA}(T_1)$ is included in the $\text{SLCA}(T)$.

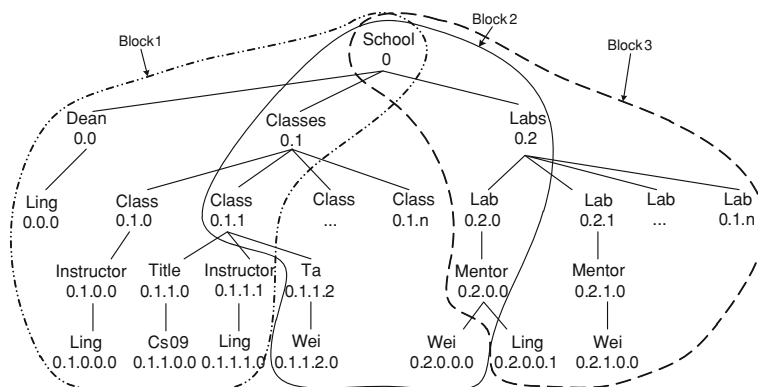


Fig. 141.2 The split of School.xml (each node is associated with Dewey number)

141.4 System Structures and Algorithm Models

141.4.1 Distributed Data Storage Model

We need to split the large XML document when storing it. Different from the generally structureless data, the Blocks split from XML document need the implementation of the structural integrity. Illustrated in Fig. 141.2, we store the Blocks of the School.xml according to the BlockSize of the distributed file system based on the premise that each Block is a SCS. The content of the Block1 will be:

```

<School> < 0
  <Dean> < 0.0
    Ling < 0.0.0
  </Dean>
  <Classes> < 0.1
    <Class> < 0.1.0
      <Instructor>< 0.1.0.0
        Ling < 0.1.0.0.0
      </Instructor>
    </Class>
    <Class> < 0.1.1
      <Title> < 0.1.1.0
        Cs09 < 0.1.1.0.0
      </Title>
      <Instructor>< 0.1.1.1
        Ling < 0.1.1.1.0
      </Instructor>
      <Ta> < 0.1.1.2
        Wei < 0.1.1.2.0
      </Ta>
    </Class>
  </Classes>
</School>
    
```

141.4.2 MapReduce-based Search Model

To facilitate the illustration, this chapter regards the MapReduce operations on HDFS as background, assuming one DataNode will only run one MapTask. Taking the distributed data shown in Fig. 141.2 as an example, Fig. 141.3 shows the computation process. MapTask1 performed in DataNode1, handling the Map operations of Block1 and Block2 which are stored on it. MapTask2 performed in DataNode2, and is responsible for Block3 and other Blocks. Through the Map operations on Block1, we will get the node set which contains the keyword “Ling”: {Ling(0.0.0), Ling(0.1.0.0), Ling(0.1.1.0)}. The execution of Map

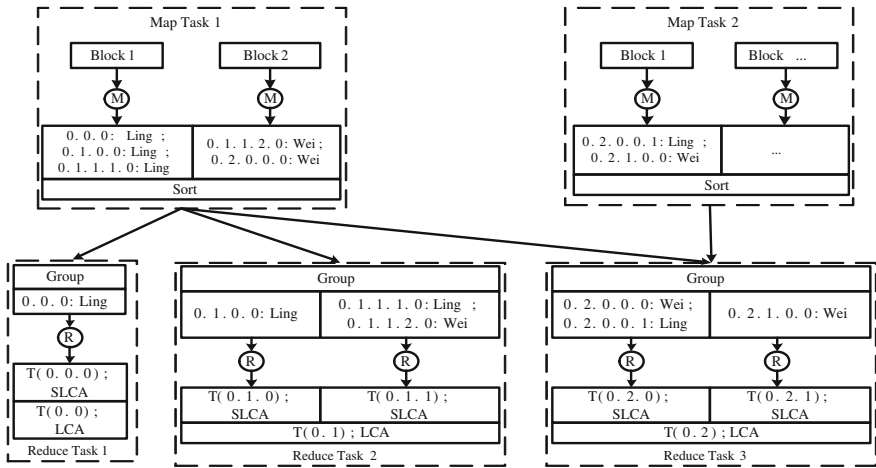


Fig. 141.3 The MapReduce-based keyword search in XML databases

operation on Block2 will get the node set which contains the keyword “Wei”: {Wei(0.1.1.2.0), Wei(0.2.0.0.0)}. Block3 gets similar results: {Ling(0.2.0.0.1), Wei(0.2.1.0.0)}. Expressed as the Key/Value form, we use Dewey code as Key, keywords and content as the Value. The purpose of doing so is to facilitate the work goes on Reduce. The last step of Map is Sort which is to summarize and sort the results of MapTask. The Sort results of MapTask1 will be: {0.0.0:Ling; 0.1.0.0:Ling; 0.1.1.1.0:Ling; 0.1.1.2.0:Wei; 0.2.0.0:Wei}.

The Map operations involve the specific problem of how to map. Here, we borrow the customizable parameter InputFormat in object JobConf of Hadoop to illustrate the problem. A calculation of the task in Hadoop is called Job, and JobConf is the configuration of this Job. InputFormat parameter is to specify how to split the input data set. We use a node sequence from the root to the node which contains the keywords or contents, or so-called sub-tree, as a small data set for Map operations rather than use the default settings with each line of text. The results after InputFormat of Block1 will be: {0.0.0:Ling; 0.1.0.0.0:Ling; 0.1.1.0.0:CS09; 0.1.1.1.0:Ling}. Block2 and Block3 will get similar results.

In the Reduce stage, Group operations are responsible for obtaining the required data and grouping it. Three Groups of ReduceTasks will obtain and group the data by the Dewey coding of its key which starts with “0.0”, “0.1” or “0.2”. Such as the Group in the ReduceTask2, it will obtain three piece of data whose Dewey coding starts with “0.1”, and the next layer data will be divided into two parts by the Dewey coding which starts with “0.1.0” and “0.1.1”, namely the result would be: {0.1.0.0:Ling; 0.1.1.1.0:Ling; 0.1.1.2.0:Wei}. Reduce function will execute multi-threaded operations on the data gotten from the above. Finally, in Fig. 141.4(1)(2), it will return the keyword sub-trees whose roots are “0.1.0” and “0.1.1”.

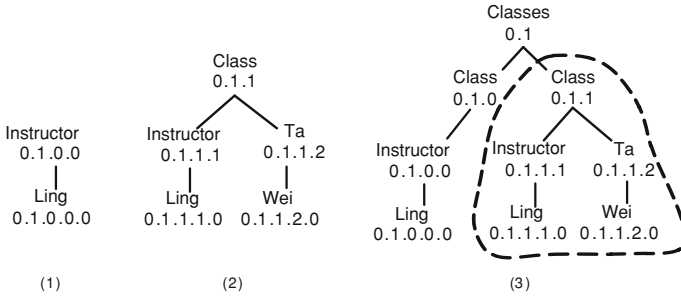


Fig. 141.4 The intermediate results of ReduceTask2

By Lemma 1 we know that the SLCA node {0.0.1:Class} which is found in the T(0.1.0) and T(0.1.1) will be one of the final results. Therefore, before going to the next step, we need to output the result set and remove the relevant nodes in the T(0.1.0) and T(0.1.1). Combined T(0.1.0) and T(0.1.1) will be T(0.1), as shown in Fig. 141.6(3) below. The LCA nodes found in T(0.1) will be output to the final summary filter, and the relevant nodes will be removed from T(0.1). Eventually, system will output: (1) SLCA node set, (2) LCA node set in addition to the SLCA nodes, (3) a keyword tree which consists of the nodes except the related of (1),(2). Finally, we execute LCA operations on (3), combine the result set with (2), find out SLCA result set in the combined LCA result set, and merge this SLCA result set with (1) to get the final result.

141.4.3 Indexing Strategies

In achieving the keyword search, we tried several indexes to improve search efficiency. (1) Global-based indexing. Global index is stored in the NameNode. When storing XML documents with Dewey coding, we set the code of node as a key of index, and set the keywords and the contents as a value of index. Such an approach can obtain a good efficiency facing the keyword search on a small amount of XML documents. However, when it faces massive data, the efficiency is low. (2) Local-based indexing. Local index is defined on all Blocks in the DataNode and stored in this machine. When the operations go to Reduce stage, index list should be passed to the machines which execute Reduces, or needs to translate the results of Map stage according to index list and pass it to the machines which execute Reduces. This operation will increase the load of the information transmission between Map and Reduce greatly. Thus it impacts the efficiency of overall system seriously. (3) The keyword search contains the path information of the keywords. This strategy is adopted by this chapter and it can address the issues mentioned above. We execute SCS operations for each Block to ensure the integrity of the sub-tree when storing XML documents. When the Map executes such Block, its path information will be

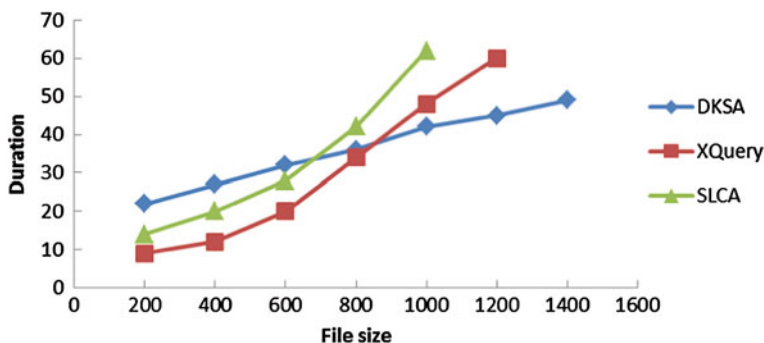


Fig. 141.5 The influence of file size

passed to the next step as a result of Map. For example, the Map execution on Block1 in Fig. 141.3, in addition to generating the result set of code or keyword, there is another result set whose key is a constant such as: “KeywordTreeStructure” and whose value is a keyword path information tree. The result of its left traversal is: {Ling(0.0.0), Dean(0.0), School(0), Ling(0.1.0.0.0), Instructor(0.1.0.0), Classes(0.1.0), Class(0.1), Ling(0.1.1.1.0), Instructor(0.1.1.1), Class(0.1.1)}. The results will be merged in the phase of Reduce. This approach, which is equivalent to the improvement of the local indexing, achieves better efficiency by passing the index information of the search results simply.

141.5 Experiments and Evaluations

Experimental environment: The Hadoop cluster configuration under LAN environment, 10 commercial PCs (Intel Core 2 Duo 2.5G CPU, 4G memory, 500G hard drive), the operating system is Red Flag Linux Desktop 6.0. The specific configuration is that 1 NameNode server and 6 DataNode servers. MapReduce setting is that every DataNode has a Mapper and two of them also are Reducer. The remaining three machines configured as DataNode server and are not added to sever group temporarily. The Test data: Simulation program generated 20 semi-structured XML documents which are named School.xml and each size is 100 M.

Figure 141.5 shows how the size of the file which is to be calculated influences the query duration of each search algorithm when the number of keywords is 2. It can be seen from the figure, when the file size is less than 700 M, XQuery is the most efficient, SLCA second and DKSA(Distributed Keyword Search Algorithm) worst. When the file size is between 700 M and 800 M, XQuery is the most efficient, DKSA second and SLCA worst. When the file size is more than 850 M, DKSA is the most efficient, XQuery second and SLCA worst. With the growth of the file size, the advantage of DKSA is more obvious.

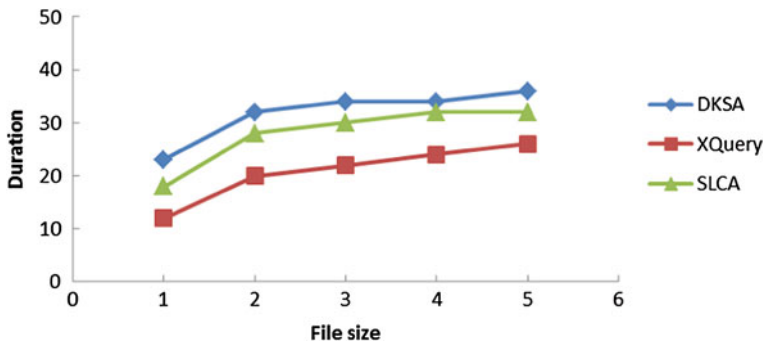


Fig. 141.6 The influence of keyword numbers

Figure 141.6 shows how the number of search keywords influences the query duration of each search algorithm when the file size is 600 M. It can be seen from the figure that there is a significant difference between one keyword based search and multi-keyword based search. And the influence of multi-keyword based search is limited. It can be considered as no effect essentially.

141.6 Conclusions

To do keyword search in XML databases is of great significance that can provide a friendlier user interface. Existing studies mainly focus on a single XML document while the XML database gets little attention. In this chapter, we combine the concept of keyword search algorithm-SLCA in XML documents and the characteristics of MapReduce to propose a distributed keyword search algorithm in XML databases, and implement it by open-source framework-Hadoop. The results show that the use of distributed parallel computing and the application of the appropriate index strategy make the system have a better overall performance. Furthermore, the larger the query scale is, the more obvious the effect is. Of course, there are some deficiencies in the algorithm, such as query efficiency, indexing strategy, etc. We will do in-depth studies on these deficiencies in the future research works.

References

1. Yanlei D, Altinel M, Franlin M et al (2003) Path sharing and predicate evaluation for high-performance XML filtering. *ACM Trans on Database Syst* 28, pp 467–516
2. Chen Y, Davidson SB, Zheng YF (2005) An efficient XPath query processor for XML streams. In: *Proceedings of ICDE*. Georgia: IEEE Press, 2005

3. Chan CY, Felber P, Garofalakis MN, Rastogi R (2002) Efficient filtering of XML documents with XPath expressions. In: Proceedings of ICDE, vol 27. IEEE Computer Press, California, pp 235–244
4. Li Y, Yu C, Jagadish HV (2004) Schema-Free XQuery. In: Nascimento MA, (ed) Proc. of 13th Int'l Conf. on Very Large Data Bases (VLDB). Toronto: Morgan Kaufmann Publishers, pp 72–83
5. Guo L, Shao F, Botev C (2003) XRANK: Ranked keyword search over XML documents. In: Proc. of the 22th ACM SIGMOD Conference. San Diego, California, USA: ACM Press, pp 16–27
6. Xu Y, Papakonstantinou Y (2005) Efficient keyword search for smallest LCAs in XML databases. In: Proc. of the 24th ACM SIGMOD Conference. Baltimore, Maryland, USA: ACM Press, pp 527–538
7. Bao Z, Ling TW, Chen B, and Lu J (2009) Effective XML keyword search with relevance oriented ranking. In: ICDE, 2009
8. Xiaofeng W, Xin Z, Min X, Xiaofeng M, Junfeng Z (2006) Keyword search on XML streams. *J Comput Res Develop* 23(03):484–489
9. Dean J, and Ghemawat S (2004) MapReduce: simplified data processing on large clusters. In: OSDI'04: Proceedings of the 6th conference on Symposium on Operating Systems Design & Implementation (Berkeley, CA, USA), USENIX Association. 10-10
10. Hadoop 0.21.0 api documentation. <http://hadoop.apache.org>. (2011)
11. Cloudera, Hadoop training and support. <http://www.cloudera.com>. (2011)
12. Al-Khalifa S, Jagadish HV, Patel JM, Wu Y, Koudas N, and Srivastava D (2002) Structural joins: A primitive for efficient XML query pattern matching. In: Proc. of ICDE Conference, pp 141–152
13. Bhalotia G, Nakhey C, Hulgeri A, Chakrabarti S, Sudarshan S (2002) Keyword searching and browsing in databases Using BANKS. In: Proceedings of ICDE. California: IEEE Computer Press, pp 431–440
14. Goldman R, Shivakumar N, Venkatasubramanian S, Garcia-Molina H (1998) Proximity search in databases. In: Proceedings of VLDB. New York: Morgan Kaufmann, pp 26–37
15. Hristidis V, Papakonstantinou Y, Balmin A (2003) Keyword proximity search on XML graphs. In: Proceedings of ICDE. Bangalore: IEEE Computer Press, pp 367–378

Chapter 142

A Combined Data Mining Method and Its Application in Water Quality Trends Analysis

Ji-Min Wang, Ding-Sheng Wan, Yue-Long Zhu and Shi-Jin Li

Abstract A combined method with moving average smoothing method, reverse test method and linear regression model is presented to analyze water quality trends in this chapter. The method is applied to the trends analysis for Yangtze-Taihu Water Diversion, the conclusions are verified by the practice and application of Wuxi Hydrology and Water Resources Monitoring Council, and the method is proved effective.

Keywords Moving average smoothing method · Reverse test method · Yangtze-Taihu water diversion · Water quality trends

142.1 Introduction

With the development of economy, water quality gets worse in recent years in some area. Water resources management departments have paid more and more attention to the water quality monitoring, and obtained huge volume of data. But how to identify implied, potentially useful knowledge, which can describe the rule of environmental evolution, from the huge volume of monitoring data becomes a problem. The water quality trends are often concerned about by the managers.

J.-M. Wang (✉) · D.-S. Wan · Y.-L. Zhu · S.-J. Li
College of Computer & Information, HoHai University, Nanjing, 210098, China
e-mail: wangjimin@hhu.edu.cn

J.-M. Wang · D.-S. Wan
National Engineering Research Center of Water Resources Efficient Utilization
Engineering Safety, Nanjing 210098, China

Many methods have been proposed to analyze hydrological time series trends, such as, linear regression, runs test method, reverse test method, seasonal Kendall, Mann–Kendall and R/S. In this chapter, we present a combined method which includes moving average smoothing method, reverse test method and linear regression to analyze water quality trends.

142.2 Related Work

Researchers have done a lot of work in the hydrological time series trends analysis. Hirsch et al. [1]. used the seasonal Kendall test for trend applicable to data sets, and then used the seasonal Kendall slope estimator for trend magnitude. At last, analyzed the relationship between constituent concentration and flow. Wang et al. [2] used R/S method analyzed fractal characteristics of hydrological time series, and pointed out that the variability of hydrological time series of the past and future was in the same or opposite characteristics. Xu et al. [3], Ichiyangl et al. [4] used Mann–Kendall method to analyze the rainfall trends. Naddafi et al. [5] used the Kolmogorov–Smirnov test to select the theoretical distribution which best fitted the water quality data, and then used simple regression, exponential model and logarithmic model to examine the concentration–time relationships of different stations. Yu et al. [6] combined R/S and Mann–Kendall to analyze future trend characteristics of hydrological time series. Kauffman et al. [7] used the seasonal Kendall test, scatter plots and box plots to detect statistical significance of water quality trends, and then compared water quality changes with watershed influences such as stream flow, seasonality, drainage basin, land use, and point source pollutants.

142.3 The Combined Method

In the combined method, moving average smoothing method is used to smooth the noise in water quality time series, reverse test method is used to test the existence of trends, and linear regression is used to analyze the trend direction and annual amplitude.

142.3.1 Moving Average Smoothing Method

Noise is made occasionally by equipment malfunction, input and data transmission errors during the water quality monitoring. Moving average smoothing method could smooth short time fluctuation and discover hidden trends. Moving average smoothing method includes: simple moving average (SMA), weighted moving

average (WMA) and exponential moving average (EMA). To compute moving average of m days, window with length m is moved from the trial to head of series $S = \{d_1, d_2, \dots, d_n\}$ with step 1, meanwhile computing the average of data in window. At the end, series with $n - m + 1$ averages obtained. WMA smoothing method sets different weight to data in the window. Equation 142.1 is used to compute WMA with window length 5.

$$WMA = d_{i-2}w_1 + d_{i-1}w_2 + d_iw_3 + d_{i+1}w_4 + d_{i+2}w_5 \quad (142.1)$$

where w_i is the weight. When $w_1 = w_2 = w_3 = w_4 = w_5 = 0.2$, Eq. 142.1 can be used to compute SMA with window length 5. Moving average smoothing method could smooth short time fluctuation and discover hidden trends.

142.3.2 Reverse Test Method

Reverse test method can tell whether there exists trend in time series, the steps are as follow [8]:

- 1: Create an average series Y from the original time series S . Split the original time series into M segments, and compute the average y_i of every segment, $i = 1, 2, \dots, M$, then series $Y = \{y_1, y_2, \dots, y_M\}$ includes M averages.
- 2: Compute the total number of reverse of series Y . To $y_i (i = 1, 2, \dots, M - 1)$, if there exists $y_j (j > i)$ and y_j greater than y_i , then y_i has a reverse. The total number of reverse of y_i is represented by A_i . Equation 142.2 gives the total number of reverse of Y represented by A .

$$A = \sum_{i=1}^{M-1} A_i \quad (142.2)$$

- 3: Make statistical test by A . The average $E(A)$ and variance $D(A)$ of A are as follow:

$$E(A) = \frac{1}{4}M(M - 1) \quad (142.3)$$

$$D(A) = \frac{M(2M^2 + 3M - 5)}{72} \quad (142.4)$$

where M is the number of segments. For statistic $z = \frac{A + \frac{1}{2} - E(A)}{\sqrt{D(A)}} \sim N(0, 1)$ called reverse coefficient, if $|z| < 1.96$ under 5% of significance level, then obvious trend exists in S , otherwise does not.

142.3.3 Linear Regression Model

Regression analysis is a statistical method used to identify quantitative relationships of two or more interdependent variables. One-dimensional linear regression function between y and x is as follow:

$$\hat{y} = b_0 + b_1x \quad (142.5)$$

where b_0, b_1 should make the quadratic sum of deviation between observed value y and estimated regression value \hat{y} smallest.

142.3.4 The Combination of Three Methods

Above three methods are combined to analyze long-term water quality trends and uniform seasonal water quality trends.

142.3.4.1 Long-Term Water Quality Trends Analysis

For water indicator W , the methods of Sects. 142.3.1–142.3.3 are combined to analyze the long-term water quality trends. The trends include increase, balance and decrease three patterns, The steps are as follow:

- 1: Smooth the noise of water quality time series with SMA.
- 2: Compute the reverse coefficient Z of water quality time series.
- 3: If $|Z|$ is less than 1.96, then time series keep balance, and the analysis terminates, otherwise turns to 4.
- 4: Build one-dimensional linear regression model between water quality concentration and time, then analyze the trends direction (increase or decrease) and annual amplitude by the sign and absolute value of b_1 respectively. If b_1 is greater than 0, then the trend is increase, if b_1 is less than 0, then trend is decrease.

142.3.4.2 Uniform Seasonal Water Quality Trends Analysis

Uniform seasonal water quality trends are the uniform trends of the water quality indicator when the seasons change in a year. The steps of uniform seasonal water quality trends analysis are as follow:

- 1: Symbolize the water quality seasonal average time series. For water quality time series $T = \{x_1, x_2, \dots, x_n\}$, the difference series is $T' = \{x_2 - x_1, x_3 - x_2, \dots, x_n - x_{n-1}\}$, $x_{i+1} - x_i$ may be greater than 0, less than 0 or equal 0

which means increase, decrease or balance respectively. Replace $x_i + 1 - x_i$ with U, B or D which represent increase, balance or decrease respectively, then symbol series $S = \{m_1, m_2, \dots, m_{n-1}\}$, $m_i \in \{U, B, D\}$ obtained.

- 2: Split S by year into series $S' = \{s_1, s_2, \dots, s_k\}$, where s_i is the seasonal trend subsequence of one year.
- 3: Compute the probability of the trend of subsequence in S' . The trend whose probability is greater than 60% is the uniform trend when the seasons change, otherwise, there is no obvious uniform trend when the seasons change.

142.4 Experiments and Data Analysis

Yangtze-Taihu Water Diversion is an important project [9] to improve the water environment in Taihu Basin by water conservancy. The Yangtze River water is extracted at Changshu hydro-junction, and flows into Taihu Lake through Wangyu River. The water quality of Yangtze River directly affects the water quality of Taihu. In this section, the long-term and seasonal water quality trends of water diversion are analyzed.

142.4.1 Experimental Data

In experiment, data monitored during the period from 2002 to 2008 are used. Four water quality indicators COD_{Mn} , NH_3-N , TN and TP are included. After smooth data with SMA, then calculate monthly and seasonal average.

142.4.2 Long-Term Water Quality Trends Analysis of Water Diversion

Yangtze-Taihu Water Diversion administrators are usually concerned about the suitability of the water quality of some months for water diversion. As follows, the four indicators monthly long-term trends are analyzed. Reverse coefficients of monthly average time series of four indicators from January to December are shown in Table 142.1.

The bold numbers mean trends exist under 5% of significance level. So trends exist in March, July, August, December for TN, December for TP, February for NH_3-N . Linear regression models are established for the water quality time series with trends, and regression coefficients b_1 are shown in Table 142.2. It can be seen from the Table 142.2, increase trends exist in time series for all indicators and

Table 142.1 Reverse coefficients of monthly average time series of four indicators from January to December

Month	TP	TN	NH ₃ -N	COD _{Mn}
January	0.734	-0.244	-0.734	-1.224
February	1.698	1.698	2.377	0.339
March	1.224	2.204	1.224	-0.244
April	0.375	0.751	-0.375	-1.127
May	-0.244	-0.734	-0.244	-0.244
June	0.244	-0.244	-0.734	-0.244
July	1.019	2.377	1.698	0.339
August	-1.127	2.254	-0.75	-1.127
September	1.878	0.375	-0.375	1.127
October	1.127	0.0	-0.375	1.127
November	-0.339	1.698	1.019	1.698
December	2.070	2.377	0.339	1.698

Table 142.2 Regression coefficients b_1

Indicator	Month	b_1	Annual amplitude(%)
TN	March	0.023	2.3
	July	0.058	5.8
	August	0.028	2.8
	December	0.062	6.2
TP	December	0.048	4.8
MH3-N	February	0.036	3.6

months, that means the water quality indicators get worse, but annual amplitude are not high.

Indicators concentration curves with trends are shown in Fig. 142.1, the same trends can be found in the figure.

142.4.3 Seasonal Water Quality Trends Analysis of Water Diversion

As a example, TN seasonal trends is analyzed. First, TN seasonal time series is symbolized to symbol series, and then symbol series is divided into subsequences by the year, at the end, analyze the subsequences and get the probability shown in Table 142.3.

TN seasonal water quality trends is DDDU, that means TN decreases from the Winter at beginning of the year to the Autumn and increases in the Winter of next year. Sometimes, there is no trend with significantly higher probability than the

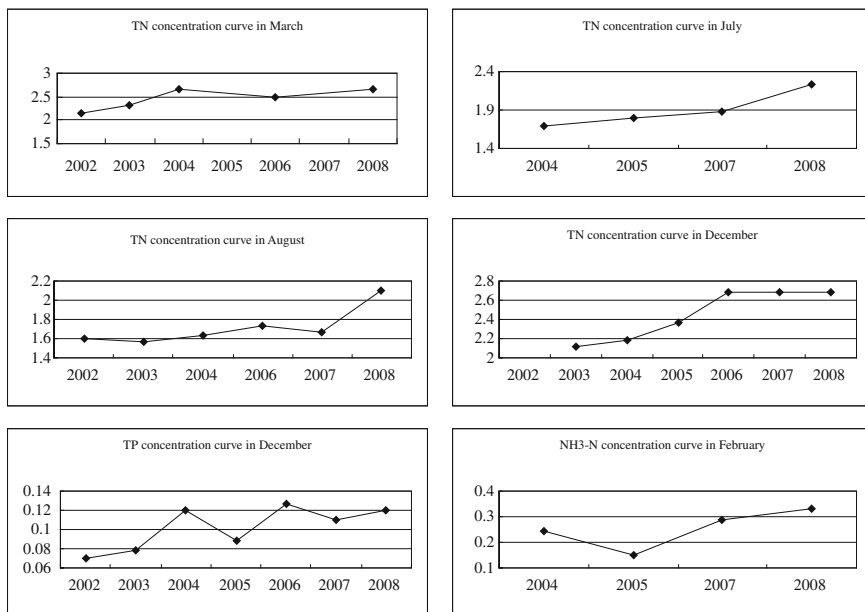


Fig. 142.1 Indicator concentration curves with trends

other trends, such as TP. Below, TP seasonal trends will be analyzed. the TP trends probability shown in Table 142.4.

In Table 142.4, the probability of D and U is very similar, all close to 50%, when they are the cases of the turn of Winter and Spring, Spring and Summer, so they are all the main trends. So TP seasonal water quality trends are (DIU)(DIU)UU, which means when Winter turns to Spring and Spring turns to

Table 142.3 TN seasonal water quality trends

Season alternation	Trend	Probability	Trend with the highest probability
Winter → Spring	D	0.71	D
	B	0	
	U	0.29	
Spring → Summer	D	1	D
	B	0	
	U	0	
Summer → Autumn	D	0.67	D
	B	0	
	U	0.33	
Autumn → Winter	D	0	U
	B	0	
	U	1	

Table 142.4 TP seasonal water quality trends

Seasonal alternation	Trend	Probability	Trend with the highest probability
Winter → Spring	D	0.57	DIU
	B	0	
	U	0.43	
Spring → Summer	D	0.57	DIU
	B	0	
	U	0.43	
Summer → Autumn	D	0.33	U
	B	0	
	U	0.67	
Autumn → Winter	D	0.33	U
	B	0	
	U	0.67	

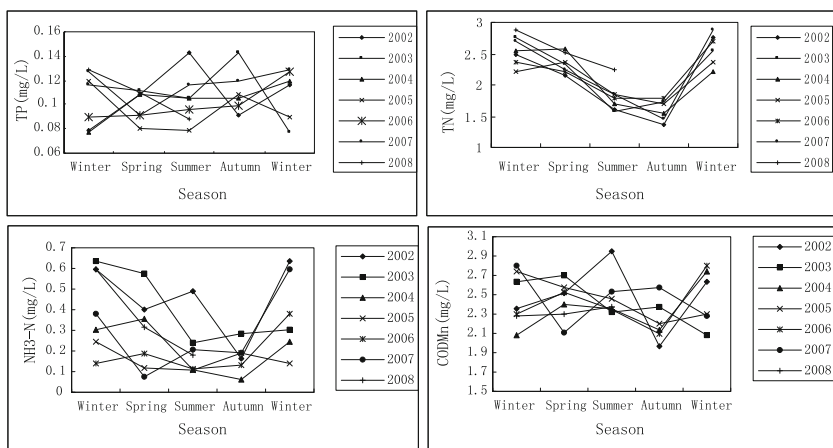


Fig. 142.2 Seasonal average concentration curves of years

Summer the uniform trend is not obvious, but when Summer turns to Autumn and Autumn turns to Winter, the trends are all increasing.

The seasonal trends of NH₃-N and COD_{Mn} are DDDU and UDDU respectively. Seasonal average concentration curves of years are shown in Fig. 142.2, the analysis could be verified from the figure.

142.5 Conclusions

This chapter presents a combined method with moving average smoothing method, reverse test method and linear regression model to analyze water quality trends, and the method is applied to the water quality trends analysis of Yangtze-Taihu

Water Diversion. Conclusions are as follow: water quality of Water Diversion generally does not significantly increase or decrease, but some indicators in some months demonstrate trends which are obviously increasing. In a year, TN and $\text{NH}_3\text{-N}$ decrease gradually from the Winter to Autumn, and increase in next Winter. The trends of TP are not obvious when Winter turns to Spring, and Spring turns to Summer, but TP increases from Summer to Autumn and the next Winter. COD_{Mn} increases from Winter to Spring, then decreases from spring to Autumn, finally, increases in next Winter. The conclusions have been validated by practice and application of management departments of water resources of Taihu.

Acknowledgments This research was supported by the Fundamental Research Funds for the Central Universities (No. 2009B22014), the National Natural Science Foundation of China (No. 51079040), 948 Project of Ministry of Water Resources (No. 201016).

References

1. Hirsch RM, Slack JR, Smith RA (1982) Techniques of trend analysis for monthly water quality data [J]. *Water Resour Res* 18(1):107–121
2. Wang XL, Hu BQ, J Xia. (2002) R/S analysis method of trend and aberrance point on hydrological time series [J]. *J Wuhan Univ Hydraul Electr Eng* 35(2):10–12
3. Xu ZX, Zhang N (2006) Long-term trend of precipitation in the Yellow River Basin during the past 50 years [J]. *Geogr Res* 25(1):27–33
4. Ichiyangl K, Yamanaka MD, Muraji Y et al (2007) Precipitation in N.Cpell between 1987 and 1996 [J]. *Int J Climatol* 15(2):245–256
5. Naddafi K, Honari H, Ahmadi M (2007) Water quality trend analysis for the Karoon River in Iran [J]. *Environ Monit Assess* 134:305–312
6. Yu YS, Chen XW (2008) Analysis of future trend characteristics of hydrological time series based on R/S and Mann–Kendall methods [J]. *J Water Res Water Eng* 19(3):41–44
7. Kauffman GJ, Belden AC (2010) Water quality trends (1970 to 2005) along Delaware streams in the Delaware and Chesapeake Bay Watersheds, USA [J]. *Water Air Soil Pollut* 208:345–375
8. Xiao ZH, Guo YM (2009) Time series analysis and SAS application [M]. Wuhan University Press, Wuhan, pp 24–28
9. Wu HY (2008) Research of key technology for Yangtze-Taihu water diversion [J]. *China Water Resour* 1:1–3

Chapter 143

Model of Enterprise Innovation Based on Data Warehousing

Jianwei Guo, Bingru Yang, Guangyuan Li and Nan Ma

Abstract With the increasing complexity of economic globalization, the enterprises of China would face new challenge and new pressure. Various problems have sprung up such as environmental pollution, low rewards, labors' cost rising, and so on. The innovation for enterprise has become both inevitable and necessary. In this paper a new method of building a data warehouse has been adopted and a data warehouse has been designed to evaluate innovativeness of enterprises for the first time.

Keywords Innovation score · Innovation field · Innovation indicator · Data warehouse

143.1 Introduction

With the increasing complexity of globalization, there has been a rapid growing importance and discussion for more research on how to improve enterprise innovation ability to gain competitive business advantages. Furthermore, Chinese government is keeping a close watch on how to improve ability for enterprise innovation, and to enhance competitiveness for Chinese enterprise around world. Therefore, evaluating method for enterprise innovation is very important. According to the evaluation, managers of enterprise can learn the effect of resource utilization and operation. Additionally, government officers can evaluate the

J. Guo (✉) · B. Yang · G. Li · N. Ma
School of Computer and Communication Engineering, University of
Science and Technology Beijing, 100083 Beijing, China
e-mail: darkandwhite2002@163.com

innovation degree for an enterprise or a kind of specific industry. For the above reasons, enterprise innovation score should be evaluated by innovation experts using formulas and measures based on the data. Experts would collect a number of data from various enterprises as much as possible. And they can design formulas and algorithms to measure the enterprise innovation scores. Then, all the managements or officers may see the innovation scores at those enterprises. In this paper, we do not focus on how to measure the innovation score. We put our attention on how to organize a mass of information of data, and how to show them to users. What using data warehousing integrate data innovation-oriented, helping users use their data to make strategic decisions is the particular description in this paper.

In summation, the enterprise innovation based on data warehousing works for enterprise and government, especially for government to evaluate enterprise innovative performance.

143.2 Data Warehouse

Data warehousing is a collection of decision support technologies [1–3], aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions. A data warehouse is a “subject-oriented, integrated, time varying, non-volatile collection of data that is used primarily in organizational decision making” [4–6].

Typically, the data warehouse is maintained separately from the organization’s operational databases. There are many reasons for doing this. The data warehouse supports on-line analytical processing (OLAP) [6, 7], the functional and performance requirements of which are quite different from those of the on-line transaction processing (OLTP) applications traditionally supported by the operational databases.

Data warehouses, in contrast, are targeted for decision support. Historical, summarized and consolidated data is more important than detailed, individual records. Since data warehouses contain consolidated data, perhaps from several operational databases, over potentially long periods of time, they tend to be orders of magnitude larger than operational databases; enterprise data warehouses are projected to be hundreds of gigabytes to terabytes in size. The workloads are query intensive with mostly ad hoc, complex queries that can access millions of records and perform a lot of scans, joins, and aggregates. Query throughput and response times are more important than transaction throughput. Consequently, three new basic concepts have been proposed.

Facts are a set of linked data (metrics and context data). The data can be either measures permitting to determine the activity (business transaction) or events used for the analysis.

Dimensions are the fact context. There are analysis axis related to the measures to study into the facts.

Table 143.1 Innovation fields for enterprise

Innovation field
F1. Innovation strategy
F2. Core competencies and knowledge
F3. Innovation process
F4. Culture of innovation
F5. Organizational structure and network of partners
F6. Market development
F7. Key technology
F8. Project management
F9. Production and service

Metrics called measures correspond to numerical data of facts allowing to evaluate performances.

143.3 Measurement of Innovation

The evaluation for enterprise innovation degree was issued in techniques in the past. However improving technique is not enough for an enterprise to compete successfully. Because improving competitive ability for enterprise need fully support, like strategic deployment, service ability, customer relationship, and so on. Here, we evaluate enterprise innovation in a means which is designed by professor Zhang in Tongji University. In his design, the enterprise innovation include nine innovation fields which are showed in Table 143.1, and any field include some innovation indicators. The relation between fields and indicators will explain in later chapters. Now, it's enough for us to know that the score of innovation fields is figured from indicator scores, and total score of enterprise innovation is figured by nine innovation fields.

143.4 Data Acquisition

Data is basic factor in data warehouse. Innovation score is basic factor in enterprise innovation data warehouse. Of course, the innovation score is acquired by some evaluations which innovation experts figure out. Let us see a procedure about how to figure out score in key technology field by indicators of key technology. First, innovation experts should make a table including five indicators. They are Number of third party patents, Number of owed patents and key techniques, Research degree for trends of production, Research and popularization of innovative approaches, and R&D proportion from sales. And enterprise executives fill out specific information about these five indicators. Then, experts calculate a total score from five indicators. The calculation is a specific weighted algorithm, and it

Table 143.2 Key technology field scores in company xxx and yyy using a percentage grading system

Key technology field	Company xxx	Company yyy
i1. Number of 3rd party patents	60	70
i2. Number of owed patents and key techniques	70	50
i3. Research degree for trends of production	80	80
i4. Research and popularization of innovative approaches	50	60
i5. R&D proportion from sales	100	90

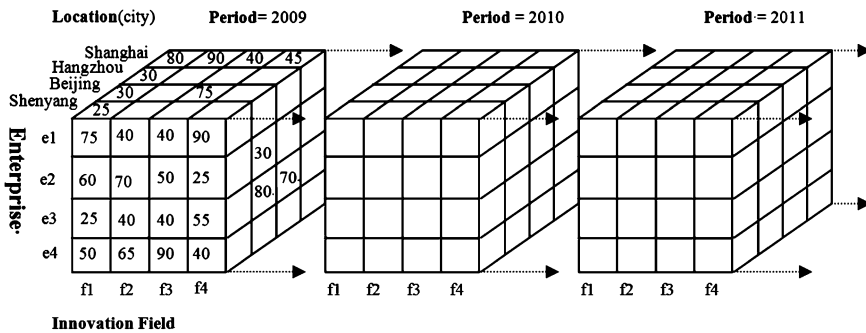


Fig. 143.1 4×4 4D cube of innovation degree score. The measure displayed is innovation scores

is not focus in this paper. Finally, the total score is measurement for field of key technology. In the same way, experts can get scores in any other field.

For example, every score of *key technology* in of company xxx and yyy are displayed in Table 143.2.

143.5 Multidimensional Cube Modeling

In order to evaluate degree of enterprise innovation, the data warehousing model must be designed by innovation-oriented, then to research related attributes around it [8, 9]. Each of the attribute depends on a set of dimensions, which provide the context for the measure. We propose that dimensions associated with enterprise innovation degree can be the innovation field, enterprise location, enterprise type, and the period when the strategy was made Fig. 143.1.

The Innovation field dimension may consist of nine field attributes: innovation strategy, Core competencies and knowledge, Innovation process, and other fields aforementioned. The total score of nine fields represent the innovation ability

Enterprise location dimension consist of cities, like Shanghai, Beijing, Shenyang, etc.

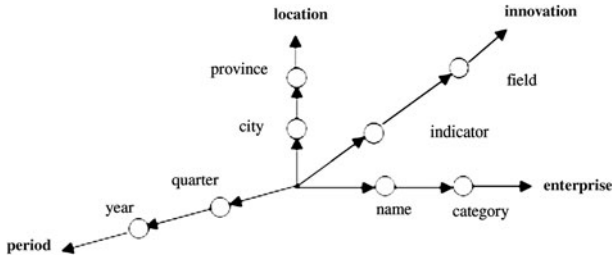


Fig. 143.2 Hierarchy for the dimension in form of star network model

Enterprise type are made of companies of food industry, companies of IT industry, manufacturing industry, etc.

Period represent enterprises make decision about innovation strategies in the different periods.

143.5.1 Hierarchies Modeling

A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts. Let us take location for an example. City can be mapped to the province. For example, Shenyang can be mapped to the Liaoning province. In while of course, higher-level can be mapped to the low-level concepts. Nine enterprise innovation fields can decompose into 77 indicators. The relationship between Innovation field and the indicators attribute are defined through below hierarchical relationship. Due to limited space, this paper only takes Strategy-field and key technology as example Fig. 143.2.

Strategy-field: (1) A clear long-term innovation strategic objectives. (2) Cooperative innovation strategy and overall strategy. (3) Correspondence between innovation method and innovation goal. (4) Correspondence between innovation model and internal environment. (5) Proportion of R&D sales revenue in project fund. (6) Proportion of research in R&D. (7) Proportion of research in pre-research project i1. (8) Proportion of project without implemented.

Key Technology: (1) Number of 3rd party patents. (2) Number of owed patents and key techniques. (3) Research degree for trends of production. (4) Research and popularization of innovative approaches. (5) R&D proportion from sales.

143.5.2 Multidimensional Database Modeling

Our enterprise innovation data warehouse uses a star schema to represent the multidimensional data model [10, 11]. The database consists of a single fact table and a single table for each dimension. Each tuple in the fact table consists of a

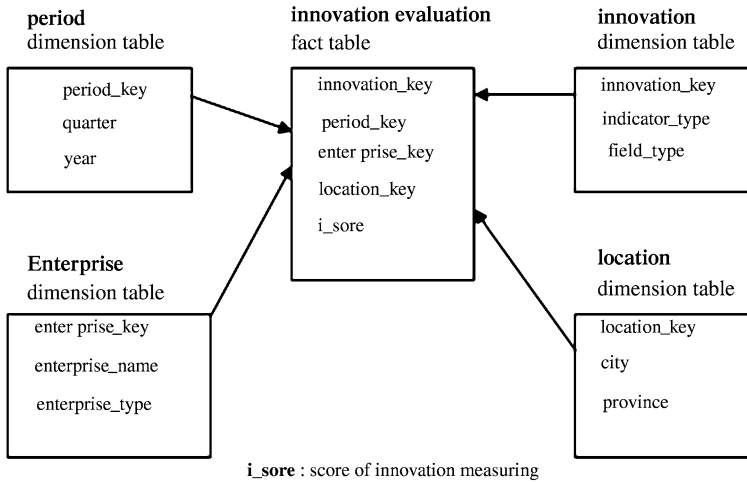


Fig. 143.3 Data warehousing star schema for innovation evaluation

pointer (foreign key—often uses a generated key for efficiency) to each of the dimensions that provide its multidimensional coordinates, and stores the numeric measures for those coordinates. Each dimension table consists of columns that correspond to attributes of the dimension.

Innovation scores are considered along four dimensions, namely, innovation, period, enterprise, and location. The schema contains a central fact table for innovation evaluation that contains keys to each of the four dimensions. Figure 143.3 shows data warehousing star schema for innovation evaluation.

The star schema of innovation data warehousing is defined in DMQL [12] as follows:(The score of innovation measuring is not simple addition and subtraction, so function sum is replaced by `i_sum` which is defined by innovation evaluation algorithm.):

```

define cube innovation_evaluation_star [period, innova-
tion, enterprise, location]:
i_score = i_sum(innovation_scores),
define dimension period as (period_key, quarter, year)
define dimension innovation as (innovation_key, indica-
tor_type, field_type)
define dimension enterprise as (enterprise_key, enter-
prise_name, enterprise_type)
define dimension location as (location_key, city, province)
  
```

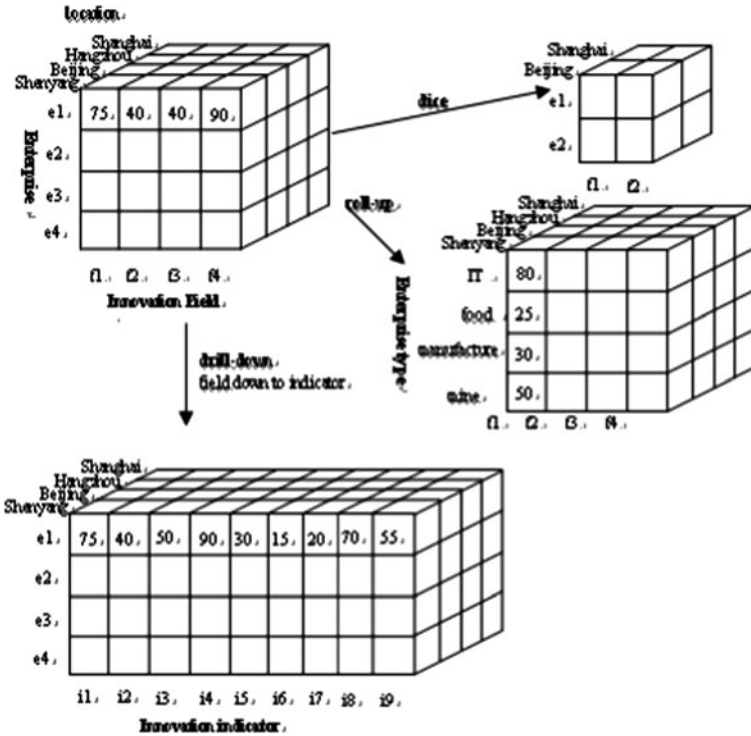


Fig. 143.4 OLAP operations on multidimensional data for enterprise innovation

143.6 OLAP Operations in the Innovation Multidimensional Data Model

Let’s look at some typical OLAP operations for enterprise innovation multidimensional data.

The roll-up operation (also called the drill-up operation by some vendors) performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction. Figure 143.4 shows this hierarchy was defined as the total order “specific enterprise < a kind of industry type.” The roll-up operation shown aggregates the data by ascending the enterprise hierarchy from the level of specific type to the level of all type. In other words, rather than grouping the data by specific an enterprise name, the resulting cube groups the data by sum of all enterprise.

The drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data. drill-down can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions. Figure 143.4 shows the result of a drill-down operation performed on the central cube by stepping down a concept hierarchy for innovation defined as “indicator < field.”

Drill-down occurs by descending the innovation hierarchy from the level of field to the more detailed level of indicator. The resulting data cube details the total score of innovation evaluation per indicator rather than summarizing them by innovation field.

Slice and dice: The slice operation performs a selection on one dimension of the given cube, resulting in a subcube. The dice operation defines a subcube by performing a selection on two or more dimensions. Figure 143.4 shows a dice operation on the central cube based on the following selection criteria that involve three dimensions: (location = “Beijing” or “Shanghai”) and (enterprise = “e1” or “e2”) and (field = “f1” or “f2”).

Using OLAP operations users can get innovation score for all aspect from year to quarter, from province to city, from a kind of industry to a specific company. No matter who, enterprise executive or government officer can get comprehensive information for innovation evaluation they need.

143.7 Conclusion

In this chapter, we design a data warehousing model for enterprise innovation. Making innovation field, enterprise location, period and enterprise type into sets of dimensions, evaluate enterprise innovation score. Furthermore, clarify an important notion about enterprise innovation in nine innovation fields, and relating to innovation indicators. Finally, we hope more and more Chinese enterprises should pay more attention to improve individual innovation degree, and to enhance competitiveness for Chinese enterprise around world.

Acknowledgments This work originated from pioneering project of the National Department of Technology (Grant No. 2010IM020900).

References

1. Widom J (1995) Research problems in data warehousing. In: Proceedings of the 4th international conference information and knowledge management, pp 25–30, Baltimore, MD, Nov 1995
2. Kimball R, Ross M (2002) The data warehouse toolkit: the complete guide to dimensional modeling, 2nd edn. Wiley, New York
3. Shoshani A (1997) OLAP and statistical databases: similarities and differences. In: Proceedings of the 16th ACM symposium. Principles of database systems, pp 185–196, Tucson, May 1997
4. Imhoff C, Galemno N, Geiger JG (2003) Mastering data warehouse design: relational and dimensional techniques. Wiley, New York
5. Gupta A, Mumick IS (1999) Materialized views: techniques, implementations, and applications. MIT Press, Cambridge

6. Chaudhuri S, Dayal U (1997) An overview of data warehousing and OLAP technology. *SIGMOD Rec* 26:65–74
7. Shoshani A (1997) OLAP and statistical databases: similarities and differences. In: *Proceedings of the 16th ACM symposium, principles of database systems*, pp 185–196, Tucson, May 1997
8. Sarawagi S, Stonebraker M (1994) Efficient organization of large multidimensional arrays. In: *Proceedings of the 1994 international conference data engineering (ICDE'94)*, pp 328–336, Houston, Feb 1994
9. Gray J, Chaudhuri S, Bosworth A, Layman A, Reichart D, Venkatrao M, Pellow F, Pirahesh H (1997) Data cube: a relational aggregation operator generalizing group-by, cross-tab and sub-totals. *Data Min Knowl Disc* 1:29–54
10. Han J, Fu Y, Wang W, Chiang J, Gong W, Koperski K, Li D, Lu Y, Rajan A, Stefanovic N, Xia B, Za OR (1996) DB miner: a system for mining knowledge in large relational databases. In: *Proceedings of the 1996 international conference data mining and knowledge discovery (KDD'96)*, pp 250–255, Portland, Aug 1996
11. Beyer K, Ramakrishnan R (1999) Bottom-up computation of sparse and iceberg cubes. In: *Proceedings of the 1999 ACM-SIGMOD international conference management of data (SIGMOD'99)*, pp 359–370, Philadelphia, June 1999
12. Han J, Fu Y, Wang W, Chiang J, Gong W, Koperski K, Li D, Lu Y, Rajan A, Stefanovic N, Xia B, Za OR (1996) DB miner: a system for mining knowledge in large relational databases. In: *Proceedings of the 1996 international conference data mining and knowledge discovery (KDD'96)*, pp 250–255, Portland, Aug 1996

Chapter 144

Classifying Imbalanced Dataset Using Local Classifier Fusion

Tong Liu, Weijian Ni and Yongquan Liang

Abstract Imbalanced class distribution is a pervasive problem in real applications of classification. Traditional classification approaches tend to be poorly performed due to the important class overwhelmed by the majority but less important ones. In this chapter, we proposed a novel approach to handle the class-imbalance problems where local-instances distributions are well examined to make up the global classification model. In this work we perform a broad experimental evaluation on real world datasets show that our approach outperforms traditional classification approaches and several existing imbalanced classification approaches.

Keywords Classification · Class imbalance · Local distribution · Model fusion

144.1 Introduction

In wide range of real applications of classification, classes are often distributed highly imbalanced, i.e., the overwhelming portion of samples comes from only some classes. For example, in information retrieval, the relevant documents retrieved by user's query are much less than the irrelevant ones [1]. In automatic

T. Liu (✉) · W. Ni · Y. Liang
College of Information Science and Engineering, Shandong University
of Science and Technology, 266510 Qingdao, Shandong, People's Republic of China
e-mail: lovebonat@gmail.com

W. Ni
e-mail: niweijian@gmail.com

Y. Liang
e-mail: lyq@sdust.edu.cn

acronym extraction, the truthful pairs of acronym and its explanation are very sparse in the whole candidate pairs [1]. Since the learning target of most of the traditional classification algorithms is a prediction of accuracy on training set, the resultant classification model can be easily biased toward the majority class [2], and thus tend to perform poorly on the minority class that is often representative of interesting patterns, and decrease the generalization of the learning models.

Therefore, classification over class-imbalance distribution has been one of the most important issues in machine learning. During the past few years, a number of approaches to imbalance classification have been proposed. Generally speaking, the main idea of these approaches is to re-balance the original distribution such that the majority and minority classes can be equally treated. For example, a balanced data set for learning can be generated by either oversampling upon the minority class or undersampling upon the majority class. However, the oversampling and undersampling approaches often suffer from overfitting and losing useful information, respectively. Another type of re-balancing strategy is cost sensitive learning where the minority and majority classes are associated with comparable costs during learning.

In this chapter, we propose a novel imbalanced classification approach on the basis of local classifier fusion. Specifically, we first cluster the imbalanced dataset into several subsets, each of which is used for training the local classifiers. During the inference stage of our approach, the global prediction model is constructed by the combination of these local classifiers. There are mainly two key features of our approach. First, a new clustering algorithm is proposed to partition the imbalanced dataset into several “balanced” parts while the natural local distribution of the dataset is captured. Second, the local models are combined in a supervised manner, in which a variety of features of each model are taken into consideration.

144.2 Classification on Imbalanced Distribution

144.2.1 Algorithm

Recently mining imbalanced datasets has been the focus of much research, which can be divided into the data level and the algorithmic level.

The data level, mainly refer to re-sampling, including oversampling upon the minority class, undersampling upon the majority class and both of them; they can alleviate the degree of imbalanced distribution between the two classes to some extent. One of the most famous oversampling methods is SMOTE [3]. It beats the random oversampling with its informed properties, by adding the “new” minority instances and without suffering from overfitting. However, SMOTE blindly generalizes synthetic minority instances without regarding the majority instances and may cause overgeneralization. Several undersampling approaches are proposed to make the selected samples more representative. Wilson’s Editing [4] is one of the

Table 144.1 Confusion matrix

	Predicted as positive	Predicted as negative
Actually positive	True positives (TP)	False negatives (FN)
Actually negative	False positives (FP)	True negatives (TN)

undersampling techniques that attempts to remove only majority class instances that are likely to contain noise which is determined by the k -NN algorithm with $k = 3$. The undersampling approach uses several distances between the majority and the minority, including the nearest, the farthest, the average nearest, and the average farthest distances, to sample the majority instances. However, potentially useful information contained in some ignored samples is neglected. Current literatures report that the performances of oversampling approaches are worse than that of undersampling approaches [2]. With regard to the unfavorable results either oversampling or undersampling alone in the imbalanced classification, recently some researchers proposed integration both of them [5] to re-balance the original distribution.

The algorithmic level, previous research mainly focused on modify properly the existing classify learning algorithms according to the defects of dealing with imbalanced classification [6, 7]. The main view of cost sensitive learning is assigning different cost factors to false negatives and false positives [6]. The main learning objective of cost sensitive Boosting algorithm is the cost of misclassifying a positive instance out-weights that of misclassifying a negative one. However, the success of the unequal weights approach may rely on parameter tuning processes, and misclassification costs are often unknown.

144.2.2 Evaluation Measures

As for classification on imbalanced dataset, the traditional evaluation measures such as accuracy is far below satisfactory when the imbalance ratio reaches 99:1. Therefore, new measures are designed to evaluate the classification results of imbalanced dataset. Generally speaking, these measures are calculated using the confusion matrix shown in Table 144.1. Specifically, sensitivity (recall) is the probability of the positive class of all instances, and the specificity is the probability of negative class of all the instances, and they are usually trade-off. Hence, F-Measure and G-Mean will often be used due to they all combine sensitivity and specificity.

$$G - Mean = \sqrt{sensitivity * specificity} \quad (144.1)$$

$$F - Measure = \frac{2 * precision * recall}{precision + recall} \quad (144.2)$$

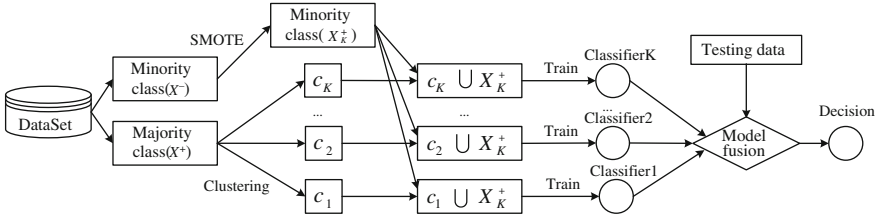


Fig. 144.1 Overview of the approach

$$Sensitivity(Recall) = \frac{TP}{(TP + FN)}, Specificity = \frac{TN}{(TN + FP)}, Precision = \frac{TP}{(TP + FP)} \tag{144.3}$$

144.3 Local Classifier Fusions for Imbalanced Data

Inspired by the description of feature vector, a new description of an instance is proposed. That is, make the descriptions of vectors convert feature space to distance space. Because of the variations of vector descriptions, we introduce clustering algorithm into dividing the majority class. First oversampling the minority class with SMOTE to some extent. The number of clusters depends on the number of the whole minority examples after oversampling by SMOTE. The minority class after SMOTE are used to combine with new clustering results to train local classifiers. Then cluster majority class into K clusters based on the distance space descriptions, and train several balanced parts. However, individual classifier is weak, so it is necessary to combine those local classifiers into one classifier. We integrate these local classifiers in a supervised manner. The overview of our approach is graphically illustrated in Fig 144.1.

144.3.1 Inspiration from Feature Space

In this section, we will make great use of multi-view strategy to majority class. Let $\vec{x}_i^- (i = 1, \dots, m)$ and $\vec{x}_j^+ (j = 1, \dots, n)$ denotes the feature vectors of the instances of the majority and minority class, respectively, both \vec{x}_i^- and \vec{x}_j^+ are d dimension vector in R^d , where d represent the number of features, and m and n denote the sizes of these two groups respectively. Based on the feature space, introduce a variation in distance space. Basically, the original descriptions of vectors in feature space are described as:

$$\vec{x}_i^{-a} = (x_{i1}, x_{i2}, \dots, x_{id})(i = 1, \dots, m) \quad (144.4)$$

$$\vec{x}_j^+ = (x_{j1}, x_{j2}, \dots, x_{jd})(j = 1, \dots, n) \quad (144.5)$$

On the other hand, suppose the distance between the i th majority instance and the j th minority instance is d_{ij} . Then convert feature space into distance space of the same vector, and the whole majority instances can be described as \vec{x}_i^{-b} .

$$d_{ij} = \sqrt{\sum_{k=1}^d (x_{jk}^+ - x_{ik}^-)^2} \quad (144.6)$$

$$\vec{x}_i^{-b} = (d_{i1}, d_{i2}, \dots, d_{in})(i = 1, \dots, m) \quad (144.7)$$

The pseudo-code of our approach is shown in algorithm 1. We apply the clustering in algorithm 2 to the new representation vectors in the distance space of the majority class. By now, $\{f_i\}_{i=1}^K$ are K decision functions of the local-classifiers $\{T_i\}_{i=1}^K$, respectively. Then, we can represent \vec{x} by its function values $(f_1(\vec{x}), \dots, f_k(\vec{x}))$.

144.3.2 Local Model Fusion

As mentioned above, we get K local classifiers through algorithm 1. However, the final output of a test instance must obtained by model fusion. Recently, ensemble technique have been applied in a broad spectrum of scenarios in order to improve the performance of weak classifiers [2, 7, 8]. Here, we take the distances between a test sample and each hyper-plane into consideration. Of course, the shortest distance of that hyper-plan is the test sample belongs to. In this section, the key in the process of model fusion is how to calculate the distance between a test sample and a hyper-plane. By now we have got K different models from above process, the classification abilities of those models are different for those test samples, so we consider this kind of ability as classification quality of the hyper-plane function f_i to a given test vector \vec{x} , denoted as q_i . As usual, given a test instance \vec{x} , we can get its prediction function value though the following equation (144.8):

$$f(\vec{x}) = \sum_{i=1}^K \omega_i \cdot f_i(\vec{x}) \quad (144.8)$$

Based on the difference of prediction ability of each hyper-plane for the same instance, we induce another transformation of equation (144.8) into (144.9).

$$f(\vec{x}) = \sum_{i=1}^K q_i \cdot d(f_i, \vec{x}) \cdot f_i(\vec{x}) \quad (144.9)$$

Algorithm 1. Local Classifier Training

Input: Majority training set $X^- = \{\vec{x}_1^-, \dots, \vec{x}_m^-\}$; The number of k nearest neighbors;
 Minority training set $X^+ = \{\vec{x}_1^+, \dots, \vec{x}_n^+\}$
Output: local-classifiers $\{f_i\}_{i=1}^k$

1. $X_K^+ = \text{SMOTE}(X^+, k)$
2. $K = |X_K^+|$
3. $\{c_1, c_2, \dots, c_K\} = \text{clusteringnbpsp}(X^-, K)$
4. For $i = 1$ to K do
5. $T_i = c_i \cup x_k^+$
6. Build a component classifier f_i with T_i
7. End for

Algorithm 2. Clustering algorithm

Input: Negative instance Set in two views $S^- = \{(\vec{x}_1^a, \vec{x}_1^b), (\vec{x}_2^a, \vec{x}_2^b), \dots, (\vec{x}_n^a, \vec{x}_n^b)\}$,
 Distance metric d^a and d^b , Clusters number k
Output: Clustering Result $C = \{C_1, C_2, \dots, C_k\}$

1. Initialize $C_i = (\vec{x}_i^a, \vec{x}_i^b)$, $i = 1, \dots, n$
2. For $t = 1$ to $(n - k)$
3. $(C_r, C_s) = \arg \min_{i,j=1,\dots,n} \{d^a(C_i, C_j), d^b(C_i, C_j)\}$
4. Merge C_r and C_s
5. End For

Surely, the lager the value q_i , the more important the f_i is with regard to the test vector \vec{x} . where $d(f_i, \vec{x})$ is the distance between the test sample \vec{x} and the hyper-plane f_i . It can be solved by KNN. Suppose n_{i1}, \dots, n_{ik} denote the number of the i th test sample's k neighbor samples, belonging to different clusters, respectively.

$$d(f_i, \vec{x}) = \min\{n_{i1}/k, \dots, n_{ik}/k\} \tag{144.10}$$

From the above algorithm 1, we get a series of decision functions $\{f_i\}_{i=1}^k$. The test \vec{x} instance is also a d -dimensional vector, i.e. $\vec{x} = (x_1, x_2, \dots, x_d)$. Thus $f_i(\vec{x})$ can be obtained by training SVM on \vec{x} . It can be computed as follows:

$$\begin{aligned} f(\vec{x}_1) &= \sum_{i=1}^K q_i \cdot d(f_i, \vec{x}_1) \cdot f_i(\vec{x}_1) \\ &= q_1 \cdot d(f_1, \vec{x}_1) \cdot f_1(\vec{x}_1) + \dots + q_K \cdot d(f_K, \vec{x}_1) \cdot f_K(\vec{x}_1) \end{aligned}$$

.....

$$\begin{aligned} f(\vec{x}_k) &= \sum_{i=1}^K q_i \cdot d(f_i, \vec{x}_k) \cdot f_i(\vec{x}_k) \\ &= q_1 \cdot d(f_1, \vec{x}_k) \cdot f_1(\vec{x}_k) + \dots + q_K \cdot d(f_K, \vec{x}_k) \cdot f_K(\vec{x}_k) \end{aligned}$$

Table 144.2 Characteristics of datasets

Data set	Size	Features	Ratio
Wdbc	569	9	1.9
Wpbc	198	9	3.1
Sick	3772	29	15.3
Letter	20000	16	24.3
Abalone	4280	8	40.1
Mcd	29508	6	100.1

$$\begin{pmatrix} f(\bar{x}_1) \\ \vdots \\ f(\bar{x}_k) \end{pmatrix} = (q_1, \dots, q_K) \times \begin{pmatrix} d(f_1, \bar{x}_1), \dots, d(f_K, \bar{x}_1) \\ \vdots \\ d(f_1, \bar{x}_k), \dots, d(f_K, \bar{x}_k) \end{pmatrix} \times \begin{pmatrix} f_1(\bar{x}_1) \\ \vdots \\ f_K(\bar{x}_k) \end{pmatrix} \tag{144.11}$$

We get q_1, \dots, q_k by solving these equations. The function in (144.12) is used to predict the label of a new example.

$$f(\vec{x}) = \sum_{t=1}^K q_t \cdot d(f_t(\vec{x}), \vec{x}) \cdot f_t(\vec{x}) \tag{144.12}$$

144.4 Experimental Results

Information about experiment data sets [9] is summarized in Table 144.2. The imbalance ratio varies from 1.9 to 100.

We randomly divide each dataset into two subsets through stratified sampling. We use 80% of the instances to train the classifiers and 20% to evaluate the performances of the classifiers. Such process is repeated 10 times for each data set. In the experiment, we effectively compared the performance of our proposal against three traditional techniques under the same experimental conditions: (1) a single SVM without any preprocessing step, denoted as SVM; (2) a single SVM with over-sampling minority class, denoted as SVM + SMOTE; (3) a single SVM with our proposed method, denoted as LCSVM; (4) a single SVM with our proposed method and over-sampling the minority, denoted as LCSVM + SMOTE.

It can be seen in Tables 144.3 and 144.4, our approaches return the best results not only in sensitivity but also in specificity. It does not have a significant impact in terms of sensitivity and specificity while the imbalance ratio is low, as their values are already high in the traditional SVM method. However, compared with the high imbalanced datasets, such as Letter, Abalone and Mcd, the imbalanced ratio is vary from 24 to 100, we can see that a single SVM is worst of all, this is because the class-boundary learned by the SVM is skewed towards the majority

Table 144.3 Performance in terms of sensitivity

Data set	SVM	SVM + SMOTE (k = 1)	SVM + SMOTE (k = 9)	LCSVM	LCSVM+ SMOTE (k = 1)	LCSVM+ SMOTE (k = 9)
Wdbc	0.9463	0.9469	0.9469	0.9486	0.9511	0.9511
Wpbc	0.4997	0.6711	0.6711	0.6839	0.7012	0.7012
Sick	0.7851	0.8702	0.8809	0.9076	0.9289	0.9460
Letter	0.6572	0.6607	0.7269	0.7091	0.7597	0.7648
Abalone	0.4435	0.4237	0.6518	0.7909	0.8417	0.7835
Mcd	0.1929	0.2234	0.3632	0.4121	0.4219	0.4408

Table 144.4 Performance in terms of specificity

Data set	SVM	SVM + SMOTE (k = 1)	SVM + SMOTE (k = 9)	LCSVM	LCSVM+ SMOTE (k = 1)	LCSVM+ SMOTE (k = 9)
Wdbc	0.9632	0.9618	0.9618	0.9620	0.9638	0.9638
Wpbc	0.7859	0.8100	0.8100	0.8018	0.8109	0.8109
Sick	0.8996	0.9019	0.9063	0.9187	0.9247	0.9388
Letter	0.8094	0.8129	0.8306	0.7844	0.8488	0.8461
Abalone	0.9299	0.9299	0.9334	0.9335	0.9335	0.9374
Mcd	0.9177	0.9031	0.9266	0.8988	0.8999	0.9298

class. Note that among the benchmark data sets, the imbalance ratios in Abalone and Mcd are more than 40, both SVM and SVM + SMOTE have high specificity but low sensitivity in the two highly imbalanced data sets, while LCSVM + SMOTE has balanced values between Sensitivity and Specificity.

Figure 144.2 visualizes comparison results on F-Measure analysis and G-Mean analysis. The value on the horizontal axis is formatted as data-name. Notice that in Fig. 144.2, there is a rarely difference between our method and the other methods on Wdbc data, because of the ratio in the Wdbc is only 1.9. With the imbalance ratio getting large, ours is more effective than the other methods. In addition, SVM + SMOTE approach seems to perform as well as our method in the Abalone, it is largely depend on the value of parameter k, from the experiment results, it is obviously that it is effective method only when k is assigned the highest value, here we test k equal to 9. It shows that our method outperforms the benchmark methods when imbalance ratio is high.

144.5 Conclusion

This chapter introduces a novel approach to handle the class-imbalance problems where local-instances distributions are well examined to make up the global classification model. We perform a comprehensive set of experiments; the results show that the global prediction model is not only capable of improving the

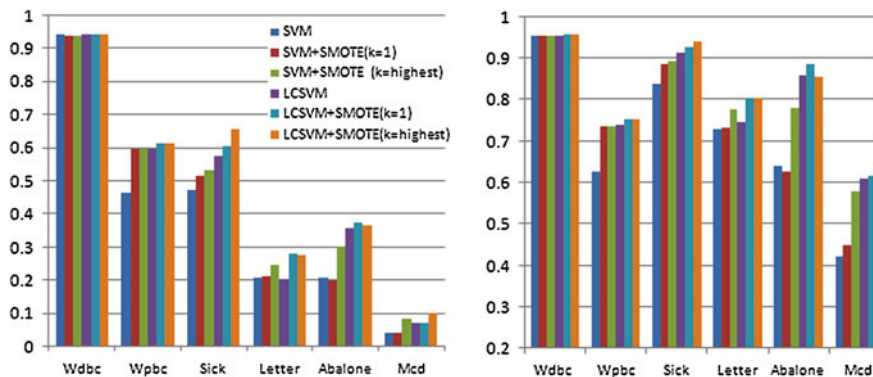


Fig. 144.2 Effectiveness comparison on F-measure and G-mean

performance on rare classes, but also considering the true local distribution of imbalanced data. Intuitively, the algorithm almost stable and even better than the benchmark SVM + SMOTE classifier in terms of the G-mean and the F-measure, which are important measures used in detection of rare events.

Acknowledgments This work is supported by research project of “SUST Spring Bud” (2010AZZ179).

References

1. Weiss GM (2004) Mining with rarity: a unifying framework. SIGKDD Explor Newsl 6(1):7–19
2. Ni W, Xu J, Liu T et al (2010) Acronym extraction using SVM with Uneven Margins. In: Proceedings of the 2nd IEEE Symposium on web society, pp 132–138
3. Chawla NV, Japkowicz N, Kotcz A (2004) Editorial to the special issue on learning from imbalanced data sets. ACM SIGKDD Explor 6(1):1–6
4. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer et al (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
5. Wilson DL (1972) Asymptotic properties of nearest neighbor rules using edited data sets. IEEE Trans Syst Man Cybern 2:408–421
6. Li DC, Liu CW, Hu SC (2010) A learning method for the class imbalance problem with medical datasets. Comput Biol Med 40(5):509–518
7. Zhou ZH, Liu XY (2006) On multi-class cost-sensitive learning. In: Proceedings of the 21st national conference on artificial intelligence, Boston, pp 567–572
8. Chen C, Liaw A, Breiman L (2004) Using random forest to learn imbalanced data. Technical Report 666, Department of Statistics, Berkeley
9. Latecki LJ, Wang Q, Oknar-Tezel SK, Megalooikonomou V (2007) Optimal subsequence bijection. IEEE Int Conf Data Min 0:565–570

Chapter 145

Access Frequency Based Energy Efficiency Optimization in Data Centers

Jie Wang, Yangang Cai, Chunxiao Liu, Jiabin Luo and Xinyi Ding

Abstract With the rapid growth of various data, large-scale data centers are built to store and analyze them. These data centers are consuming enormous energy every year, while most of them are wasted due to the low utilization of servers. In recent years, many efforts are carried out to reduce the energy cost as well as ensure the QoS, while they do not distinguish different data with different access frequency. In this chapter, an efficient energy optimization and management system is proposed based on the concept of different data weights caused by their access frequency. We collected the data's frequency of being requested in large-scale data centers and found that some data are frequently visited and always in an active state, while some data are seldom visited. We assume that data with different access frequencies have different weights, thus different management

J. Wang · Y. Cai · C. Liu (✉) · J. Luo · X. Ding
College of Computer Science and Information Engineering,
Zhejiang Gongshang University, Hangzhou 310018, China
e-mail: cxliu@mail.zjgsu.edu.cn

J. Wang
e-mail: wangjie1206@126.com

Y. Cai
e-mail: caiyangang@126.com

J. Luo
e-mail: 1535976313@qq.com

X. Ding
e-mail: dxywill@yahoo.cn

Y. Cai
School of Computer and Information Engineering,
Peking University, Shenzhen 518055, China

strategies will be adopted to save energy. In order to guarantee the availability of data, two redundancy policies: Replication and erasure coding will be discussed in this chapter. At last, we bring forward a policy about how to avoid hot spot caused by our method.

Keywords Energy efficiency optimization · Data centers · Access frequency · Servers

145.1 Introduction

A report conducted by the IDC shows that the total amount of data in the world is 0.18 ZB in 2006 and the number will become 1.8 ZB or 10 billion TB in 2011 [1]. In order to deal with these huge data, many large-scale data centers are built. Such data centers consist of thousands of servers running all the time, therefore consuming a great deal of energy.

The hardware, the energy and the management are three main parts in the running cost of these data centers. In this chapter, we focus on the system optimization of energy efficiency. When designing a data center, for the reason of peak queries and extensibility, more servers will be configured than the normal needs. As a result, a huge number of electricity is consumed by data centers every year. Study shows that in 2006 the cost of energy consumed by IT infrastructures in US was estimated as 4.5 billion dollars and it is likely to double by the year 2011 [1]. The fact is that most of the energy is wasted because most of the time the servers are in idle status waiting for queries. The average utilization is disappointingly below 20% [2].

The energy cost in data centers consists of servers and cooling system. A report shows every 1 W of power used to operate servers often correspond to 0.5–1 W of power required by the cooling system. To improve the efficiency of energy in servers, many practices were conducted during the passed few years. Dynamic voltage and frequency scaling (DVFS) [3] provides different voltages and frequencies for the servers according to their different status checked; Virtual machines (VM) divides a physical machine to several VM with different kinds of applications running on them; Chassis consolidation [4] is another widely used technique, which saves energy by running the chassis as little as possible. It can assign the incoming tasks to the minimal number of chassis and shutdown the unused chassis, thus save energy for us.

Another significant part of cost is the energy consumed in cooling system such as fans and air conditioning system in the computer room. However, this topic is beyond our chapter and the interested reader can refer the hierarchical control scheme proposed by Zhikui for this problem [5]. All the above approaches mentioned have one common characteristic that they treat the data in all servers equal. However, we found that the probability of a piece of given data being visited

during a period of time varies greatly. As a result, we propose to take different strategies for these different data, in which we set their weight according to their frequencies of being visited.

In data centers, we also found that compared with the CPU or memory management cost, the power used to drive chassis is more obvious [4]. Therefore, we try to arrange those data with similar access frequencies into the servers of the same chassis, then the high active chassis will be frequently visited and those are in low spirit can be set to the state Nap [2]. The main challenges are as follows: (1) what kind of strategies should be taken to measure the weight of data in servers? (2) How to organize the data according to their weight, and what kind of efficiency mechanism should be taken for different weight. (3) How to avoid the hot spots?

145.2 Related Work

During the passed few years, many approaches and models are proposed to optimize the energy efficiency in data centers. DVFS [3] is such a kind of technology that it can dynamically adjust the server's voltage and frequency based on the different demands of computation. As we all know, only decrease the frequency or voltage cannot save energy for us. DVFS collects the information about the current status of the system and computes its current load, and then makes a prediction about the computing ability the system needs next stage, thus the system can adjust the frequency and voltage accordingly. DVFS can be implemented by means of hardware or software.

VM is another method used to improve the utilization of energy efficiency in data centers. It divides a physical machine into many VM with different kinds of operating systems reside on them, then applications that require different platforms can run on a physical machine with the guarantee of resources violation provided by the VM technology. The Xen [6] can host up to 100 VM instances simultaneously on a modern server with little overhead.

Pakbaznia and Pedram [4] proposed that the difference between power consumption values for different utilizations is small compared to the total power consumption of the chassis. It is desirable to assign the incoming tasks to the minimum number of chassis so that the remaining ones can be turned off. This is just the concept of chassis consolidation.

To guarantee the reliability and availability, the redundancy mechanism is needed and replication and erasure codes [7] are two main solutions. Chained declustering [8] is a replication scheme that stripes the partitions of a data set two times across the nodes of the system, thereby doubling the required disk space. If the faults do not occur in adjacent nodes, the arrangement of the replication allows for balanced workload distribution and high availability when some nodes are offline. The key idea of the erasure codes is that the data at the sender is divided into m blocks and by multiplying a $m \times n$ matrix (a Vandermonde matrix

Table 145.1 The top ten queried keywords in Google data center during a period

Rank	Keyword	Frequency
1	People	18818
2	(Blank)	16332
3	Music	8971
4	Fruit	8922
5	Christmas	7589
6	Business	5508
7	Food	5413
8	Woman	5008
9	Family	4787
10	Computer	4647

for example) these m blocks are coded to n blocks ($n > m$). If we can get at least k blocks ($k \geq m$), the data can be reconstructed. Compared with the method of Replication, the Erasure codes can theoretically provide the same QoS with less redundancy.

Ideally, we just want to simply turn off the idle system, so that we can save energy. To our disappointment, a large fraction of servers exhibit frequent but brief burst of activity. And we can hardly predict what the users are going to do. The power nap [2] designed the entire system to transit rapidly between a high performance active state and a minimal-power nap state in response to instantaneous load. Then energy wasted during the idle state can be reduced. What is more, we can set the chassis in which less used data (for example, from Table 145.1 we can see that compared with the keyword “people”, the “computer” is relatively less visited) with less weight reside to Nap state, because these data are not frequently visited and they can stay in Nap for a relative long and continuous time, thus much power can be saved.

However, we should also understand that some data are highly frequently visited in some period. For example, during the FIFA world cup, the data about football must be more frequently queried than the data about the Astrology. Hence during these special periods we also can reorganize these data across the servers to save energy.

145.3 Access Frequency Based Energy Efficiency Optimization

A data center is typically a facility with several rows of server racks, each row consists of several racks and each rack consists of several chassis in which are multiple (blade) servers. The servers can be single-core or multi-core processors. Cold air supplied by the air conditioning unit comes from the floor [9], cools down the servers, and then is extracted from the room. As Ehsam [4] pointed out that although the utilization level of a server varied by the type of the task running on

it, the difference between the power computation of a fully-utilized server and a 40–60% utilized one is less than 2% for a typical blade server. The difference between power consumption for different utilizations is small compared with the total power consumption of the chassis. Consequently, more attention will be paid to the distribution of chassis in this chapter. In our proposed energy efficiency management system, we arrange the related data that usually have similar access frequency to the same chassis with the VM technology, thus these data can be visited simultaneously by different users. The remained ones with less weight will be set to different stages according to some strategies.

The system operation mainly consists of the following steps: (1) All the data in the servers has a counter indicating its weight. (2) Requests are sent to the data center. (3) These data being visited increase their weight by one. (4) During a period of time, local monitor collect the information about the weight of the data that reside on servers and send them to the global monitor. (5) Transition starts. Global monitor rearrange the location of data and reorganize them according to their weight. Those that have similar weight will be organized in the same chassis. (6) After a fixed time, those high weight data and low weight data will exchange their chassis to avoid hot spots. We discuss our weight setting scheme, the redundancy policies and the hot spots avoidance strategy in detail at the following subsections.

145.3.1 Weight Setting Scheme

Following describes how to set different weight for the data with different access frequencies. One alternative way is that every piece of data has a weight counter whose initial value is 0, which represents how many times the data has been visited during a given period of time. Every time the data is visited, the counter pluses by 1. The bigger the counter, the higher weight it has. And, there is a monitor on every chassis takes charge of the local data's weight. Every a period of time, the local monitor will send these weight information to the global monitor [10], and then the data transition begins. These with similar weight will be organized into the same chassis (may be different servers). When the weight of data is below a fixed number, we can set the chassis that these data reside into a Nap state.

Although support for sleep states is widespread in handhold, laptop and desktop machines, these states are rarely used in current server systems. Unfortunately, the high restart latency typical of current sleep states renders them unacceptable for interactive services; current laptops and desktops require several seconds to suspend using operating system interfaces. Thus the availability cannot be guaranteed.

Meisner et al. [2] studied the idle/busy intervals in actual data centers, and concluded that power nap is effective if state transition is below 10 ms. They model power nap as an M/G/1 queuing system and compare it with the DVFS model theoretically. From Table 145.2 we can see that using the power nap is much more efficient than the DVFS under some condition. To implement this

Table 145.2 Per-workload energy savings of power nap and DVFS

Workload	Power nap	DVFS
Web	59	23
Mail	35	21
DNS	77	23
Shell	55	23
Backup	61	23
Cluster	34	18

system there are mainly two requirements. (1) Fast transitions. Transition speed is the dominant factor in determining both the power savings and the response time. 10 ms transition speed is required to keep the benefits of the power nap, and a faster speed may reduce the overhead of power nap. (2) Minimizing power draw in Nap state. One of our research topics is to extend the continuous time in nap state, so we can provide little power to the servers.

In order to fulfill these two requirements, there are still a lot of efforts need to be done considering one study shows that these idle periods in servers are usually less than 1 s [2]. In our future work, we will study the cost of the power nap and whether it is suitable to be applied to our strategy. After all, this technology challenges the hardware at the same time.

145.3.2 Our Redundancy Policies

When reducing the energy waste in data center, we should also keep the QoS to guarantee the reliability and availability of the data. There are mainly two policies when considering redundancy: Replication and Erasure codes (a kind of software RAID [11]). Of course there are many other approaches like the LPCA [12, 13] algorithm, but its main idea is similar to Erasure codes, so we do not plan to make a detailed discrimination. As we talked above, Replication is very simple and widely used in industry. In some degree it has become the default standard for storing data.

In Google data center, every file is divided into several chunks (normal 64 MB, but the size can be adjusted to meet different requirements). Every chunk has three replications and they are located in different racks. It is nearly impossible that the three servers with the same data reside on will break down at the same time. This architecture is easy to understand and implement, and can offer a high QoS when the query request is sent to the data center. The server master will find the nearest replication. After a serial computation and reconstruction, the result will be sent back to the user within just a few ms.

In this chapter we only focus on the linear block codes as they are simple and appropriate for many applications. Besides we put forward the drawbacks exist in

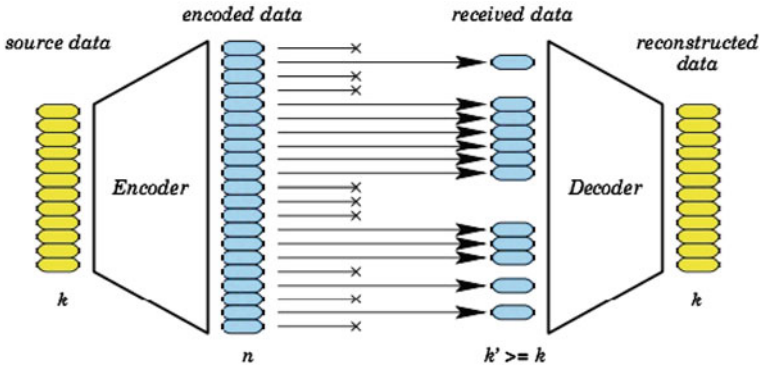


Fig. 145.1 Erasure codes [6]

Erasure codes and the corresponding solutions for future research. Figure 145.1 describes the process of the encoding and decoding of data. The source data at the sender are first divided into k blocks, then encoded to produce n blocks. In such a way any subset of k encoded blocks suffices to reconstruct the source data. Linear codes can be analyzed using the knowledge of linear algebra. Suppose the source data $x^T = x_0 \dots x_{k-1}$. G is an $n \times k$ matrix, then a (n, k) linear code can be represented by

$$y^T = Gx^T \tag{145.1}$$

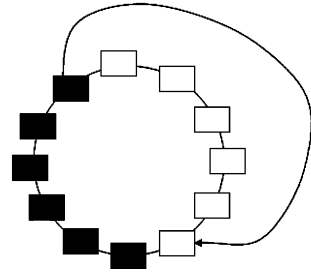
For a proper definition of the matrix G , if we can find that k components of y are still available across the data center, then source data can be reconstructed. The problem here we should consider is how to find the matrix G . Since G is an $n \times k$ matrix with rank k , any subset of k encoded blocks should convey information on all the k source blocks. As a consequence, each column of G can have at most $k-1$ zero elements, an alternative choice is the Vandermonde matrix and its determinant

$$\prod (x_j - x_i), \quad j > i, \quad n \geq i, \quad j \geq 0 \tag{145.2}$$

Here what we should also pay attention to is that every y_i in y is much bigger than x_i . If every x_i has b bits and an element of matrix a , the width of each y_i is $a + b + \lceil \lg 2 k \rceil$ ($\lceil \lg 2 k \rceil$ is the indicator of every block). This information will be used for reconstruction because it is definitely inefficient to transfer great amount of data among the servers.

In order to overcome the expansion of data, Luigi [7] proposes to work in a finite field. Roughly speaking, a field is a set in which we can make add, subtract, and multiplication operations, which means that the result of sums and products of field is characterized by having a finite number of elements. Most of the properties of linear algebra apply to finite fields as well; the main advantage of using a finite field is the closure property which allows us to make exact computation on field

Fig. 145.2 Avoiding hot spots



elements without requiring more bits to represent the results. And what we should do is to map the data into field elements. Operate these data according the rules of field and then apply the inverse mapping to reconstruct the desired results.

Till now, we know that a software implementation of Erasure codes is expensive. Erasure codes also have the following drawbacks: (1) it will take great cost to upgrade data.(2) A query request needs many servers to cooperate with each other. In our future work we will study further about the Erasure codes or other algorithms that are more efficient.

145.3.3 Avoiding Hot Spots

We may find that if we organize these high level data into the same chassis, if we do not take some strategies, these chassis's life will be short because they will never get time to have a rest. Upgrading devices will add costs. So we may want these chassis in data center take turns to support these high active data.

One alternative solution is illustrated in Fig. 145.2, all these chassis are logically linked like a ring, and these chassis that are high active form a queue, after a period of time, the head active chassis will be inserted to the tail, thus avoiding the problem.

145.4 Conclusion

To improve the energy efficiency in data centers, this chapter proposes an access frequency based optimization and management system that organizes those data that have the similar frequency of being visited into the same chassis, thus those data that are not frequently visited can be set to the Nap state. We realize that the access frequency of different data varies greatly, they should be arranged according to their weight when distributing them across the data center, thus we can elongate the continuous time of being Nap of those data that are not always visited.

There are much efforts need to be done to promote the algorithm, we try to find a new method that are less complex and can offer the same QoS and may adopt different redundancy strategies for different level chassis in our future work.

Acknowledgments This work is funded by the National Natural Science Foundation of China under Grant No. 61003188, and the Technology Plan Program of Zhejiang Province under Grant No. 2009C11034, and the Zhejiang Provincial Natural Science Foundation of China under Grant No. Y1111159, No. Z1101243 and No. Z1111051.

References

1. White T (2010) Hadoop: the definitive guide
2. Meisner D, Gold BT, Wenisch TF (2009) PowerNap: eliminating server idle power. *ACM SIGPLAN Notices* 44(3):205–216
3. Semeraro G, Magklis G, Balasubramonian R, Albonesei DH, Dwarkadas S, Scott ML (2002) Energy-efficient processor design using multiple clock domains with dynamic voltage and frequency scaling. In: *Proc. of the eighth international symposium on high-performance computer architecture (HPCA'02)* pp 29–40
4. Pakbaznia E, Pedram M (2009) Minimizing data center cooling and server power costs. In: *Proc. of the 14th ACM/IEEE international symposium on low power electronics and design* pp 145–150
5. Wang ZK, Tolia N, Bash C (2010) Opportunities and challenges to unify workload, power, and cooling management in data centers. *ACM SIGOPS Oper Syst Rev* 44(3):41–46
6. Dong KJ, Feng JH, Yan BP (2005) A distributed file system model based on erasure code. *Comput Eng* 31(20):93–95
7. Rizzo L (1997) Effective erasure codes for reliable computer communication protocols. *ACM Comput Commun Rev* 27(2):24–36
8. Lang W, Patel JM, Naughton JF (2009) On energy management, load balancing and replication. *SIGMOD Record* 38(4):35–42
9. Hamann HF (2008) A measurement-based method for improving data center energy efficiency. In: *Proc. of the IEEE international conference on sensor network, ubiquitous and trustworthy computing* pp 312–313
10. Beloglazov A, Buyya R (2010) Energy efficiency resource management in virtualized cloud data centers. In: *Proc. of the tenth IEEE/ACM international conference on cluster, cloud and grid computing* pp 826–831
11. Patterson DA, Gibson G, Katz RH (1988) A case for redundant arrays of inexpensive disks (RAID). In: *Proc. of the ACM SIGMOD international conference on management of data* pp 109–116
12. Zhang W, Ma JF (2007) LPCA—Data distribution algorithm in distributed storage. *Syst Eng Electron* 29(3):453–458
13. Zhang TX, Pan F, Yang XY, Liu Z, Zhang W (2010) Improved algorithm of LPCA. *Comput Eng Appl* 46(9):83–84

Chapter 146

Application of Data Mining Technology in Jewelry Design

Cen Qin

Abstract This paper describes the status of jewelry design briefly, and points out the relationship between the jewelry design and the data accumulated by enterprises, then proposes the project that brings the data mining technology into the jewelry design. Forecasting the enterprise's future sale with Microsoft Time Series model furthermore validates the mining model through analysis of historical data and existing data. It provides a new way for decision-making of jewelry enterprises, which will help the leadership do a more accurate market positioning and grasp the future development trend of jewelry, has a certain practical value.

Keywords Data mining · Microsoft time series · Jewelry design · Sales forecast

146.1 Introduction

Jewelry and clothing have appeared at the same time since the beginning of human. How to integrate the modern culture into jewelry, and how to design new fashion jewelry, and what kind of design can be accepted by consumer, which is a new question for jewelry designs. Jewelry is designed by designers, but it is according to the annual business plan for different regions and ethnic groups, and the style, material and the total production of jewelry are also different. The enterprise leaders must understand the jewelry's sale of previous

C. Qin (✉)
School of Information and Engineering,
Wenzhou Medical College, Wenzhou, Zhejiang, China
e-mail: cenqin8899@126.com

years as well as the characteristics of various regions in order to position jewelry's fashion trends accurately. Business leaders should master the data of every department before decision-making. With the development of informationization the enterprise has accumulated a large amount of data, is fragmented and non-standard such as noise data, vacancy data and inconsistent data, which brings trouble to enterprise's decision-making. Nowadays, the common problem of enterprises is that the amount of data is very large but there is little valuable information. The appearance of data mining gives decision-making support for enterprises. Enterprises can make use of data mining and business intelligence technology to forecast the fashion trend of jewelry, improve the market hit rate and reduce the rate of poor sale via analysis and mining of the historical information.

146.2 Microsoft Time Series Algorithm

Data mining is a new commercial information-processing technology, which main feature is extraction, conversion, analysis and treatment for other models on large number of data in business database to extract the auxiliary data to support decision-making. Making use of the existing data, forecasting can analyze the future jewelry's demand. Enterprise leader can make the decision for future demand according to the forecasting, which can get maximize profits with the least resources. The exact forecasting can help enterprise do the better market positioning and the more efficient production plans, which reduces the phenomenon of lost sale and inventory, and strengthens production, materials planning to make full use of corporate resources [1].

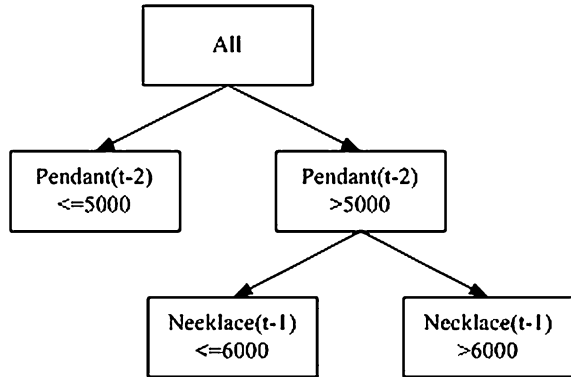
The Time Series algorithm is a rational sales forecasting model, which can simulate the seasonality [2], periodicity and the trend of jewelry sale based on the earlier sales data, so can revise the subjectivity of forecasting for staff to improve the forecasting accuracy effectively. The basic thought of Time Series is that arrange the sales data according to the order of annual, month to observe the trace usually.

The Microsoft Time Series algorithm is a regression algorithm provided by Microsoft SQL Server 2005 Analysis Service, which combines auto regression technology with decision technology, and also be called Auto Regression Tree (ART) [3]. It can build mining model to forecast continual lines, such as the jewelry's sale.

146.2.1 Using Multiple Time Series

In terms of Data Mining eXtension (DMX), a time series is a single case. The weekly sale amount of jewelry during the past year becomes a single case of a time

Fig. 146.1 The ART of necklace-pendant



series. A mining model may contain multiple time series. For example, a model can contain the time series of all jewelry products, including earring, necklace, ring, pendant, and so on. Series are not always independent, for example the sale of necklace and pendant may be strongly correlated. The Microsoft Time Series algorithm recognizes cross-series correlations when they exist. This is one of the unique features of this algorithm [4].

146.2.2 Auto Regression Tree

Microsoft Time Series algorithm is an autoregressive model in which the function f corresponds to a regression tree [5], in other words, in ART, the function f is represented by a regression tree. As Fig. 146.1 shows the ART of necklace-pendant built by the time series data. The first tree split is on pendant sale two months ago. If pendant sale were more than 5,000 two months ago, another split attribute would be necklace's sale value of previous month. In the case of previous month's necklace sale were less than 6,000, and the regression formula for necklace as follows:

$$\text{Necklace}(t) = 3.02 + 0.72 * \text{Pendant}(t - 1) + 0.31 * \text{Necklace}(t - 1). \quad (146.1)$$

146.2.3 Seasonality

The most of time series have seasonal mode. The ART uses the process of case transform to deal with the seasonality, and there also use the seasonal parameter Periodicity_Hint to add historical data point. Jewelry is a seasonal product, and its periods are three months in a general way. In the case of the periods being three

months, the ART includes the observational data as $\text{Necklace}(t-3)$, $\text{Necklace}(t-6)$, ..., $\text{Necklace}(t-8*3)$, $\text{Pendant}(t-3)$, $\text{Pendant}(t-6)$, ..., $\text{Pendant}(t-8*3)$, and then the regression tree will make use of these observational data when it is split, and also use in the regression formula. When there are several periodicities, ART algorithm will add many lines in the case table based on seasonality. If there is no appointed periodicity, the Microsoft Time Series algorithm will measure the seasonality automatically with fast Fourier transform.

146.2.4 Making Historical Predictions

After a time series model is processed, it can perform forecasting as well as historical prediction. In DMX, there is a predefined prediction function call `PredictTimeSeries`. For example, `PredictTimeSeries(Necklace, 5)` returns a nested table of five rows, representing the Necklace sale in the next five months.

Microsoft Time Series algorithm can use the model to predict the jewelry's sale over the past, for example, `PredictTimeSeries(Necklace, -10, -5)` returns the historical prediction of necklace sale past 5–10 months. The time series algorithm predicts on historical data points using `Historical_Model_Count` and `Historical_Model_Cap` parameter, here `Historical_Model_Count` specifies the number of historical models which need to be built, while `Historical_Model_Gap` specifies the time increments of historical models.

146.3 Application of Data Mining Technology for Jewelry Design

It classes the jewelry according to country, area, material, style and worn bearings based on the analysis of historical data and existing data, and predict the jewelry demand and develop trend of different country and area with data mining technology. Enterprise can make the appropriate stock plans, inventory plans, market strategies, and get better market position.

Microsoft time series algorithm makes use of linear regression decision tree to analyze the time-related data, such as sale of month and yearly profit. Jewelry is seasonal product, and the sale change is related to time, so this paper built a new sale forecasting model of jewelry with the trait of time series forecasting model. Time series forecasting model analyzes the future sale trend of jewelry objectively and effectively, and avoids the subjective judgment for historical data during sales forecasting according to sales historical data, and simulates the seasonality, periodicity and trend effectively. There are many factors influence jewelry's sale, such as sales area, jewelry category, customer structure, seasonality and the competition of similar companies, furthermore enterprise produces a large number of sales data

everyday. How to make use of the huge data to analyze sales status, and get implied useful information, which has important meaning.

146.3.1 Data Ready

The first step of data ready is collecting data and making sure the analytical data source, which may need to capture some outer data.

The data of data warehouse is from inner business system and outer data source. The business system of enterprise is built in different background, and it faces different application and developer, which data structure, store platform and system platform is also different, furthermore the format of outer data source is more intricate. As the data are captured by many data sources, there exists mistakes and inconsistency before the data transform into data warehouse, which needs to detect those data and correct the abnormal data by certain means, which transforms the business model into analysis model that is the process of data clean and transform. The data should be clean as follows: the inconsistency of field's length and data value, the violation of integrality restraint and so on. For example, the time's format is adopted as 2006-01-01 in the ERP (Enterprise Resource Planning) system, but it is 1st, January 2006 in OA (Office Automation) system, so there needs to transform the time's format into same format.

In order to carry out data mining, there needs to merger multiple data sources into the uniform data store to solve the question of semantic fuzzy that is called data integration, which refers mode integration, the handling of redundant data and the handling and detection of conflict data.

The data of XG jewelry Co., Ltd come from the ERP system and OA system, such as product storage, which comes from the jewelry storage table of the ERP system, the jewelry production table and the material storage table of OA system. After preprocessing the data, merged them into data warehouse, the integration as shown in the Fig. 146.2.

When the data have prepared, run the interface program, put the cleaned and transformed data into data warehouse. This system adopts Integration Services owned by SQL Server 2005 as ETL tool, which is new enterprise-level data integration platform, can extract and transform the data from several sources(such as XML data file, flat file and relation source), then load these data to one or more goals. Usually, when the business system is not busy, it will automatically load data regularly. The data of fact table are loaded from the ERP system and OA system. The quantity of data is very great in central fact table with star mode, which will add continuously along with time goes. Here need to concern the performance of loading tool, meanwhile should make sure finish the loading in certain time.

The sales log is stored in the data warehouse, which provides the basic information of customers' purchase log. The main attributes are product code, customer code, time code, sales region code, sales quality, selling price and so on.

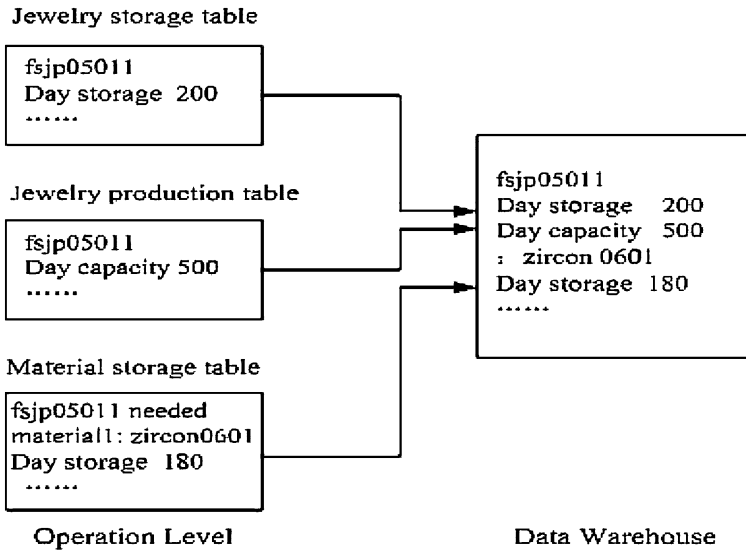


Fig. 146.2 Data integration from the operation layer to the data warehouse layer

146.3.2 The Realization of Model

This paper makes use of the historical sales data of January 2005 to June 2007 of XG jewelry Co., Ltd to predict the sale of necklaces and pendants in autumn of that year. The month sale act as the predicted time series data, stored in relational databases with each line representing a time series, which sectional drawing as shown in the Fig. 146.3.

According to the data of time series table, using DMX to build the time series model, the part code show as following:

```

Create Mining Model NecklacePendantTsa
(
Time date key time,
Necklace long continuous predict,
Pendant long continuous predict
) Using Microsoft_Time_Series
    
```

In time series model, there is not need new example, the predicted value derived directly from the training examples. After the analysis and handling of mining algorithm, there can get the forecasting figure of necklace-pendant as Fig. 146.4.

The real line represents the actual sale, and the dashed represents the predicted sale. The dashed shows the predicted sale of history before July 2007, and the later dashed shows autumn's predicted sale in this year, and the vertical shows the forecasting deviation, and usually, the further for future's forecasting, the larger the prediction's deviation is.

Fig. 146.3 The input table of time series

	Time	Necklace	Pendant
▶	2005-1-1 0:00:00	2221500	2131000
	2005-2-1 0:00:00	2225000	2125000
	2005-3-1 0:00:00	2245000	2099000
	2005-4-1 0:00:00	2185000	2071000
	2005-5-1 0:00:00	2167000	2176000
	2005-6-1 0:00:00	2170000	2061000
	2005-7-1 0:00:00	2553000	2576000
	2005-8-1 0:00:00	2550000	2489000
	2005-9-1 0:00:00	2562000	2491000
	2005-10-1 0:00:00	2659000	2556000
	2005-11-1 0:00:00	2651000	2550000
	2005-12-1 0:00:00	2697000	2557000
	2006-1-1 0:00:00	2335000	2352000
	2006-2-1 0:00:00	2346000	2246000
	2006-3-1 0:00:00	2387000	2280000
	2006-4-1 0:00:00	2191000	2071000
	2006-5-1 0:00:00	2187000	2065000
	2006-6-1 0:00:00	2176000	2063000
	2006-7-1 0:00:00	2643000	2572000
	2006-8-1 0:00:00	2656000	2582000
	2006-9-1 0:00:00	2655000	2577000
	2006-10-1 0:00:00	2777000	2652000
	2006-11-1 0:00:00	2737000	2632000
	2006-12-1 0:00:00	2783000	2657000
	2007-1-1 0:00:00	2342000	2230000
	2007-2-1 0:00:00	2390000	2216000
	2007-3-1 0:00:00	2337000	2231000
	2007-4-1 0:00:00	2211000	2062000

The main aim of data mining is gain knowledge to support decision-making, so it is important to explain and estimate result.

- (1) The predicted data is consistent with the actual fact basically according to the predication result of historical sale, so it is effective to predict future sale with time series algorithm.
- (2) While there are a large number of data, analyzing the sales data with data mining tool can get surprise result, which is vital to enterprise’s decision-making and is key information for succeeding in market competition.
- (3) Here making use of time series algorithm to study and analyze a large number of historical sales data, then gain the sales change trend, and predict the sale, sales price and sales profit of future with training time series model. According

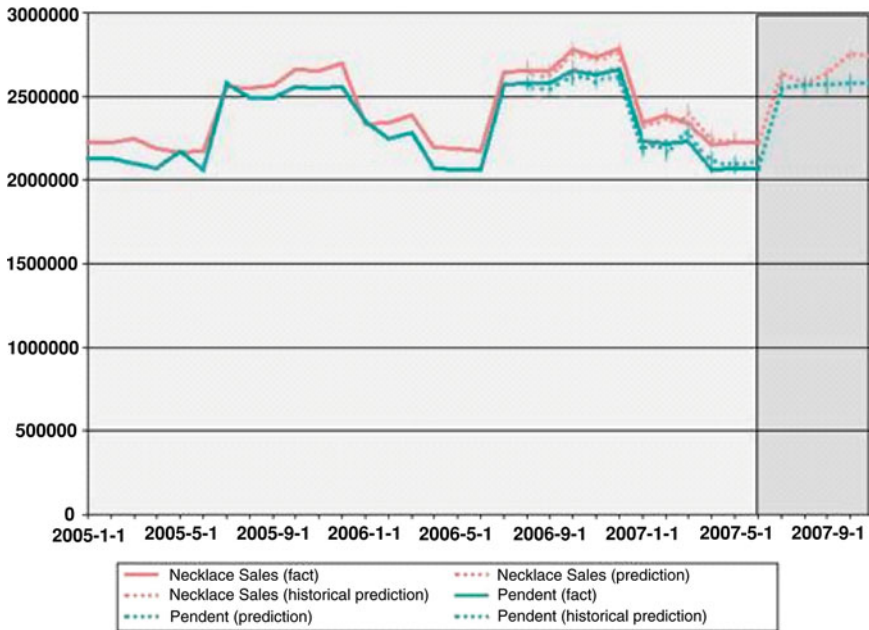


Fig. 146.4 The forecasting map of necklace-pendant

to sales prediction, enterprise can make the appropriate stock plans, inventory plans, and market strategies and so on.

146.4 Conclusion

The paper proposes that brings data mining technology into jewelry design, and takes the data of January 2005 to June 2007 of XG jewelry Co., Ltd as example to predict the jewelry sale with Microsoft time series model, and validate the accuracy of adopted mining model. It provides the new way to enterprise's decision-making, helps the leadership do a more accurate market positioning, holds future development trend of jewelry, has a certain practical value.

References

1. Liu R (2006) The application and research of data mining for sales forecast. Zhejiang University, Hangzhou
2. Wan Y, Chen S, Dai S (2006) Study on forecast method for garment sales based on time series and PERT. Res Technol 11:60–63

3. Yang Y (2006) Automobile inventory sales prediction model based on time series data mining. Wuhan University of Science and Technology, Wuhan
4. Tang Z, Jamie M (2007) Data mining with SQL Server 2005. Tsinghua University Press, Beijing
5. Wang Y (2003) Business forecasting methods. University of International Business and Economics Press, Beijing

Chapter 147

Logic Algebra Method for Solving Theoretic Problems of Relational Database

Zhang YiShun and Ju ChunHua

Abstract A logic algebra method is explored for solving key problems of sets of functional dependencies in the theory of relational database. For a given set of functional dependencies, we first get its corresponding logical function and then determine all prime implicates of the logical function in logic algebra. Each prime implicate represents a nontrivial and simplest functional dependency implied by original set of functional dependencies, and vice versa. By using these prime implicates all equivalent minimal covers, all candidate keys and closure of arbitrary sets of attributes for the given set of functional dependencies are computed unitedly and directly.

Keywords Relational database · Logic algebra · Functional dependency · Prime implicate · Candidate key

147.1 Introduction

Although the general theory of relational database has been established since Codd [1] published the fundamental paper on this subject 40 years ago, it is not hard to find that many key algorithms [2, 3] in the theory, such as determining all keys, closure of arbitrary sets of attributes, and minimal covers, are time-consuming and tedious for often testing each item in sets of data dependencies many times, and moreover some algorithms could not output complete results.

Z. YiShun (✉) · J. ChunHua
Computer and Information Engineering College, ZheJiang
GongShang University, Hangzhou

In this chapter we explore an efficient method for dealing with these problems. Due to limited space, we only discuss its application to issues about sets of functional dependencies. Its basic principle is the equivalence between the inference rules for functional dependencies and operation rules for logic algebra [4]. Its procedure is follow: for a given set of functional dependencies, first write out its corresponding logical function, then determining all prime implicates of the logical function, finally transform each prime implicate into corresponding functional dependency (abbreviated FD), which is nontrivial and the simplest, namely, without redundant determining factors. A very important point is that these FDs consist of a set with a specific property: if and only if such a FD is implied by original set of FDs, it is in the set. By using this set the above problems can be solved easily.

There are many researches on this similarity or equivalence between operations rules in logic algebra and inference rules in sets of FDs, but most results is limited to change proves in FDS to corresponding problems in logic algebra [4, 5]. This only provides a new and probably simpler prove method.

Partial basic idea discussed in this chapter has been proposed in some papers [6, 7] many years ago. Because for a given set of FDs the relation of one to one correspondence between prime implicates of its corresponding logical function and implied simplest FDs has not been clearly revealed, this thinking has not arisen due attention and in-depth study.

147.2 Preliminary Notions

A lot of knowledge about logic algebra and theory of relational database are needed in this chapter. Primary definitions listed below are just for subsequent discussion.

147.2.1 Logic Algebra

Letters stand for variables in logic algebra. A literal is the appearance of a variable or its complement. Logic variable and result of operation in logic algebra has just two value labeled 0 and 1. 0 and 1 do not represent amount, but two opposite logic states.

Definition In logic algebra there are three basic operations: AND, OR, and NOT.

“AND” operation of variables A and B is denoted as AB , namely product form. If A, B both have value 1, the result is 1, otherwise 0. A product term is one or more literals connected by AND operators, such as ABC

“OR” operation of variables A and B is denoted as $A + B$, namely sum form. If A, B both have value 0, the result is 0, otherwise 1. A sum term is one or more literals connected by OR operators, such as $A + B + C$.

“NOT” is a unary operator, “NOT” operation of variable A is denoted as \bar{A} . If A is 0, the result is 1, otherwise 0.

Definition If values of variables A_1, A_2, \dots, A_n are given, the value of variable Y is determined uniquely, then Y is called logical function of A_1, A_2, \dots, A_n , denoted as $Y = F(A_1, A_2, \dots, A_n)$.

Definition For a logical function with n variables, a standard product term, also called a minterm, is a product term that includes each variable, either uncomplemented or complemented.

Each variable appears in the form of either uncomplemented or complemented in a minterm, so n variables can consist of 2^n minterms.

Definition For two logical functions with same variables, if for any different values of variables the results of the two functions are always equal, the two logical functions are called equivalent.

Any logical function can be equivalently changed into a sum of products, namely product terms connected by OR operators. Each logic function has equivalent one and only one sum of products made of minterms that is called as sum of standard product terms. So equivalent logic functions have the same sum of standard product terms. For example A, B, C are variables, then $AB + BC = ABC + ABC\bar{C} + \bar{A}BC$. Right side of equality is a sum of standard product terms. Obviously the right part is more complex than left part.

Definition In the sum of standard product terms of a logic function, each minterm and product term obtained by merging minterms are called implicants. A prime implicate is a implicate that cannot be simplified further by merging minterms.

Definition Simplifying a logic function means to get an equivalent simplest sum of products that are made of prime implicants and has not redundant product terms. Generally there are more than one equivalent simplest sum of products for a given logic function.

147.2.2 Functional Dependency

Definition In a relational schema $R(U)$, U is a set of attributes here, a functional dependency $X \rightarrow Y$ holds if, in any legal relation, the value of $X \subseteq U$ determines the value of $Y \subseteq U$

Definition A minimal set of FDs F must satisfies following three conditions:

- (1) Every right side of a dependency in F is a single attribute.
- (2) Any $X \rightarrow A$ in F can't be derived from $F - \{X \rightarrow A\}$.

- (3) For every $X \rightarrow A$ in F and any proper subset Z of X , $Z \rightarrow A$ can not be derived from F .

Above condition 2 indicates minimal set of FDs has no redundant FD, and condition 3 indicates every FD in F has no extraneous attributes. Any set of FDs has one or more equivalent minimal sets of FDs called as its minimal cover.

Definition $R(U)$ is a scheme over attributes U , $X \subseteq U$. X is called a super key of R if functional dependency $X \rightarrow U$ holds over R . If does not exist any $X' \subset X$, such that $X' \rightarrow U$ holds, then X is a candidate key of R .

Definition Suppose that F is a set of FDs over attributes U , and $X \subseteq U$, $X_F^+ = \{A | X \rightarrow A \text{ is implied by } F\}$, X_F^+ is called the closure of X over F .

147.2.3 Equivalent Theorem

The discussion below does not consider two extreme cases for a FD, namely $X \rightarrow \phi$ and $\phi \rightarrow X$.

Definition Supposing that right side of each FD in a set of FDs is a single attribute. If there is more than one, it can be split up into a few FDs of which every FD has a right side with a single attribute. For each FD $P_1P_2, \dots, P_n \rightarrow A$ in the set, transform it into a product term $P_1P_2, \dots, P_n\bar{A}$, named corresponding product term of the FD. The sum of all such product terms consist of the corresponding logic function for the set of FDs. Conversely $P_1P_2, \dots, P_n \rightarrow A$ is the corresponding FD of product term $P_1P_2, \dots, P_n\bar{A}$, the corresponding set of FDs for a logic function.

The following equivalence theorem [5] is the start point for exploration in this chapter.

Equivalence Theorem Let σ be a functional dependency, Σ be a set of FDs, and σ' and \sum' be the corresponding product term and the corresponding logic function, if and only if σ' is an implicate of \sum' , σ is implied by Σ .

For convenience we call the following theorem prime implicate theorem [8] that reveals a close and useful relation between prime implicates and functional dependencies.

Prime Implicate Theorem Any prime implicates implied by the corresponding logic function of a set of FDs contains one and only one complemented variable.

147.3 Application of Logic Algebra Method

In this section we will study how to use logic algebra method for computing all minimal covers, all candidate keys and closure of arbitrary set of attributes for a given FDS.

147.3.1 Determine Minimal Cover

According to Equivalence Theorem and Prime Implicate Theorem we can get following conclusions:

Corollary Suppose Σ is a FDS, Σ' is corresponding logical function, and σ' is a prime implicate of Σ' , then corresponding FD σ of σ' is implied by Σ and σ' has not redundant attributes in its determining factors.

Corollary There exist a one to one correspondence between simplest and-or formula of corresponding logical function of a set of FDs and equivalent minimum set of FDs.

How to compute minimum set of FDs is reduced to simplify corresponding logical function. First getting related logical function of original set of FDs, then simplifying the logical function to obtain simplest and-or formula. Its corresponding set of FDs just is equivalent minimum set of FDs. Different simplest and-or formulas correspond different minimum sets of FDs.

Example Determine all equivalent minimum FDS for the given set of FDs as follow:

$$F = \{ AB \rightarrow D, CD \rightarrow B, CD \rightarrow A, A \rightarrow B, D \rightarrow C \}$$

Answer The corresponding logical function is

$$Y = AB\bar{D} + A\bar{B} + D\bar{C} + CD\bar{B} + CD\bar{A}$$

All prime implicates of Y are $A\bar{B}$, $A\bar{C}$, $A\bar{D}$, $D\bar{A}$, $D\bar{B}$, $D\bar{C}$

All simplest and-or formulas and its corresponding minimum sets of FDs of Y are:

$$A\bar{D} + D\bar{A} + D\bar{B} + D\bar{C} \text{ corresponds } \{A \rightarrow D, D \rightarrow A, D \rightarrow B, D \rightarrow C\}$$

$$A\bar{B} + A\bar{D} + D\bar{A} + D\bar{C} \text{ corresponds } \{A \rightarrow B, A \rightarrow D, D \rightarrow A, D \rightarrow C\}$$

$$A\bar{C} + A\bar{D} + D\bar{A} + D\bar{B} \text{ corresponds } \{A \rightarrow C, A \rightarrow D, D \rightarrow A, D \rightarrow B\}$$

$$A\bar{B} + A\bar{C} + A\bar{D} + D\bar{A} \text{ corresponds } \{A \rightarrow B, A \rightarrow C, A \rightarrow D, D \rightarrow A\}$$

147.3.2 Determining All Candidate Keys

Determine all candidate keys for relational scheme $R = \langle U, F \rangle$, where U is a set of attributes and F is a set of FDs on U . According to definition of candidate key, the following procedure for finding a key is correct obviously.

Add a new FD $U \rightarrow T$ in F , denote as F' , where T is a arbitrary attribute not in U , eliminate redundant attributes in the left hand of $U \rightarrow T$ to get $K \rightarrow T$, then K is just a candidate key.

In logical algebra the procedure is as: get corresponding logical function of F' , determine all prime implicants just as $K\bar{T}$, every K is a candidate key.

The procedure can be simplified. Assume $U - K = \{A_1, A_2, \dots, A_n\}$, the addition of minimum items of $K\bar{T}$ is

$$K\bar{T} = K\bar{T}(A_1A_2 \cdots A_n + A_1A_2 \cdots \bar{A}_n + \cdots + \bar{A}_1\bar{A}_2 \cdots \bar{A}_n)$$

Except for $KA_1A_2 \cdots A_n$, other items, from $KA_1A_2 \cdots \bar{A}_n$ to $\bar{K}\bar{A}_1\bar{A}_2 \cdots \bar{A}_n$, are all minimum items implied by corresponding logical function of F , so only need to add minimum item $KA_1A_2 \cdots A_n$ to corresponding logical function of F and determine all prime implicants that do not contain any complemented variables. Every such prime implicate correspond a candidate key.

Example (continued the previous example) Determine all candidate keys F

Answer Add minimum item $ABCD$ to corresponding logical function of F

$$Y = AB\bar{D} + A\bar{B} + D\bar{C} + C\bar{D}\bar{B} + CD\bar{A} + ABCD$$

All prime implicants of Y , without any complemented variables, are A, D , corresponding candidate keys are A, D , that are all the candidate keys of F .

147.3.3 Determine Closure of a Set of Attributes

If all prime implicants of corresponding LF for set of FDs F are obtained, it is easy to compute closure X_F^+ of any set of attributes X over F . The brief proof and algorithm are given below.

Let $X_F^+ = Y$, then for any attribute A in $Y, X \rightarrow A$ is implied by F . According to the conclusion in [3], $X\bar{A}$ is a implicate term in the FD map of F . If $X\bar{A}$ is not a prime implicate, it can be changed into a prime implicate $P\bar{A}$, where $P \subset X$. So the algorithm for computing X_F^+ is

- (1) $Y = X$
- (2) To exhibit FD map of F , and determine all prime implicants.
- (3) For each prime implicate $P\bar{A}$, if $P \subseteq X$ and $A \notin Y$, then $Y = Y \cup \{A\}$.
- (4) Return Y , it is X_F^+

We give a simple example to explain the above method.

Example $F = \{AC \rightarrow B, AD \rightarrow C, BD \rightarrow C, C \rightarrow A\}$ is a set of FDs, compute C_F^+, BD_F^+, AB_F^+

Answer corresponding LF of F is $Y = AC\bar{B} + AD\bar{C} + BDC\bar{C} + C\bar{A}$

All prime implicants of Y are $AD\bar{B}, AD\bar{C}, BD\bar{A}, BDC\bar{C}, C\bar{A}, C\bar{B}$

so $C_F^+ = \{A, B, C\}$

$$BD_F^+ = \{A, B, C, D\}$$

$$AB_F^+ = \{A, B\}$$

147.4 Conclusion

For a given set of FDs, first transform it into a corresponding logical function, then determine all prime implicates of the logical function. By use of these prime implicates, all equivalent minimum sets of FDs, all candidate keys, and closure of arbitrary sets of attributes over the set of FDs can easily be computed. These important problems in the theory of FDs are solved in a efficient and united manner.

This method can be extended to apply on set of data dependencies containing multivalued dependencies and FDs. This fully indicates its theoretical basis is consistent with deep part of theory of relational database. It is worthwhile to further research in this field.

References

1. Codd EF (1970) A relational model of data for large shared data banks [J]. *Commun ACM* 13(6):377–387
2. Maier D (1983) *The theory of relational databases* [M]. Computer Science Press Inc, Rockville
3. Maier D (1980) Minimum covers in the relational database model [J]. *J ACM* 27(4):664–674
4. Fagin R (1977) Functional dependencies in a relational database and propositional logic [J]. *IBM J Res Dev* 21(6):534–544
5. Sagiv Y, Delobel C, Parker DS et al (1981) An equivalence between relational database dependencies and a fragment of propositional logic [J]. *J Assoc Comput Mach* 28(3):435–453
6. Delobel C, Casey RG (1973) Decomposition of a data base and the theory of Boolean switching functions [J]. *IBM J Res Dev* 17(5):374–386
7. Russomanno DJ, Bonnell RD (1999) A pedagogical approach to database design via Karnaugh maps [J]. *IEEE Trans Educ* 42(4):261–269
8. Zhang Y (2009) A key point on the equivalence between functional dependencies and logic algebra [J]. *Int Conf Inf Eng Comput Sci V*:3657–3660

Chapter 148

Design History Knowledge Management System Based on Product Development Process Management and Its Implementation

Peisi Zhong, Jiandong Song, Mei Liu, Xueyi Li and Shuhui Ding

Abstract Design history knowledge (DHK) management is the key part of knowledge-based product development process management (PDPM) system. Product development and knowledge management including product development process, design history and domain knowledge are reviewed. The first step of DHK management is knowledge-based product development process modeling. A method for the knowledge-based multi-view process modeling is presented including process implementation, process monitoring, knowledge management and so on. The integrated framework of DHK management system and the hierarchical model of DHK are built. The method for acquisition and management of process history, design intent and domain knowledge is presented. The functional modules of knowledge-based PDPM system is described and the architecture of DHK management system based on PDPM is developed and used successfully during the lifecycle of a new product development in a company.

Keywords Product development process · Design history knowledge · Process history · Design intent · Domain knowledge · Knowledge management

P. Zhong (✉) · J. Song · M. Liu · X. Li · S. Ding
Advanced Manufacturing Technology Center, Shandong University of Science and
Technology, Qingdao 266510, People's Republic of China
e-mail: pszhong@163.com

J. Song
e-mail: jdsong07@163.com

M. Liu
e-mail: liumei.cn@163.com

X. Li
e-mail: xueyi_l@163.com

S. Ding
e-mail: shdingcn@163.com

148.1 Introduction

PDPM is one of the key enabled technologies for the implementation of concurrent engineering (CE) including process modeling, process implementation, process monitoring, process analysis and optimization, process improvement and reengineering, process simulation etc. In the most cases of product development process, the experience of the product development is acquired passively. Because designers and managers of the project are busy in developing the product, and no time to build the knowledge for product or improve the development process.

The product development process is made up of a series of processes and sub-processes and the sequence of activities done by many roles to finish a certain product development. The process model is based on activities. The process history records the status changing of activities and status changing of flow, which means the input and output of activities.

The main purpose of the study on DHK management is to explore the theory and methods on process management, history management and domain knowledge management during the product development process, that is, to integrate the DHK management into the lifecycle of the product development process including mainly the acquisition and usage of process history, design intent, domain knowledge etc.

It is necessary to develop an integrated supporting environment for knowledge management during the life cycle of product development including process history, design intent, domain knowledge and so on. The following is the theory, methods and implementation for DHK management and used in a company successfully.

148.2 Review

With the embedded study on CE, the cooperative product development, integrated supporting environment for concurrent design become the current hotspot [1, 2]. The integrated product development team for cooperative product development is made up of designers who are in different site and related by network, as if they are in one place. Thus it can reduce the developing cost and improve product quality to use fully the software and hardware resource of CAD/CAM and all kinds of knowledge and experience of correlative designers. How to acquire and reuse the knowledge during the development process becomes a research hotspot now. The research can be divided into the following aspects: product development process, process history, design intent, domain knowledge and integrated support environment.

Product Development Process. The foundation on research of product development process management is process modeling, process analysis, process improvement and process reengineering, process monitoring and conflict

management. As the deference of requirement and applied background, researchers present many kinds of process methods and technologies. Summing up all methods, there are IDEF method, structured analysis, petri-net modeling method, real-time structured analyzing and process modeling method, process programming method (including rule-based method) and systematic dynamic method [3]. In these methods, some of modeling methods are supported by corresponding tools such as SADT/IDEF0, petri-net and so on, others are in study and lack of supporting tools. In a word, all these methods and their supporting tools are focus on one or two views of many views of product development process, especially lack of the knowledge view.

PDPM can assort with product development activities, monitor and attemper product development process, build unified product process information model, set up the checking mechanism of product development process, establish the mechanism of constraint and conflict resolution among activities, provide decision support tool for product development process and design and so on [4].

Process History, Design Intent and Domain Knowledge. The research object on design history is mainly the domain knowledge gathered during product development including design decision knowledge and process decision knowledge. The knowledge for design decision includes knowledge of requirement of market and users, design standard, design principle, design modification, manufacturing, assembly, cost, quality and so on. The knowledge for process decision includes information on manpower and resource, knowledge of process planning, coordination and so on. All these knowledge reflects the history of product development and holds an important role in designing high-quality product.

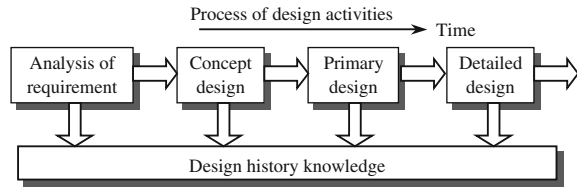
In the existing models for design history, data of product, process (including operation) and decision (including intent, principle and constraint) are the main contents. The whole process from the initially definition to the finally geometry and topology design can be backdated according to the decision, and the information of product, information of development process and their relationship can be retrospected according to the process.

It is difficult to integrate design history, process model and product structure [5]. Most research considers process development process as task flow, pays more attention to process itself and doesnot think over how to integrate process, domain knowledge and numerical product model.

148.3 Process Based Integrated Framework DHK Management Sytem

The PDPM system is used to coordinate activities in concurrent product development process, control and attemper the product development process, set up a union information model for product and process, mechanism of examine and approve for product development, mechanism of constraint and conflict coordination among activities, and provide toolkits for PDPM system.

Fig. 148.1 Relationship of DHK and development process



The DHK management system is based on the PDPM system. The DHK management system captures the design history by the module of process monitoring and organizes the design intent and domain knowledge according to product development process.

Basic Concepts. The DHK includes the following main parts:

Process history is the history of the design tasks, including the steps of process implementation, the statuses of the implementing condition, the statuses of the flows, the distribution of organization and resource, the schedule of the project, as well as the record of evaluation, decision and so on.

Design intent is the sum of information about the design method and the decision causation, including the status changing process of the design information, the design steps and scene which caused the above change, as well as design history information of design decision, the choice of design schemes, design intent and so on.

Domain knowledge is the sum of the design principle, design method and design experience which are exist in professional books, manuals, as well as in the brain of human.

Integrated Framework. Design history is the recorder and elucidation about the development process of the product design object. It incarnates all important information and knowledge in the process of product development life cycle. Figure 148.1 is the relationship of design history and development process.

In general, the model of design process includes concept design, primary design, detail design, process planning and so on. In these stages, the design intent changes from abstract to concrete form, and the design data generates and is perfected step by step. The interaction among the designer, design intent, design data, design tool, resource, and environment changes and forms the design pathway. The recorder to the thinking process of designer forms the design rationale. Figure 148.2 is the hierarchical model of DHK.

148.4 Acquisition and Management of Design History Knowledge

The design intent and domain knowledge can be acquired by knowledge acquisition tool and the correlative attribute of design is automatically acquired by the integrated framework, such as time, correlative activities, roles and so on.

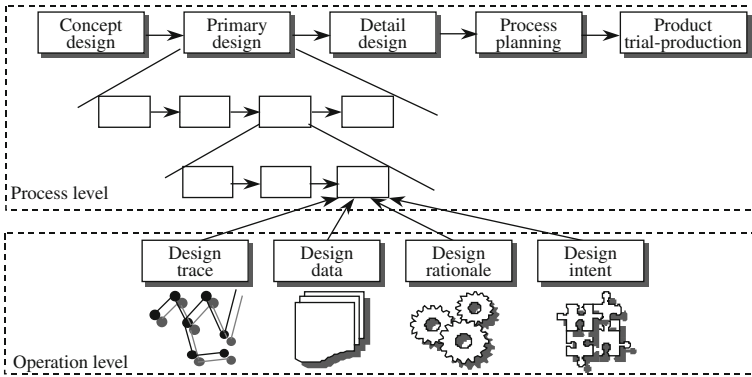
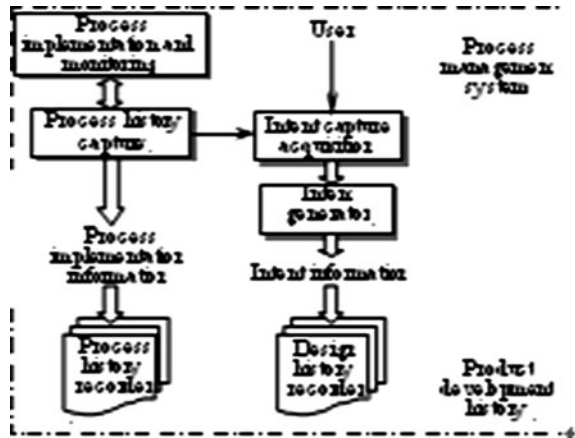


Fig. 148.2 Hierarchical model of DHK

Fig. 148.3 Acquisition of design history knowledge



The process history mainly uses automated acquisition supporting by computers except acquiring mechanism of process management similar to design intent. By integrating the process implementation and monitor system, all events in the development process can be captured. And the module of process history will capture the event automatically, and save it in the process history base, as shown in Fig. 148.3.

Process History. Process history is the record of the actual product development process according to the technology management. The content of process history includes three parts.

Part 1 is the process model of the special product development. It includes mainly process architecture, activities, descriptions of input and output flow, constraint condition and rules for process execution, status of flow, distribution of role and resource. The information above can be acquired when process is modeled.

Part 2 is the executing history of design tasks, including changes of development process, sequence and time of task execution, status change of process execution condition, as well as changes of activities and flow, dynamic usage of role and resource, reason of design change and iterative process, and dynamic calculation of workload. In addition, public message and special message are provided in the environment of concurrent product development process.

Part 3 includes coordination and decision of task or project and its principle, and content corresponding to the project, such as time and cost of designers training, method or technology involved in the project, the domain of developers, cost of evaluation, time of document writing.

All the three parts are acquired automatically during the process execution as shown in Fig. 148.3. The module of process history capture acquires the process implementation information from the module of process implementation and monitoring, and input all these information to process history base.

Design Intent. The capture of design intent is a semiautomatic process with part autonomic capability and participation of users. There are two input method, one is that the system activates the intent capture module through the process monitoring module according to the property and result of tasks, and requires developers to input related information. Another method is that the developer may activate the intent capture module at any time according to his requirement, as shown in Fig. 148.3.

Design intent uses the nature representation. It supports the non-formalized input tools such as text editor, notebook, electronic panel, e-mail and multimedia device. Thus the developer can describe his ideas more clearly and freely, not be constrained by rules.

Domain knowledge. The product design is a complex process and must be researched with the point of system engineering. Its components possess very strong independency. So the goal of product design can be decomposed into several sub-goals and each has stated independency. The relationship among sub-goals forms a public knowledge base (KB), and the knowledge for each sub-goal forms its own KB which can inherit the knowledge from the public KB.

The describing model of domain knowledge for concurrent design of complex product is shown in Fig. 148.4 in 3 dimensions—process, component and sub-knowledge base. Different design process, different component (or part) and different supporting tool for concurrent design has its own sub-knowledge base (Sub-KB).

148.5 Implementation

In the process-centered system, the DHK management system is attached to the PDPM system. Integrated to engineering design system, the knowledge system gets the design intent and history information of product with the help of PDM, and acquires the interactive information from designer to supporting tools in the

Fig. 148.4 Describing model of domain knowledge

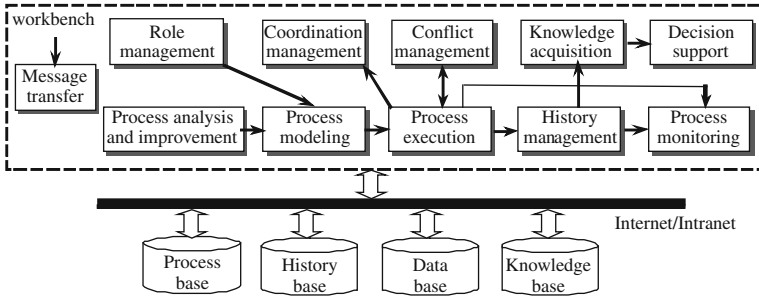
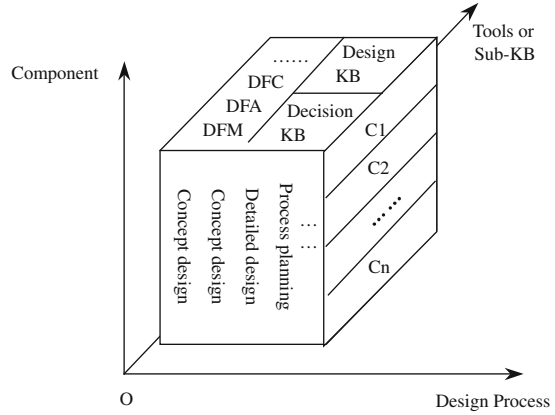


Fig. 148.5 Functional modules of PDPM system

environment of PDPM. It is the main part of PDPM to realize general DHK management.

The knowledge-based PDPM system consists of role management, process analysis and improvement, process modeling, process executing, process monitoring, history management, coordination management, conflict management, knowledge acquisition, decision support and so on which are integrated together on the workbench of the system. The functional modules of the PDPM system are shown in Fig. 148.5.

In the system, the sub-systems of process execution and process monitoring drive the developing process and update the process status. The sub-system of history captures and records the process history, design history, the status of using resource and so on. It can improve the leader’s insight to the product development and provide a foundation to improve the process later to record the actual process, decision and principle used. Figure 148.6 is the modules of the DHK management with the support of PDPM system.

The core module of the HKM system is the capture of process history, intent manager, history and knowledge browser, KB editor and DBMS/KBMS.

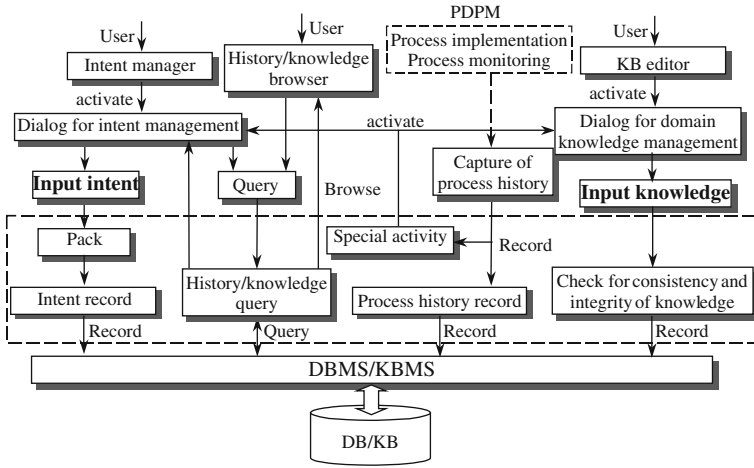


Fig. 148.6 Modules of DHK management system

The capture of process capture takes charge of monitoring the status changing of activities and flows, recording important process information, distinguishing the abnormality of process and providing the basis for the recording of design intent and domain knowledge. The intent manager takes charge of acquiring intent information from designers to integrates related information such as activity, time and role. The history and knowledge browser provides an interactive interface for designers to query related information of history, intent and knowledge. KB editor provides an interactive tool to acquire domain knowledge and assure the consistency and integrity of knowledge.

Supported by the DHK management system, designers can browser and query the history of design history including process history, design intent and domain knowledge. For example, designers can query the system by time to know the status of the product development at a certain time, query the system by activity to know when the activity started and ended or other important information about the activity, query the system by role to know the working status of one designer.

The system supports designers to backdate the design process. As the DHK management system records the history of the lifecycle of the product development, the actual status of key design, certain activity, all design process at that time can be replayed to know why and how the product is designed and direct the leader to evaluate accurately the whole product development process.

The integrated supporting environment of the PDDM system including the integrated knowledge management system is developed and has a B/C/S structure. It provides an ideal support environment for the integrated knowledge management and cooperative design with Internet/Intranet. The system has been used successfully during the life cycle of a new product development in a company and obtained satisfactory results.

148.6 Summary

The integrated knowledge management in lifecycle of product development process is one of key enabled technologies to implement CE. Using the system correctly, the knowledge can be saved and reused including the process history, design intent and domain knowledge in Enterprise. It is an effective tool to support the knowledge-based PDPM system and intelligent design.

Acknowledgments The research is supported by National Natural Science Foundation of China (No. 50875158), Science and Technology Development Program of Shandong Province (No. 2010GGX10408) and Applied Basic Research Program of Qingdao, China (No. 09-1-3-51-jch).

References

1. Jiang P, Shao X, Qiu H et al (2009) *Comput Ind* 60:416
2. Ming XG, Yan JQ, Wang XH et al (2008) *Comput Ind* 59:154
3. Jiang P, Shao X, Qiu H et al (2008) *Concurr Eng Res Appl* 16:114
4. Zhong P, Liu D, Liu M et al (2005) *LNCS* 3168:175
5. Barbosa CAM, Feijo B, Dreux M et al (2003) *Adv Eng Softw* 34:621

Chapter 149

The Application of Series Importance Points (SIP) Based Partition Method on Hydrological Data Processing

Haixiong Chen

Abstract China has accumulated a large amount of valuable hydrological data, and the descriptive physical variables can be categorized into various types of hydrological time series. Time-series data usually contains huge amounts of high-dimension data that being continuously updated, thus it is difficult to directly mine the original time series data. This chapter adopts a time series segmentation algorithm based on series importance point (SIP)—PLR_SIP, to approximately describe time series with line segments based on SIP. SIPs are used as the splitting point to reflect the main features of time series and reduce the dimensions of the time series data, thus minimizing the overall error.

Keywords Series importance point · Hydrological data · Time series · Time series segmentation algorithm

149.1 Introduction

China is a country with serious water shortage, but also frequently suffers from flood to drought. Therefore, there is a great demand for accurate hydrological forecasting to control flood and drought, plan, design and manage water projects

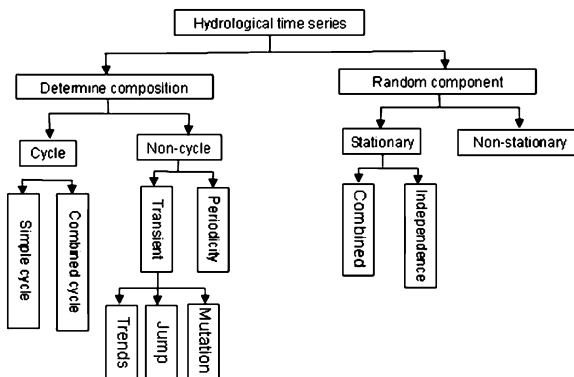
H. Chen (✉)
Zhejiang Water ConserVancy And Hydropower College,
Hangzhou, China
e-mail: chenhx@zjwchc.com

and rationally and efficiently use of water resources. Since 1980, during the high-tech waves driven by computer, the operational department of the Ministry of water resources has fully adopted high technology and invested heavily on hydrological information infrastructure, since then much beneficial research has been carried out for various aspects of hydrological science [1]. The most prominent achievement is the construction of national hydrological database system and the national real-time rainfall and flood control system. Up to 2004, the amount of the national hydrological data has accumulated over 10TB, and with the relevant weather, land resources information systems interconnected, China has formed a huge amount of relatively integrated information resources for scientific research, and provided a sound hydrological information environment. With these large quantities of complex data, the traditional hydrology theory and statistical methods would be incompetent to analyze because of insufficient computing capacity, storage capacity and lack of algorithms. The introduction of data mining theory and technology to design hydrological data mining theories, techniques and methods may provide new ideas and possible solutions to scientific hydrological problems research.

149.1.1 Hydrological Time Series

Hydrological database contains a lot of data, such as regularly river level records and water flow observation of river station and rainfall observations and so on. Hydrological phenomenon is known as time-varying, and the changing process is called hydrological processes. The hydrological data is the discrete record of the process. Therefore, according to the descriptive physical variables, the hydrological data can be classified as various types of hydrological time series data. In general, the hydrological time series are classified as the following type [2]: continuous state and discrete time, the observations from certain period (usually taken equidistant) is consisted of continuous changing phenomenon, and the average value (or total) for each period is also commonly used. Hydrological time series can be divided into two categories according to the types of variable: single variable time series or multivariate time series. Examples of single variable time series include the annual rainfall or water flow of a given station, or the annual or monthly water flow of a certain river basin; while multivariate time series include the quantity and quality sequences of a station, or the annual or monthly precipitation series of a river basin. At present, many hydrologists believe that the hydrological time series are generally composed of determined and random components. The determined component has specific physical concept and could be divided into periodic and non-periodic process, while the random component is caused by the irregular oscillation and random effects and could be divided into

Fig. 149.1 The structure of hydrological time series



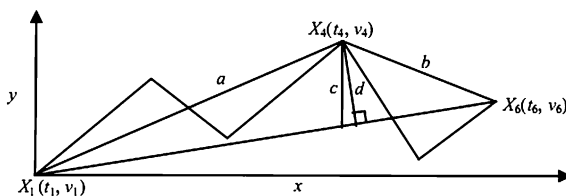
stationary and non-stationary process. Figure 149.1 describes the structure of hydrological time series.

149.1.2 Time Series Processing Technology

Time series data is usually high dimensional multivariate data. If certain operations such as query, sequence clustering and classification are carried out directly on the raw data, then the efficiency would be extremely low. The solution to the problem is to segment the long-time series into several relatively short and non-overlapping sub-series, and each sub-series are converted into a advanced data representation. The segmentation of time series has three distinct usages: (1) extracting the main features of time series data and removing the interference of details, thus keeping only the main form of time series to better reflect the characteristics of time series and improve the efficiency and accuracy of time series query, (2) transforming the high-dimensional time series into low-dimensional space, thus reducing the data dimension, and realizing data compression and improving query efficiency, (3) discovering the abnormal patterns within the time series data.

Keogh et al. referred all algorithms with time series as input and line as output as piecewise linear algorithm [3]. This method is consistent with human common experience, and usually has low-dimensional index structure and fast computing speed, so it is widely adopted. In recent years, domestic and foreign scholars has carried out in-depth study on the piecewise linear algorithm and put forward many new algorithms after absorbing the advantages of other methods and techniques. At present, the PLR algorithm can be divided into: (1) piecewise determined by the

Fig. 149.2 Three distances from data point X_4 to X_1 and X_6



fitting error: straight line is used to fit the original time series, until it reaches the predetermined error threshold. Such as the sliding window PLR_SW [4], top-down PLR_TD [5], bottom-up PLR_BU [6] and so on. These algorithms only focused on the minimum local error, and in the fitting process, some of the key features of the original data are smoothed out, thus the overall sequence variation is much ignored. (2) Segment determined by special points: methods include landmark model, edge point segmentation method, and the piecewise importance point model. Through a number of special points, such as local extreme points, edge points or importance points, these methods segment the time series, and avoid the loss of important information of straight-line fitting. But these available methods are relatively complicated, and don't take the overall error into consideration in the process, and the effectiveness of dimension reduction is inferior to the fitting method.

In this chapter, series importance point (SIP) based piecewise linear algorithm PLR_SIP is proposed. When choosing the piecewise point, this method takes both the overall characteristics and the global minimum error of the time series into account, resulting in higher algorithm efficiency.

149.2 Distance of Time Series Data

Time series is an ordered set of elements composed of recording time and value, denoted as: $X = \langle x_1 = (t_1, v_1), x_2 = (t_2, v_2), \dots, x_i = (t_i, v_i), \dots, x_n = (t_n, v_n) \rangle$, where element $x_i = (t_i, v_i)$ denotes the record value v_i at the time t_i of the time series. Human vision usually processes the smooth curve as several straight lines, and the most impressive point s are the sequence of extreme points. Extreme points of the time series data have different effects on the shape. Some extreme points exert great influences, while some points have barely any impacts on the overall shape. Pratt et al. [7] reviewed three point-to-end distance measurement methods: Euclidean distance, orthogonal distance, vertical distance. Figure 149.2 shows the three distances from data point X_4 to X_1 and X_6 , where $a + b$ is the Euclidean distance; c is the vertical distance; and d is the orthogonal distance.

Fig. 149.3 The comparison of vertical and orthogonal distance

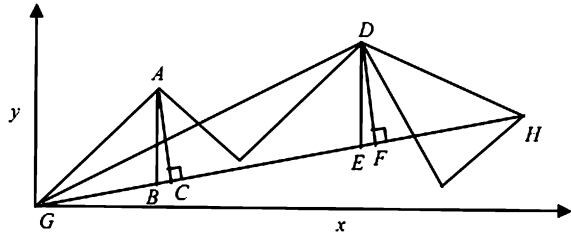
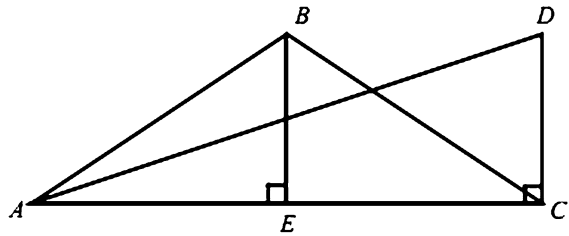


Fig. 149.4 The comparison of the Euclidean distance, orthogonal distance metric



Euclidean distance:

$$D = a + b = \sqrt{(t_4 - t_1)^2 + (v_4 - v_1)^2} + \sqrt{(t_4 - t_6)^2 + (v_4 - v_6)^2} \tag{149.1}$$

Vertical distance:

$$D = c = \left| (v_1 + (v_6 - v_1)) \times \frac{t_4 - t_1}{t_6 - t_1} - v_4 \right| \tag{149.2}$$

Orthogonal distance:

$$s = \frac{v_6 - v_1}{t_6 - t_1} \tag{149.3}$$

$$tc = \frac{t_4 + s \times v_4 + s \times v_6 - s^2 \times t_6}{1 + s^2} \tag{149.4}$$

$$vc = s \times tc - s \times t_6 - s^2 \times t_6 \tag{149.5}$$

$$D = d = \sqrt{(tc - t_4)^2 + (vc - v_4)^2} \tag{149.6}$$

These three distance measurements focus on different aspects, Fu Tak-chung [8]. With experimental data, it could be demonstrated that SIPs generated from the vertical and orthogonal distance are of the same order, but SIPs generated from Euclidean distance have different sequence order. Vertical distance and Orthogonal distance are shown in Fig. 149.3. The point-to-end orthogonal distance from

A to GH is $|AC|$ and the vertical distance is $|AB|$, while the orthogonal distance from D to GH is $|DF|$ and the vertical distance is $|DE|$. Since the triangle ABC and triangle DEF are apparently similar, then AB and DE, AC and DF are scaled proportionally. In the time zone $[G, H]$, if the orthogonal distance $|DF|$ from D to GH is the maximum orthogonal distance of all points, then it can be concluded that the vertical distance $|DE|$ from D to GH is also the maximum vertical distance of all points. Therefore, SIPs sequence generated from the vertical distance and the orthogonal distance is identical.

The comparison of the Euclidean distance and the orthogonal distance is shown in Fig. 149.4, where $|BE|$ is the orthogonal distance from B to AC, and $|DC|$ is the orthogonal distance from D to AC. If $|BE| = |DC|$, it can be concluded that $|AB| + |BC| > |AD| + |DC|$. In the time region $[A, C]$, if two points have the same orthogonal distance to AC doesn't result in the conclusion that the accumulated distance of these two points to the endpoints of $[A, C]$ are equal. Therefore, SIPs sequence generated from the Euclidean distance and the orthogonal distance may be inconsistent.

Definition: series importance point (SIP) is point that has the maximum error within the zone and the furthest distance from the zone ends.

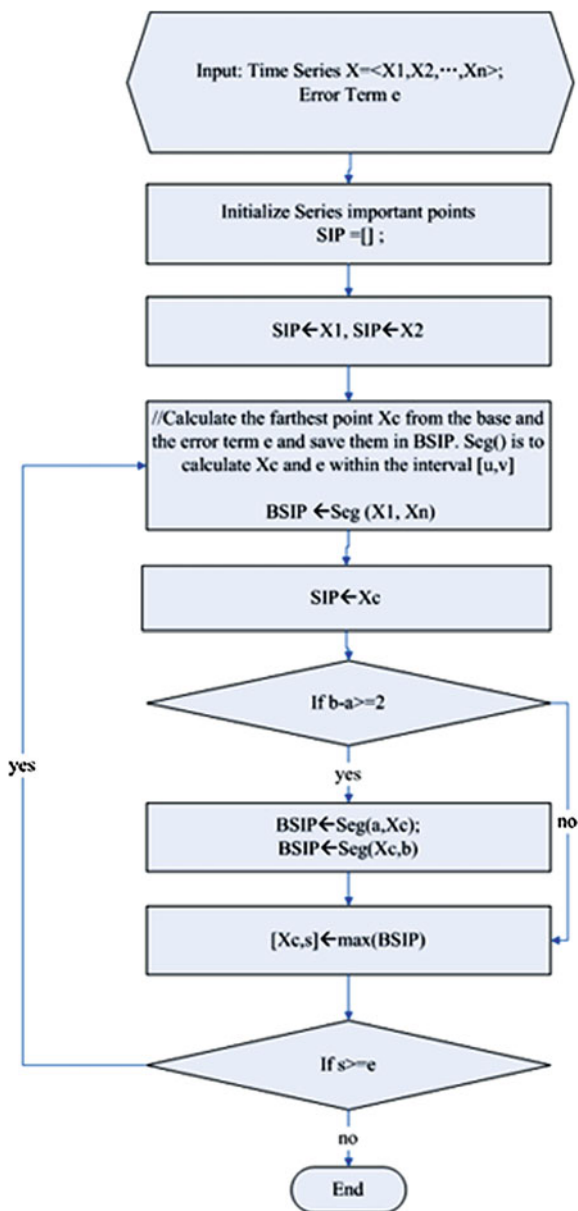
When segmenting time series data, previous researchers minimize the error when each piecewise line is fitted with the original data and commonly orthogonal distance from point to line is adopted to calculate the fitted error. The above has proved that SIPs generated from the vertical distance and orthogonal distance has the same order, and the way to calculate vertical distance is much simple [9].

149.3 Segmentation Algorithm Based on the Series Important Point

From the definition of SIP, there are two rules to be followed when choosing the SIPs of time series:

- (1) The point should be the furthest point from the adjacent SIPs;
- (2) Before the point is added to the SIPs, global maximum error of the zone should be taken into account.

The selection algorithm of SIPs sequence is described as follows:



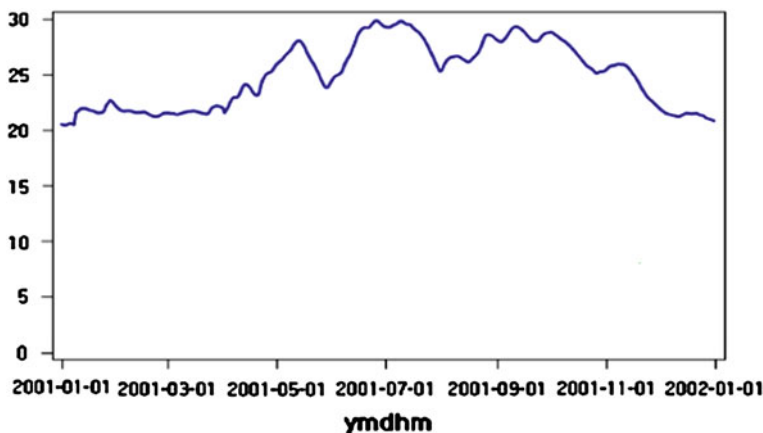


Fig. 149.5 The water level line of Chenglingji station of Yangtze river

149.4 The Experiment on Chenglingji's Water Level of Yangtze River

The experimental data is the water level data from the upstream Chenglingji (station code: 65071) of the Yangtze River, spanning from 1997 to 2005. Since there are few tributaries between stations, the similarity of the lines of water level is quite high. Figure 149.5 depicts the water level line of Chenglingji stations in 2001, where the horizontal axis represents the date (in days) and the vertical axis represents the value of water level (in meters). Flood alert level is a very important basis for flood forecasting. According to the annual Yangtze River Hydrological Report, the alert level of Chenglingji station is 32.5 meters. When the water level is relatively low, the impact of upstream on downstream is not so obvious. This study selects the values over the threshold of 85% of the alert level, namely the time series water level higher than 19.625 meters of Chenglingji station. When the input error is 6, the time series is divided into 72 segments, the compression ratio is 365/72, and the time series data before and after segmentation are shown in Fig. 149.6.

For the same time series data, PLR_SIP, PLR_IP, PLR_TD and PLR_SW linear segmentation method are adopted to segment the water level line respectively. The experimental results show that PLR_SIP splitting well reflects the overall time series features; PLR_IP emphasizes too much on the importance of local points without considering the problem of global optimization; PLR_TD and PLR_SW segmentation method smooth out some important information in the fitting process and fail to describe the abnormal data within the time series data. The time complexity of algorithm PLR_SIP and PLR_IP is close to $O(l \times \log^n)$, the time complexity of PLR_SW is $O(l \times n)$, and that of PLR_TD is $O(n \times n)$. The CPU

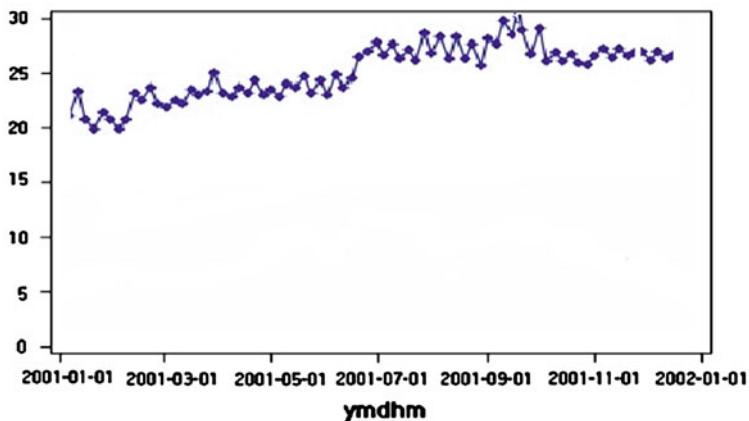


Fig. 149.6 The water level line of Chenglingji station of Yangtze river after segmentation

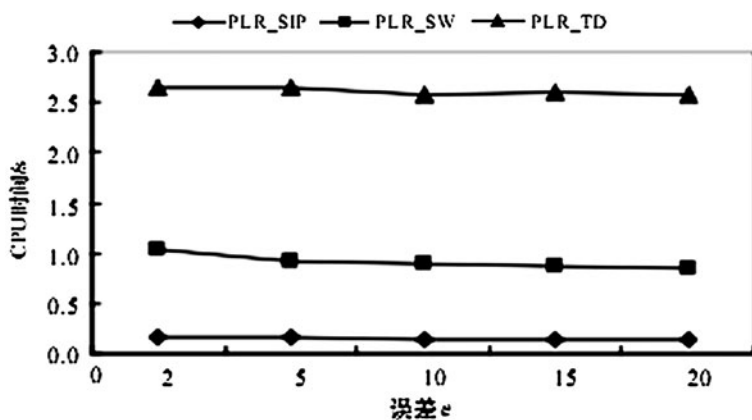


Fig. 149.7 The comparison of CPU time of these three segmentation algorithms

time consumption of these segmentation algorithms with different input errors are shown in Fig. 149.7.

149.5 Conclusion

Compared with other PLR segmentation algorithm, the proposed algorithm effectively and efficiently segments the time series, and well describes the overall characteristics of time series data. In the SIPs selection process, both the local and

global optimization problems are taken into account, and the overall shape of the original time series is preserved, and piecewise error is minimized. This algorithm provides researchers a basis to describe time series data, and furthermore to discover time series pattern.

The experimental data of the article is based on the units of days with large time granularity, while generally flood spreads quickly. According to the collection of hydrological data, usually the hydrological data (water level, flow, etc.) are collected daily or every 12 h, but when a big storm flood occurs, the need for flood control demands the hydrological data to be collected in real time or hourly. So a finer granularity of water level data is practically significant to estimate the flood transmission. Obviously, unless an effective segmentation algorithm is adopted, the calculation would be very huge.

The study supported by the Foundation of Zhejiang College of Water Conservancy and Hydropower, grant No. 200715099.

References

1. Baldonado M, Chang C-CK, Gravano L, Paepcke A (1997) The Stanford digital library metadata architecture. *Int J Digit Libr* 1:108–121
2. Bruce KB, Cardelli L, Pierce BC (1997) Comparing object encodings. In: Abadi M, Ito T (eds) *Theoretical aspects of computer software. Lecture notes in computer science*, Vol 1281. Springer, Berlin, pp 415–438
3. van Leeuwen J (ed) (1995) *Computer science today. Recent trends and developments. Lecture notes in computer science*, Vol 1000. Springer, Berlin
4. Wu S-Y (2007) *Research and application of pattern mining on hydrological time series [D]*. Master's degree thesis, Hohai University
5. Liu Deping J (1991) *Hydrological time series models and forecasting methods [M]*. Hohai University Press, Nanjing
6. Keogh E, Chakrabarti K, Pazzani M et al (2001) Dimensionality reduction for fast similarity search in large time series databases [J]. *J Knowl Inf Syst* 3(3):263–286
7. Qu Y, Wang C (1998) Supporting fast search in time Series for movement patterns in multiples scales [C]. In: *Proceedings Of the 7th ACM CIKM International conference on information and knowledge management*, Bethesda
8. Keogh E, Pazzani M (1998) An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback [C]. In: *Proceedings of the 4th International conference on knowledge discovery and data mining*, New York
9. Park S, Lee D (1990) Fast retrieval of similar subsequences in long sequence databases [C]. In: *Proceedings of the 3rd IEEE knowledge and data engineering exchange workshop*, Chicago

Chapter 150

A Novel PSO k-Modes Algorithm for Clustering Categorical Data

Lu Mei and Zhao Xiang-Jun

Abstract The k-modes is a classical clustering algorithms for categorical data set. Its principle is simple, but it is easy to converge to a local optimum. PSO (Particle Swarm Optimization) algorithm is an effective tool for optimization, so we proposal a novel clustering algorithm based on PSO for categorical data set. PSO usually used to solve continuous optimization problems, but the categorical data are non-continuous. This paper presents a novel k-p-modes algorithm to overcome this problem. Results show the method is effective.

Keywords PSO · Global optimization · Clustering algorithm

150.1 Introduction

Cluster analysis is an important tool for data mining, which can divide the known data set into meaningful or useful clusters [1]. The K-means algorithm [2–5] is well known for its efficiency in clustering large data sets. However, it may only stop at local optima of the optimization problem and working only on numeric data limits its use.

The categorical data are discrete, sometimes non-numeric. How to cluster the large categorical data sets? Huang proposed the algorithm for clustering large data

L. Mei (✉) · Z. Xiang-Jun
School of Computer Science and Technology, Xuzhou Normal University,
Xuzhou, 221116 China
e-mail: lumei@xznu.edu.cn

Z. Xiang-Jun
e-mail: xjzhao@xznu.edu.cn

sets with categorical values in 1998 [6]. It is similar to the k-means algorithm except for calculating the similarity of two elements. The k-modes algorithm also stops at locally optimal solutions. Many literatures have tried to find a globally optimal solution. For example, the fuzzy k-modes algorithm [7], the tabu searching method [8]; the genetic fuzzy k-modes algorithm [9].

In PSO, the system searches for optima by updating generations. Comparing with GA, PSO is easy to implement and has fewer parameters to adjust. PSO simulates the behaviors of bird flocking. It is initialized with a group of random particles and then searches for optima by updating generations. Each particle is updated according to the two “best” values. The first one is the position vector with its best fitness finish ever, and is denoted by p_{id} . The second one is the best fitness particle of this iteration, and is denoted by p_{gd} . After finding the two best values, the particle updates its velocity and positions with the following two equations:

$$v_{id}^{k+1} = w \times v_{id}^k + c_1 \times rand(.) \times (p_{id} - x_{id}^k) + c_2 \times rand(.) \times (p_{gd} - x_{id}^k) \quad (150.1)$$

$$x_{id}^{k+1} = x_{id}^k + v_{id}^{k+1} \quad (150.2)$$

where x_{id}^k is the value on the d th dimension of the i th particle, v_{id}^k is the velocity on the d th dimension of the i th particle, w is the inertia weight which is a parameter to control the impact of the previous velocities on the current velocity, c_1 , c_2 are positive constants, $rand(.)$ is a random number between 0 and 1.

Many literatures attempted to combine particle swarm optimization with the clustering algorithm. For example, the fuzzy clustering algorithm based on PSO [10–12]. They all use PSO to solve the continuous problems. Yin BO, et al. propose fuzzy K-Prototypes clustering based on particle swarm optimization. Similar with k-means, the method essentially uses a continuous data processing method [13].

This paper firstly converts the categorical data numerical into a high-dimension continuous data. Then we transform the k-modes algorithm into the one fitting for processing high-dimensional data. Then this paper proposed a novel PSO k-modes algorithm, which integrates the PSO algorithm and the k-Modes algorithm to find the globally optimal solution of the optimization problem. The outline of the paper is as follows. In Sect. 2, the k-modes algorithm is briefly reviewed. In Sect. 3, the novel PSO k-modes algorithm for clustering categorical data is proposed. In Sect. 4, the experimental results are presented to illustrate the effectiveness of our new algorithm. Finally, some concluding remarks are given in Sect. 5.

150.2 K-Modes Algorithm

Supposed that $D = \{X_1, X_2, \dots, X_n\}$ is a categorical data set with n objects, each of which is described by m categorical attributes $\{A_1, A_2, \dots, A_m\}$. $A_j (1 \leq j \leq m)$ has n_j categories, $DOM(A_j) = \{a_j^{(1)}, a_j^{(2)}, \dots, a_j^{(n_j)}\}$. n_j is the number of the value of A_j . The data object $X_i = [x_{i1}, x_{i2}, \dots, x_{im}]$, where $x_{ij} \in DOM(A_j), 1 \leq i \leq n$

Definition 1 The dissimilarity measure between X_i and X_j is defined by the total mismatches of the corresponding attribute categories of the two objects. Formally,

$$d(X_i, X_j) \equiv \sum_{l=1}^m \delta(x_{il}, x_{jl}) \quad 1 \leq i, j \leq n \tag{150.3}$$

where

$$\delta(x_j, y_j) = \begin{cases} 0, & x_j = y_j; \\ 1, & x_j \neq y_j. \end{cases} \tag{150.4}$$

Definition 2 Let the k cluster centers be represented by $Z = \{Z_1, Z_2, \dots, Z_k\}$, $Z_l = [z_{l,1}, z_{l,2}, \dots, z_{l,m}]$, $(1 \leq l \leq k)$ The objective of the k -modes is to minimize

$$F(W, Z) = \sum_{l=1}^k \sum_{i=1}^n w_{li} d(Z_l, X_i) \quad 1 \leq l \leq k, 1 \leq i \leq n, \tag{150.5}$$

subject to

$$w_{li} \in \{0, 1\}, 1 \leq i \leq n, 1 \leq l \leq k, \tag{150.6}$$

$$\sum_{i=1}^n w_{li} = 1, 1 \leq l \leq k, \tag{150.7}$$

where $d(., .)$ is defined in Eq. 150.5, W is the $k \times n$ partition matrix, $w_{li} = 1$ means the object X_i is a member of the cluster which center is Z_l .

Theorem 1 The quantity $F(W, Z)$ defined in Eq. 150.5 is minimized if and only if

$$z_{l,j} = a_j^{(r)} \in DOM(A_j), \tag{150.8}$$

where

$$\left| \left\{ w_{li} | x_{i,j} = a_j^{(r)}, w_{li} = 1 \right\} \right| \geq \left| \left\{ w_{li} | x_{i,j} = a_j^{(t)}, w_{li} = 1 \right\} \right| \quad 1 \leq t \leq n_j, 1 \leq j \leq k. \tag{150.9}$$

Theorem 2 Let $Z = \{Z_1, Z_2, \dots, Z_k\}$ be fixed, then the partition matrix W which minimizes the quantity $F(W, Z)$ defined in Eq. 150.5 subject to (150.5) is given by

$$w_{li} = \begin{cases} 1, & \text{if } d(Z_l, X_i) \leq d(Z_h, X_i), 1 \leq l, h \leq k; \\ 0, & \text{if others.} \end{cases} \tag{150.10}$$

Based on the two theorems described above, the k-Modes algorithm can be implemented recursively, described as follows.

- Step 1 Select k initial modes, one for each cluster.
- Step 2 Allocate an object to the cluster whose mode is the nearest to it according to (150.3). Update the mode of the cluster after each allocation according to Theorem 1.
- Step 3 After all objects have been allocated to clusters, retest the dissimilarity of objects against the current modes. If an object is found such that its nearest mode belongs to another cluster rather than its current one, reallocate the object to that cluster and update the modes of both clusters (according to Theorem 2).
- Step 4 Repeat step 3 until no object has changed clusters after a full cycle test of the whole data set.

150.3 The K-P-Modes Algorithm

150.3.1 The Data Pre-Processing

Firstly, all of the categorical data are mapped into continuous natural numbers.

Secondly, for each data object $X_i = [x_{i1}, x_{i2}, \dots, x_{im}]$ in data set, whose attribute value x_{ij} ($1 \leq i \leq n$) is now not only a single number, but an expanded set on the possible values, i.e., for the j th value x_{ij} , a set of n_j on $[0,1]$. Besides, the default value in the matrix should be added by 0. For example, an object $X_i = (2, 1, 1)$ would be now expanded as follows,

$$X_3 = (2, 1, 1) \xrightarrow{\text{expanded}} X_3 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & \\ 0 & 1 & 0 \end{bmatrix} \xrightarrow{\text{after added by 0}} X_3 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Generally, let $n_{\max} = \max(n_j) (1 \leq j \leq m)$, after pre-processing, the data $X_i = [x_{i1}, x_{i2}, \dots, x_{im}]$ is transformed into a $m \times n_{\max}$ matrix, shown as follows,

$$X_i = \begin{bmatrix} x_{i11} & x_{i12} & \cdots & x_{i1n_{\max}} \\ x_{i21} & x_{i22} & \cdots & x_{i2n_{\max}} \\ \vdots & & & \\ x_{im1} & x_{im2} & \cdots & x_{imn_{\max}} \end{bmatrix} \tag{150.11}$$

The excellency of pre-processing we can now make the data more “continuously”, for intermediate data object are defined, as soon as some coefficients are between 0 and 1. This idea is the same as the natural number are expanded to integer, etc.

Since each data object change, the dissimilarity measure between X_i and Y_i should be changed accordingly. The new form is as follows,

$$d(X_i, X_j) \equiv \sum_{l=1}^m \delta(x_{il}, x_{jl}) \quad 1 \leq i, j \leq n, \tag{150.12}$$

where

$$\delta(x_{it}, x_{jt}) = \sqrt{\sum_{k=1}^{n_{\max}} (x_{it} - x_{jt})^2} / \sqrt{2} \tag{150.13}$$

In (150.12), to divide $\sqrt{2}$ is to make our algorithm be compatible with the traditional k-modes algorithm. $F(W, Z)$ is the same as that of the traditional k-modes [as (150.5)] except for the dissimilarity measure, which is defined as (150.12).

150.3.2 The Encoding and Fitting Function

Each particle maintains a matrix $Z = (Z_1, Z_2, \dots, Z_k)$, where Z_i represents the Z_i cluster centroid vector and k is the cluster number. We use $f(Z)$ as the fitness value to evaluate the solution represented by each particle defined as (150.14).

$$f(Z) = 1 / (F(W, Z) + 1) \tag{150.14}$$

where, $F(W, Z)$ is defined as Eq. 150.5. The smaller $F(W, Z)$ is, the higher the fitness value is, and the better the cluster is.

150.3.3 K-p-Modes Algorithm

In PSO, each particle is made up of the k cluster centroids, whose form has changed as shown in (150.11), so the form of velocity should be changed accordingly. It has the same representation than a position: a list of sets. The only difference is that the coefficients can have any real value.

We apply “standardized” to make sure that all coefficients in a expanded position fall in (0, 1), and that the sum of all values in each line is 1. The example below shows how the operator works (the coefficients are between 0 and 1).

$$\begin{bmatrix} 0 & 2 & 4 \\ 0.1 & 0.3 & 0.2 \\ 4 & 1 & 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 0 & 0.33 & 0.67 \\ 0.17 & 0.5 & 0.33 \\ 0.8 & 0.2 & 0 \end{bmatrix}$$

The K-p-modes algorithm is as follows.

Step 1 Initialize

- (1) Initialize the parameters. Give the cluster number k , particle size s , c_1 and c_2 , w , particles' maximal velocity V_{\max} and the maximum iterative error.
- (2) Initialize the k cluster centroids. Choose k data objects randomly as cluster centroids to generate one particle. Repeat it for s times to generate s particles. Then encode each particle and get its position matrix.
- (3) Initialize pbest and gbest. Each particle's personal best position is initialized the same as its current position. Then calculate the global best position according to fitness, which is the highest fitness of all pbests.

Step 2 Generate partition matrix W . Calculate the distance between every particle and each centroids of every particle according Eq. 150.12, then generate the partition matrix W according to Eq. 150.10, which allocates each data object to the nearest cluster. Repeat it for s times according to the centroids provided by s particles.

Step 3 Calculate each particle's fitness.

- (1) Calculate $F(W, Z)$ according to the s partition matrix W and Eq. 150.5.
- (2) Calculate the s fitness according to Eq. 150.14.

Step 4 Update pbest and gbest.

- (1) For each particle, set current value as the new pbest if its current fitness value is better than the best fitness value (pbest) in history.
- (2) Choose the particle with the best fitness value of all the particles as the gbest.

Step 5 Update the particle, which means maybe generate new centroids. Update each particle's position and velocity according to Eqs. 150.1, 150.2.

Notation: (a) Particles' velocities on each dimension are clamped to a maximum velocity V_{\max} . If the sum of accelerations would cause the velocity on that dimension to exceed V_{\max} , which is a parameter specified by the user, then the velocity on that dimension is limited to V_{\max} .

(b) The results do not need to rounding.

Step 6 Termination. If the iteration reaches the maximum iterations or the best fitness reaches the designated value, then stop the program, else go to step 2.

Table 1 Comparison between the results of k-p-modes and GF k-modes

	Max F	Ave. F	Best γ	Ave. γ	The number of times $\gamma = 1$
GF k-modes	193.832904	209.081562	1.000000	0.771319	24
K-P-modes	120.31	126.5	1.000000	0.9815	82

150.4 Experiments

We used the corrected Rand index to assess the recovery of the underlying cluster structure. Let $P = [C_1, C_2, \dots, C_{k_1}]$ and $P' = [C'_1, C'_2, \dots, C'_{k_2}]$ be two clustering results of D. $n_{ij} = |C_i \cap C'_j|$, then the corrected rand index is defined as

$$\gamma = \left(\binom{n}{2} \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \binom{n_{ij}}{2} - \sum_{i=1}^{k_1} \binom{|C_i|}{2} \sum_{j=1}^{k_2} \binom{|C'_j|}{2} \right) / \left(\frac{1}{2} \binom{n}{2} \left[\sum_{i=1}^{k_1} \binom{|C_i|}{2} + \sum_{j=1}^{k_2} \binom{|C'_j|}{2} \right] \sum_{i=1}^{k_1} \binom{|C_i|}{2} \sum_{j=1}^{k_2} \binom{|C'_j|}{2} \right). \tag{150.15}$$

The corrected rand index γ ranges from 0 when the two clustering have no similarities, to 1 when the two clustering are identical. Because we know the true clustering of the data set, the true clustering and the resulting clustering are used to calculate γ .

Our data set is the well-known soybean data set [14]. The data set has 47 records, each is described by 35 attributes. Each record is labeled as one of the four diseases: diaporthe stem rot, charcoal rot, rhizoctonia root rot and phytophthora rot. Except for the phytophthora rot which has 17 instances, all other diseases have 10 instances each. Since there are 14 attributes that have only one category, we only selected other 21 attributes for the purpose of clustering.

Results are shown in Table 150.1. We specify $k=4, s=20, c_1 = c_2 = 2, w = 0.5$ and $\max = 200$. After running a hundred times, the misclassification rate is 1.91 using our algorithm, and 3.76 using genetic fuzzy k-modes algorithm for clustering categorical data (GF k-modes).

Compared the value of $F(W, Z)$, Ave. γ , The number of times $\gamma = 1$, we know, the one by our algorithm is better than the one by GF-k-modes. And the misclassification rate by ours is much lower.

150.5 Summary

In this chapter, we integrate the PSO algorithm and the k-modes algorithm to find the global optimization method. Thinking of that the PSO is often used to solve the continuous problem, while the categorical data are discrete, we integrate the two by expanding the categorical data into numerical and continuous data. Two

experiments show that the k-p-modes is promising and effective. It also points out a novel method to resolve the problem like this.

References

1. Tan P-N, Steinbach M, Kumar V (2006) Introduction to data mining [M]. The People's Post and Telecommunication Press, Beijing
2. Anderberg MR (1973) Cluster analysis for applications. Academic Press, New York
3. Ball GH, Hall DJ (1967) A clustering technique for summarizing multivariate data. *Behav Sci* 12:153–155
4. Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice-Hall, Englewood Cliffs
5. MacQueen JB (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th symposium mathematical statistics and probability, Berkeley, 1967, vol 1, no. AD 669871, pp 281–297
6. Huang Z (1998) Extensions to the k-means algorithm for clustering large data sets with categorical values [J]. *Data Min Know Discov* 2(3):283–304
7. Huang Z, Ng M (1997) A fuzzy k-modes algorithm for clustering categorical data [J]. *IEEE Trans Fuzzy Syst* 7(4):446–452
8. Ng M, Wong J (2002) Clustering categorical data sets using tabu search techniques [J]. *Pattern Recogn* 35(12):2783–2790
9. Gan G, Wu J, Yang Z (2009) A genetic fuzzy k-modes algorithm for clustering categorical data. *Expert Syst Appl* 36:1615–1620
10. Xu L, Zhang F (2006) Fuzzy clustering algorithm based on PSO. *Comput Eng Des* 27(21):4128–4129
11. Liu J, Han L, Hou L (2005) Cluster analysis based on particle swarm optimization algorithm. *Syst Eng-theory Pract* 6:54–58 (in Chinese)
12. Feng Z, Yan M, Zhang Z (2006) A PSO-based fuzzy clustering algorithm. *Comput Eng Appl* 27:150–151, 165 (in Chinese)
13. Yin B, Songhua H (2008) Fuzzy K-prototypes clustering based on particle swarm optimization. *Comput Eng Des* 29(11):2883–2885 (in Chinese)
14. Blake C, Merz C (1998) UCI repository of machine learning databases. <http://www.ics.uci.edu/mllearn/MLRepository.html>

Part XIV
E-Commerce and E-Government

Chapter 151

An Analysis of Influential Factors of Human Resource Allocation in Local Taxation System and the Modeling Approaches

Peiling Chen and Tian Yi

Abstract There is a talent shortage in local taxation system and an urgent need for talents has been proposed. In order to relieve the supply and demand conflicts and satisfy the needs of local bureaus, they mainly adopt the form of administrative planning, and then allocate human resource in accordance with the number and the area of households managed by per tax staff, which receive little effects but instead, worsen the mutual conflicts. What factors are exactly influencing the human resource allocation in local taxation system? How much do they impact? Taking the human resource allocation of city N as an example, this thesis further analyzes the factors such as working difficulty, working intensity and the staff quality by means of interview, questionnaire and Delphi method, which can help to determine the key factors that influence the allocation and the corresponding quantitative index, put forward the thinking of constructing human resource allocation model, and moreover, offer some suggestions, promoting a scientific human resource allocation of the local taxation system.

Keywords Local taxation system · Human resource allocation · Working difficulty · Working intensity · Delphi method

P. Chen (✉)
Teaching Affair Office, Hohai University, Nanjing, China
e-mail: ndrains@qq.com

T. Yi
Human Resource Department, Jiangsu Broadcasting Corporation,
Nanjing, China

151.1 Introduction

Tax is the main income source of a country, while the tax officials are fortune collectors of a nation. The recruitment ditch of local taxation system mainly depends on demobilized military officers and civil service examination, and the present staff largely consist of people recruited from the public and demobilized military officers in the 1980s and those graduates who entered into the system through civil service examination after 2000. In these years, since the national economy and GDP rapidly grows, the total amount of taxation has been greatly promoted, and with the establishment of new policies, the tax department's work amount and work complexity also gradually increase, which calls for a larger demand of tax workers. As an administrative department of the country, local taxation system is always restricted and motivated by a certain national personnel system during the process of dynamic human resource allocation [1]. Every year, the overall number of personnel is strictly limited by quotas, and with limited supply resources, the allocation always faces difficulties. Every local taxation bureau is confronted with heavy tasks, asking for new recruits desperately, while the total personnel number of every year has failed to meet their demands, thus the call is very urgent.

At present, local taxation system mainly allocates talents by administrative planning, decides how many new staff each bureau is supposed to get according to the number and the area of households managed by per tax staff. This kind of allocation is relatively thoughtless, giving no considerations to the influences caused by working intensity factors such as the number of enterprise tax payer and individual tax payer, nor taking account of impacts by working difficulty factors such as organization income task, prolongation declaration rate and unfinished net declaration rate, which not only fails to realize a scientific talents allocation and assign personnel to bureaus in real needs, but also makes the mutual conflicts worse. Therefore the pressing tasks that lie in the local collection are providing personnel guarantee to a successful tax collection, allocation talents scientifically, and deploying limited personnel to the department that needs them most.

The definition of human resource allocation in local taxation system is assumed to be an administrative activity that in a certain period of time, with the hope of fulfilling strategic goals, local taxation system allocate and reallocate human resource factor dynamically in various departments (mostly in local sub-bureaus), so as to make sure that the number and the structure of human resources can have an organic collocation and combination with other factors of production in the system. The influential factors are defined as those micro-environmental factors that affect human resource allocation in local taxation system, including working intensity, working difficulty, staff quality and so on [2]. What factors do have an impact on the human resource allocation? How deep have they penetrated? In search for the answers, this thesis takes local taxation system of city N as an example, carries out a positive analysis on the influential micro-environmental factors by means of interview, questionnaire, Delphi method, then on the basis of

those analytical results, it provides relevant suggestions on human resource allocation in local taxation system.

151.2 Research Process and Research Methods

151.2.1 Make a List of Influential Factors

In order to know about all kinds of micro-factors that affect human resource allocation in local taxation system, we spend two months in communicating and interviewing with some tax officials in city N. Step one: collect basic information, analyze relevant paper materials, and talk with local tax officials and staff to get familiarize with the information of human resource management and their performance index. And after getting a series of possibly comprehensive firsthand materials, we list an interview outline. Step two: according to the outline, we visit 62 officials in charge of different functions in the provincial tax bureau of city N, talk with relevant leaders and tax collection officials from 10 sub-bureaus, and then list numerous influential factors through an open question which asks about what factors they regard as influential in human resource allocation. From the perspective of tax collection difficulty, collection scale, working intensity, staff quality, we select 50 factors that might influence human resource allocation in taxation system, for instance, “the geographical coverage of tax collection”, “the geographical location of tax region”, and “the population of tax region”, etc.

151.2.2 Scale the Influential Degree of Each Factor

Based on these 50 factors, we make a questionnaire in form of a five-point Linkert item, giving five degrees of each factor (5 = very influential, 4 = influential, 3 = somewhat influential, 2 = not influential, 1 = very not influential). 500 random people were asked to finish the questionnaire and to judge the influential degree of each factor on human resource allocation.

The sample is collected in a wide range from each department of local taxation system in city N, which contains gender (71.4% male, 28.6% female), age (10.5% under 30, 73% between 31 and 45, 14.5% between 46 and 55, 2% above 56), education (14.04% college or below, 85.96% university graduate or above), political status (86.6% CPC, 0.91% democratic party, 12.49% nonparty), service year (6.6% five or five years under, 31.1% six to fifteen years, 62.3% sixteen or sixteen years above), position (3.7% provincial officials, 37.7% grass-root officials, 58.6% ordinary staff), working place (8.18% provincial bureau, 91.82% sub-bureau) and so on. 500 questionnaires are released, and 440 are retrieved, effective questionnaires occupy 88%, which proves to be a very convincing investigation.

With the method of mean calculation, we get the influential degree of each factor, among which “the total number of tax payer” has the maximum influence, while “the ability grade structure of tax staff” has the minimum influence.

Table 151.1 30 influential factors determined by questionnaire results (arranged in a descending order according to influential degree)

Serial number	Influential factor	Serial number	Influential factor
1	Total number of tax payer	16	Geographical coverage of tax collection
2	Total number of enterprise tax payer	17	Completion efficiency of tax collection
3	Enterprise tax per capita	18	Average working performance of staff
4	The number of front-line tax collector	19	This year's quota
5	Working pressure of tax collection	20	Enforcement accuracy of tax staff
6	Working complexity of the department	21	Geographical location of tax region
7	Working efficiency of tax staff	22	Present staff's capacity matching degree
8	The change of local enterprise number	23	Completion rate of income plan
9	The paying consciousness and quality of tax payer	24	Transportation of tax region
10	Comprehensive capability of tax staff	25	Individual household number per capita
11	Working experience of tax staff	26	Unfinished net-declaration rate
12	Demand for talent	27	Responsibility investigation times for enforcement improprieties
13	Present human resource allocation	28	Satisfaction of tax payer
14	Expertise of tax staff	29	Mobility of staff
15	Present staff volume's matching degree	30	Door-to-door tax declaration rate

Regarding 3 as the most widely recognized scale point, we eliminate those factors below the degree of 3, and maintain 30 influential factors (Table 151.1).

Concerning later applications of the results, responses and recognitions from leaders of taxation system toward these factors must be given a full respect. Therefore, on the basis of Chart 1, we release the second round questionnaire especially aiming at 34 officials from provincial tax bureaus, institutions and sub-bureaus. It still adopts the form of five-point Linkert, inviting them to judge the degree of 30 influential factors. With the same mean calculation, we eliminate those factors below the degree of 3, and maintain 11 influential factors (Table 151.2).

151.2.3 Decide the Weight of Influential Factors

According to Chart 2, we make the Delphi questionnaire and have two rounds of investigation with those 34 leaders.

Table 151.2 11 influential factors and weights decided by Delphi method

Type	Example	Weight
Working intensity	This year's quota	0.045
	Total number of enterprise tax payer	0.156
	Total number of individual tax payer	0.007
	Present staff volume's matching degree	0.067
Staff quality	Enforcement accuracy of tax staff	0.083
	Working experience of tax staff	0.092
	Expertise of tax staff	0.117
	Working efficiency tax staff	0.158
Working difficulty	Geographical coverage of tax collection	0.076
	Paying consciousness and quality of tax payer	0.107
	Unfinished net-declaration rate	0.092

The first round of Delphi asks 34 leaders to compare 11 indexes in pairs, and then calculate weight of each factor by Eigen value method; the second round is to give feedbacks to leaders and invite them to revise. All of them have no controversy with the factors but provide some revisions on their weights. After an average normalization of the revised results, we get 11 influential factors and their weights (Table 151.2).

151.3 Types and Index Composition of Influential Factors

Through analysis, 11 key influential factors are grouped into 3 types: working intensity factor, working difficulty factor, and staff quality factor (Table 151.2). Some factors such as the total number of enterprise tax payer, the total number of individual tax payer, geographical coverage of tax collection and unfinished net-declaration rate are quantitative indexes, and some factors cannot be directly quantified, therefore, we have to analyze those factors and replace them with quantitative indexes. The interview with officials helps us compose a list of quantitative indexes.

151.3.1 Working Intensity Factors

Containing factors such as this year's quota, the number of enterprise tax payer and individual tax payer, present staff volume's matching degree, working intensity factors indicate the work amount that tax staff have to shoulder in order to accomplish the task, reflecting work differences between each sub-bureau. It is one of all that influence human resource allocation in local taxation system. With greater working intensity comes greater demand for talents.

This year's quota refers to the total sum of various tax revenues that collected every year and other bill payment, which is closely related to GDP level and measured by annual income.

The number of enterprise tax payer refers to the amount of enterprise tax payers in the precinct of each sub-bureau, who are the main source of tax revenues.

The number of individual tax payer refers to the amount of individual tax payers in the precinct of each sub-bureau, who seem to contribute little in total sum but because of the large population, management is very complicated.

Present staff volume's matching degree refers to the congruence of present tax staff and their working task. Average tax household means the average number of tax payers that each tax official is going to manage.

151.3.2 Working Difficulty Factors

Working difficulty factors include the geographical coverage of tax, the paying consciousness and quality of tax payer, and the unfinished net-declaration rate, indicating environmental differences between each sub-bureau, namely the differences of working difficulty. Working difficulty is another influential factor that affects human resource allocation. With greater working difficulty comes greater demand for talents.

The geographical coverage of tax collection differs according to different sub-bureaus; the larger geographical area brings the greater working difficulty.

The paying consciousness and quality of tax payer is measured by prolongation declaration rate. The lower tax consciousness and quality the payer has, the higher prolongation declaration rate will be and more difficulties the tax staff will face (because they have to ask for payment door by door).

Electronic filing system reduces the complication of synthetic filing procedure. And unfinished net-declaration rate refers to the proportion of tax payers who do not finish net-declaration to the total tax payers. The higher the rate is, the more complicated procedure synthetic filing will be, and more complicated work tax staff will face.

151.3.3 Staff Quality Factors

Staff quality factors involve tax staff's enforcement accuracy, working experience, expertise and working efficiency, which reflect their working capacity and quality. Staff quality is a basic condition for allocating talents, and also is an important evidence to show each sub-bureau's staff training and motivation program, but cannot be taken as a factor that affects talent allocation in local taxation system.

Tax staff's enforcement accuracy is measured by the average accuracy; the working experience is measured by the average service year; the staff's expertise is

Table 151.3 Influential factors and quantitative indexes of human resource allocation in local taxation system

Type	InfluentiFactor	Quantitative indexes
Working intensity	This year's quota	This year's quota
	Total number of enterprise tax payer	Total number of enterprise tax payer
	Total number of individual tax payer	Total number of individual tax payer
Working difficulty	Present staff volume's matching degree	Average tax household
	Geographical coverage of tax collection	Geographical coverage of tax collection
	Paying consciousness and quality of tax payer	Prolongation declaration rate
	Unfinished net-declaration rate	Unfinished net-declaration rate

measured by the proportion of their respective working titles in economy; for their working efficiency, it is decided by last year's completion rate of the planned income. However, the low quality of staff is no excuse for adding more staff. It is no use of solving the problem by allocating more people, which might in turn make the situation even worse; vice versa, the high quality of staff is also no excuse for downsizing. If a sub-bureau has low quality staff, it reveals that this sub-bureau has some problems in introducing, training and motivating staff, which will ask for staff training and motivating program instead of a demand for staff.

Therefore, eliminating the staff quality factors in Table 151.2 and maintaining 7 working intensity and working difficulty factors, we, eventually determine the influential factors and quantitative indexes of human resource allocation in local taxation system (Table 151.3).

151.4 The Modeling Approaches of Constructing Human Resource Allocation in Local Taxation System

The human resource allocation mode of local taxation system can be built on the basis after the influential factors are fixed.

151.4.1 Collect Data and Work Out the Influential Factors' Sub-Coefficient of Local Taxation Bureaus

Firstly, collect the actual data of the quantitative index of all influential factors in each sub-bureau, including GDP historical data, organization income tasks, the number of enterprise tax payer, the number of individual tax payer, the number of

households managed by per tax staff and the geographical coverage of tax collection, prolongation declaration rate and unfinished net-declaration rate and so on.

Secondly, according to the actual data of the quantitative index of each sub-bureau, get the maximum and minimum by moderately rounding numbers upward and downward respectively and work out the sub-coefficient in accordance with the index weight.

151.4.2 Work Out the Working Difficulty Coefficient of Each Local Taxation Sub-Bureau

The working difficulty coefficient of each local taxation sub-bureau can be worked out by weighted average in light of three sub-coefficients—the covered geographical area coefficient of the local taxation sub-bureau, the postponement declaration coefficient and the unfinished net-declaration coefficient. The fixed formula is as follows:

$$k_i = \sum_{j=1}^3 \mu_j k_{ij} \quad (151.1)$$

In this formula, k_i represents the working difficulty coefficient of the number i local taxation sub-bureau; k_{i1} represents the covered geographical area coefficient of the number i local taxation sub-bureau; k_{i2} represents the prolongation declaration coefficient of the number i sub-bureau; k_{i3} represents the unfinished net-declaration coefficient of the number i sub-bureau. And μ_j represents the weight of k_{ij} , which can be gained by Delphi method earlier. Besides, $k_{ij}(j = 1, 2, 3)$ can be gained through division of the set maximum and the minimum according to the corresponding index of k_{ij} .

151.4.3 Work Out the Working Intensity Coefficient of Each Local Taxation Sub-Bureau

The working intensity coefficient of each local taxation sub-bureau can be gained by weighted average according to four sub-coefficients of organization income tasks, enterprise taxpayers, individual taxpayer and tax households managed by per tax staff. The fixed formula is as follows:

$$e_i = \sum_{j=1}^4 \omega_j e_{ij} \quad (151.2)$$

In the formula, e_i refers to the working intensity coefficient of the number i local taxation sub-bureau; e_{i1} refers to the organization income tasks coefficient of the number i local taxation sub-bureau; e_{i2} refers to the coefficient of enterprise tax payer of the number i local taxation sub-bureau; e_{i3} refers to the coefficient of individual tax payer of the number i local taxation sub-bureau; e_{i4} refers to the coefficient of households managed by per tax staff of the number i local taxation sub-bureau; ω_j represents the weight of e_{ij} , which can be gained by the Delphi method earlier. And $e_{ij}(j = 1, 2, 3, 4)$ can be gained through division of the set maximum and minimum according to the corresponding index of e_{ij} .

151.4.4 Work Out the Human Resource Demand Coefficient of Each Sub-Bureau

According to the working intensity coefficient and working difficulty coefficient, the human resource demand coefficient of each sub-bureau can be worked out by weighted sum. The fixed formula is:

$$E_i = \beta * e_i + \gamma * k_i \quad (151.3)$$

In this formula, E_i represents the human resource demand coefficient of the number i sub-bureau; e_i represents the working intensity coefficient of the number i sub-bureau; k_i refers to the working difficulty coefficient of the number i sub-bureau, and β and γ are the weights of working difficulty and intensity coefficients respectively. The previous research and analysis find that the weight ratio of working difficulty and working intensity is 1, so $\beta = \gamma = 0.5$.

151.4.5 Work Out the Human Resource Allocation Coefficient of Each Sub-Bureau

The human resource allocation coefficient can be worked out through normalization in light of the human resource demand coefficient of each sub-bureau. The fixed formula is:

$$\alpha_i = \frac{E_i}{\sum_{i=1}^{15} E_i} \quad (151.4)$$

α_i represents the human resource allocation coefficient of the number i local taxation sub-bureau and E_i refers to the human resource demand coefficient of the number i local taxation sub-bureau.

151.4.6 Decide the Human Resource Allocation Quantity of Each Sub-Bureau

The human resource allocation quantity of each sub-bureau can be decided by multiplying the human resource allocation quantum of the local taxation system with the human resource allocation coefficient of each local taxation sub-bureau. The fixed formula is:

$$Q_i = H^* \alpha_i \quad (151.5)$$

Q_i refers to the human resource allocation quantity of each sub-bureau, H means the human resource allocation quantum of the local taxation system, and α_i represents the human resource allocation coefficient of the number i sub-bureau.

151.5 Suggestions on a Scientific Human Resource Allocation in Local Taxation System

Through investigations and analysis above, we can discover that the key influential factors that affect human resource allocation are the range of working intensity and working difficulty, while the staff quality factor is just a fundamental condition for human resource allocation, which cannot be regarded as effective as others. In the light of these researches, we put forward several suggestions to human resource allocation in local taxation system.

First of all, give a full consideration to the influences caused by working intensity and working difficulty factors. They are reliable bases to predict the demand of each sub-bureau, and are references to determine each bureau's human resource allocation coefficient. In order to realize a sensible establishment of human resource allocation coefficient, we must take these two factors into account.

Secondly, establish human resource allocation coefficient scientifically. The coefficient is a crucial ground for allocating talents, an important yardstick to measure each bureau's talent demand and degree. It is a core task for local taxation system to establish a scientific talent allocation coefficient by using scientific index system, and allocate talents in each sub-bureau by strictly following the coefficient, so that the number and structure of the entire local taxation system can make an organic collaboration with other production factors.

Finally, pay attention to quality promotion of staff. Staff quality is a basic condition for human resource allocation. If staff quality and working capacity fail to meet the requirements of the work, the effects will be enormously damaged. Therefore, we should invest large forces in promoting staff's comprehensive qualities and profession skills so as to elevate effects and qualities of human resource allocation in local taxation system.

References

1. Zhao F (2008) The allocation model of government human resource. *J Jiangxi Agric Univ (Soc Sci)* 2:93–94
2. Zhao Y, Wei P (2008) The mechanism and model of human resource allocation within the organization. *Consum Guide* 7:58–60

Chapter 152

Classification Learning System Based on Multi-Objective GA and Frigid Weather Forecast

Zhang Hongwei, Xu Jingxun and Zou Shurong

Abstract A new classification learning system based on multi-objective GA is proposed in this chapter. Firstly, the continuous attributes of samples are made discretion with a supervised segmentation method, so generalization and intelligibility of machine learning are improved. Moreover, comparison and selection mechanism based on partial order in set theory are infused into multi-objective GA. These enhance the ability to choose better chromosomes. The new algorithm is used to forecast frigid weather in northern Zhejiang province. The experiment result indicates that it has unique intelligence, higher accuracy.

Keywords Supervised segmentation method · Machine learning · Multi-objective GA · Frigid weather forecast

152.1 Introduction

Now the existing classification learning systems based on GA are almost due to the single objective [1, 2]. A new classification learning system based on multi-objective GA is proposed in this chapter. It can be widely applied to solve

Z. Hongwei (✉) · X. Jingxun · Z. Shurong
College of Computer Chengdu University of Information Technology,
Chengdu, People's Republic of China
e-mail: zhang_hw99@163.com

X. Jingxun
e-mail: xjx19861008@163.com

Z. Shurong
e-mail: zousr@cuit.edu

complicated nonlinear problem. Features of the algorithm are following: (1) The continuous attributes of samples are made discretion with a supervised segmentation method [3, 4]. The value spaces of continuous attributes are divided into multiple sub-intervals based on subjective thresholds. Just do this, enhance the generalization and understandability of classification. (2) Pareto approach [5, 6], comparison and selection mechanism based on partial order are infused into evaluation operator. It can guide the direction of evolution and is easy to choose better chromosomes.

Disastrous weather forecast is always the focus of atmospheric research. It is the eternal theme of the meteorological work, especially due to major floods, drought, typhoon and frigid weather. In recent decades, the disastrous weather forecast goes through several stages of experience forecast, weather model forecast, model output and statistics forecast, numerical forecast and multi-method integration forecast [7]. The intelligent approaches based on BP neural network [8] make some progress in this field. But these methods are poor in scalability and bring over-fitting in neural network learning algorithms. They often get local optimal solution and are complex to find structure parameters. The changes of meteorological factors also show a few features, such as mutation, highly non-linear, complexity and so on. Novel factor combination method [9] with thought of logical reasoning is almost impossible to obtain success, because it has more human intervention and great uncertainty.

The new algorithm is used to forecast frigid weather in northern Zhejiang province. It gets a very satisfactory forecast result. On time this algorithm lays a preliminary foundation for other disastrous weather forecast.

152.2 Symbol and Algorithm

According to meteorological document, if the average temperature in January or in February is less than 2°C , it means that frigid weather appears this winter [9]. Major meteorological factors are below:

- X_1 : Average temperature in March this year,
- X_2 : Accumulated temperature from October to December last year,
- X_3 : Average temperature in October this year,
- X_4 : Accumulated temperature annually last year,
- X_5 : Precipitation from July to September last year,
- X_6 : Precipitation from July to August last year,
- X_7 : Precipitation from April to June last year,
- X_8 : Accumulated temperature from January to June this year.

When some factors are far from the mean state and exceed or fall below some thresholds on time, they can result in abrupt change of weather.

Table 152.1 Encoding table of frigid class

Coding value	3	1	*
Attribute			
X_1	$(-\infty, 9.8]$	$(-\infty, 8.7]$	R
X_2		$[37, +\infty)$	R
X_3		$(-\infty, 16.9]$	R
X_4		$(183.5, +\infty)$	R
X_5		$(-\infty, 280)$	R
X_6		$(-\infty, 375)$	R
X_7		$(-\infty, 426.7)$	R
X_8		$(-\infty, 74.5)$	R

Table 152.2 Encoding table of non-frigid class

Coding value	Attribute	4	2	#
X_1		$(8.7, +\infty)$	$(9.8, +\infty)$	R
X_2			$(-\infty, 37)$	R
X_3			$(16.9, +\infty)$	R
X_4			$(\infty, 183.5]$	R
X_5			$[280, +\infty)$	R
X_6			$[375, +\infty)$	R
X_7			$[426.7, +\infty)$	R
X_8			$[74.5, +\infty)$	R

152.2.1 Encoding and Decoding

Encoding Rule

When some condition attribute exceeds some thresholds, the decision attribute can abruptly change. Therefore, these thresholds are arranged in descending order. Every threshold is made as the lower bound of each sub-range, and “+∞” is made as its upper bound.

When some condition attribute falls below some thresholds, the decision attribute can abruptly change. Therefore, these thresholds are arranged in ascending order. Every threshold is made as the upper bound of each sub-range, and “−∞” is made as its lower bound.

The sub-ranges in ascending order are encoded as the integers in ascending order.

For example, the abrupt change of decision attribute can occur when $X_1 \leq 8.7$ or $X_1 \leq 9.8$. According to range, $(-\infty, 8.7]$ is encoded as 1 and $(-\infty, 9.8]$ is encoded as 3.

Frigid weather is affected by many meteorological factors. According to thresholds of the condition attributes in the document [9], encoding tables are obtained (Tables 152.1 and 152.2).

Concrete encoding rule

- (1) The sub-ranges of every condition attribute are arranged in ascending order. Sequentially judge the attribute value whether belongs to some sub-range, the

corresponding integer of the first eligible sub-range is made as its encoding value. If the attribute value does not belong to any sub-range, this means it is unrelated to the decision attribute finally. Therefore, “*” or “#” is made as the encoding value.

- (2) Different decision attributes are encoded as different integer. Such as frigid weather is encoded as “1”, and non-frigid weather is encoded as “0”. The alleles of chromosomes and sample in different class after encoding are different. The encoding values of frigid class belong to {1, 3, *}, while the encoding values of non-frigid class belong to {2, 4, #}

Take a sample in frigid class for encoding example. Before encoding: $h = (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = (7.7, 36.3, 16.4, 183.4, 494.4, 368.2, 596.3, 70.1)$.

After encoding: $e = (1, *, 1, *, *1, *, 1, 1)$.

Decoding Rule

- (1) Referring to encoding chats, the alleles are replaced by their corresponding sub-ranges. So “*” and “#” are replaced by R.
- (2) The encoding value of decision attribute is replaced by its corresponding actual conclusion.

152.2.2 Definition

Combine the knowledge of the classical set theory. There is coverage relationship among chromosomes in the same class. There is also coverage relationship between chromosome and sample in the same class. On time there is intersection relationship in the different class.

Coverage Relationship

If the alleles of chromosomes are greater, therefore the corresponding coverage space is greater. If all the alleles in a chromosome are greater than another, there is coverage relationship between them. Also “*” and “#” are greater than all the integers.

For example: two chromosomes e_2 and e_1 are in the same class. $e_1 \prec e_2 \Leftrightarrow e_{1j} \leq e_{2j}$, Called e_2 covers e_1 . Above j can reach every allele of chromosome, but it does not the last allele.

Intersection Relationship

The corresponding sub-ranges of different alleles are also different. There is intersection relation among chromosomes in different class.

The chromosomes e_2 and e_1 are in the different class. If there is intersection among the corresponding sub-ranges of their every allele, therefore e_2 and e_1 intersect. Similarly if there is intersection relation between chromosome and sample, so $e \cap h \neq \emptyset$.

Number of Samples Covered

$m_e = |\{h | e \succ h\}|$, here h is the sample set, and m_e is the number of samples covered by the chromosome e after decoding.

152.2.3 Fitness Vector Function

According to meteorological experience knowledge:

- (1) The number of condition attribute that affect decision attribute is less, and the complexity of analysis is lower. So “*” or “#” should be as more as possible.
- (2) Set m is the number of samples covered by some chromosome, and they are in the same class. Due to the same chromosome, set n is the number of samples covered by it, but they are in different class. Obviously m is greater and n is less, therefore this chromosome is better.

Based on the two points above, construct fitness vector function $\tilde{f} = \left(f_1 = \frac{n+1}{m+1}, f_2 = \frac{1}{1+v} \right)$, and v expresses the num of “*” or “#” in a chromosome. In order to take a medium, set two thresholds as $\alpha = 1, \beta = 1$. If $f_1 < \alpha \wedge f_2 < \beta$, so the chromosome is reserved in set ND . Just do this, retain the chromosomes which have forecasting ability as more as possible.

152.2.4 Complete Implementation Steps of Algorithm

Deletion Rule

In order to eliminate all superfluous chromosomes, set the deletion rule below. Choose randomly two chromosomes e_1 and e_2 from the memory pool ND , m_{e_1} and m_{e_2} are e_1 and e_2 corresponding number of samples covered.

If $e_2 \prec e_1 \wedge m_{e_2} = m_{e_1}$, then delete e_1 from ND .

If $e_2 \prec e_1 \wedge m_{e_2} < m_{e_1}$, then delete e_2 from ND .

Classification Learning Algorithm Based on Multi-objective GA

1st step: Initialize randomly the chromosome population $E = \{e|e_i, i = 1, 2 \dots K\}$. Set M is the samples set after encoding, the non-dominated chromosomes set $NDSet = \emptyset, ND = \emptyset$, the iteration T .

2nd step: Choose a chromosome e from E and delete it from E . Definite $m = 0, n = 0, v = 0$

3rd step: Select a sample h from M to compare with the chromosomes e .

- (3.1) If $e \succ h$, the condition attributes and decision attributes are consistent, then set $m = m + 1$;
- (3.2) If $e \cap h \neq \emptyset$, the condition attributes are consistent, but the decision attributes are contradictory, therefore set $n = n + 1$.

4th step: Repeat 3rd step until there is not any sample existed in M . Calculate v with the chromosome e and calculate $f_1 = \frac{n+1}{m+1}, f_2 = \frac{1}{v+1}$. If $f_1 < 1 \wedge f_2 < 1$, therefore update $ND = ND \cup \{e\}$.

5th step: Return to 2nd step until there is not any chromosome existed in E .

6th step: If the iteration T can satisfy actual demands, then go to 9th, otherwise go to next step.

7th step: Select the non-dominated chromosomes from the population ND through AP algorithm and update the set $NDSet$. Make crossover and mutation among the chromosomes in $NDSet$, therefore get a new chromosome set P . Calculate the objective function value and update ND .

8th step: Select the non-dominated chromosomes among $NDSet \cup P$ as the next generation of population with AP algorithm again.

(8.1) If $|E| < K$, so randomly generate chromosomes until the original population size.

(8.2) If $|E| > K$, then delete some chromosomes until the original population size. Return to 2nd step.

9th step: Choose the better chromosomes from ND according to deletion rule above, at last output the result.

152.3 Example Analysis

152.3.1 Examples Data

In order to validate the effectiveness and intelligence of the algorithm, the meteorological data in Northern Zhejiang among 42 years is used as the experimental data. Here the matrix $X_{13 \times 8}^1$ and $X_{13 \times 8}^2$ are used to show the training samples set M . Take $X_{16 \times 8}^3$ as the prediction samples set C .

The vertical axis shows the corresponding real values of meteorological factors $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$ in some year. Also $Y_{13 \times 1}^1$ and $Y_{13 \times 1}^2$ means the actual weather of the years above (“1” means the frigid weather, “0” means the non-frigid weather, unit of precipitation is mm, unit of temperature is °C).

$$X_{13 \times 8}^1 = \begin{pmatrix} 7.6 & 30.3 & 16.8 & 183.4 & 695.8 & 406.1 & 445.2 & 73.8 \\ 7.7 & 36.3 & 16.4 & 183.4 & 494.4 & 368.2 & 596.3 & 70.1 \\ 9.7 & 32.5 & 15.2 & 187.0 & 490.9 & 272.1 & 395.6 & 76.1 \\ 10.8 & 34.8 & 17.5 & 191.6 & 299.2 & 173.2 & 548.5 & 76.2 \\ 11.2 & 33.7 & 17.6 & 194.5 & 692.7 & 504.1 & 545.8 & 80.2 \\ 9.8 & 38.5 & 18.4 & 201.8 & 442.2 & 166.2 & 401.4 & 80.0 \\ 9.0 & 34.7 & 17.1 & 188.9 & 597.9 & 289.7 & 424.4 & 74.2 \\ 10.1 & 35.1 & 16.0 & 190.2 & 376.9 & 152.6 & 462.5 & 75.3 \\ 9.2 & 34.5 & 18.0 & 193.2 & 268.9 & 125.1 & 475.3 & 75.6 \\ 8.0 & 35.5 & 17.6 & 187.4 & 281.0 & 217.4 & 316.4 & 74.8 \\ 10.5 & 34.3 & 17.6 & 192.5 & 367.2 & 202.6 & 461.9 & 80.0 \\ 9.4 & 30.4 & 17.6 & 188.0 & 167.9 & 135.4 & 427.1 & 65.9 \\ 9.7 & 38.7 & 16.6 & 190.3 & 256.3 & 180.9 & 444.0 & 72.9 \end{pmatrix} \quad Y_{13 \times 1}^1 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$X_{13 \times 8}^2 = \begin{pmatrix} 7.4 & 32.3 & 17.4 & 184.3 & 610.6 & 395.0 & 300.7 & 71.8 \\ 6.0 & 35.4 & 17.9 & 186.0 & 525.1 & 375.1 & 426.7 & 70.7 \\ 8.0 & 32.6 & 15.6 & 189.9 & 230.1 & 67.5 & 432.3 & 75.6 \\ 9.0 & 35.2 & 17.1 & 184.7 & 244.4 & 140.2 & 294.5 & 93.2 \\ 10.3 & 33.2 & 17.3 & 191.4 & 416.9 & 116.8 & 548.5 & 80.3 \\ 8.5 & 34.8 & 17.4 & 186.4 & 257.8 & 224.4 & 426.5 & 75.4 \\ 9.2 & 34.4 & 18.7 & 193.3 & 427.3 & 233.7 & 424.5 & 77.0 \\ 8.0 & 31.0 & 17.4 & 182.6 & 280.3 & 118.1 & 353.4 & 75.3 \\ 10.2 & 37.8 & 18.9 & 188.0 & 759.4 & 385.7 & 532.1 & 71.7 \\ 8.1 & 35.5 & 17.6 & 194.4 & 133.9 & 44.3 & 335.6 & 77.2 \\ 9.1 & 35.0 & 17.4 & 192.2 & 439.5 & 305.4 & 274.8 & 79.1 \\ 7.4 & 35.0 & 17.3 & 181.2 & 782.5 & 654.3 & 334.5 & 73.1 \\ 10.5 & 30.0 & 15.5 & 185.0 & 495.0 & 431.7 & 441.0 & 77.6 \end{pmatrix} Y_{13 \times 1}^2 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

$$X_{16 \times 8}^3 = \begin{pmatrix} 9.0 & 37.2 & 18.9 & 189.6 & 422.6 & 378.6 & 302.0 & 76.8 \\ 8.6 & 33.3 & 17.5 & 189.0 & 416.8 & 191.6 & 764.8 & 76.6 \\ 8.7 & 34.8 & 17.3 & 184.0 & 310.4 & 177.2 & 592.1 & 69.8 \\ 6.7 & 33.0 & 18.2 & 188.5 & 433.2 & 248.7 & 364.0 & 75.4 \\ 9.0 & 33.9 & 16.5 & 188.1 & 385.3 & 248.7 & 498.0 & 76.7 \\ 8.3 & 36.6 & 19.1 & 191.0 & 739.8 & 603.5 & 372.6 & 76.8 \\ 7.4 & 35.4 & 18.3 & 191.6 & 393.1 & 260.2 & 355.3 & 76.6 \\ 9.5 & 35.9 & 17.9 & 190.6 & 668.2 & 393.1 & 440.8 & 78.2 \\ 11.4 & 36.5 & 17.5 & 200.7 & 480.7 & 276.7 & 394.3 & 82.6 \\ 8.1 & 34.7 & 17.3 & 190.1 & 436.4 & 338.4 & 525.7 & 76.7 \\ 8.1 & 33.6 & 16.1 & 190.6 & 414.5 & 267.7 & 272.4 & 78.4 \\ 9.1 & 34.0 & 17.2 & 188.3 & 680.3 & 593.0 & 383.0 & 77.7 \\ 8.9 & 40.3 & 17.6 & 286.2 & 220.2 & 183.5 & 448.6 & 82.6 \\ 10.5 & 33.8 & 18.3 & 193.9 & 320.1 & 296.5 & 473.9 & 78.2 \\ 8.1 & 36.9 & 18.6 & 198.1 & 522.0 & 486.0 & 429.2 & 74.5 \\ 11.3 & 36.4 & 18.2 & 198.7 & 943.0 & 420.7 & 195.9 & 84.9 \end{pmatrix}$$

152.3.2 Forecast Chromosomes Output

Set parameters $K = 50$, $|M| = 26$, $T = 100$, $|C| = 16$, the crossover rate $p_c = 0.8$ and the mutation rate $p_m = 0.1$. The matrix $Z_{5 \times 10}^1$ shows the output result. The vertical axis shows respectively the corresponding encoding value of the factors set $(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$, the forecast result and m_e .

Table 152.3 Forecast result and actual result

Year	Forecast result	Actual result
1982	0(non-frigid)	0(non-frigid)
1983	0(non-frigid)	0(non-frigid)
1984	1(frigid)	1(frigid)
1985	0(non-frigid)	0(non-frigid)
1986	un certain	0(non-frigid)
1987	0(non-frigid)	0(non-frigid)
1988	0(non-frigid)	0(non-frigid)
1989	0(non-frigid)	0(non-frigid)
1990	0(non-frigid)	0(non-frigid)
1991	0(non-frigid)	0(non-frigid)
1992	un certain	0(non-frigid)
1993	0(non-frigid)	0(non-frigid)
1994	un certain	0(non-frigid)
1995	0(non-frigid)	0(non-frigid)
1996	0(non-frigid)	0(non-frigid)
1997	0(non-frigid)	0(non-frigid)

$$Z_{5 \times 10}^1 = \left\{ \begin{array}{cccccccccc} 1 & * & * & 1 & * & * & * & 1 & 1 & 2 \\ 1 & * & 1 & * & * & 1 & * & 1 & 1 & 1 \\ \# & \# & 2 & \# & 2 & \# & \# & 2 & 0 & 9 \\ \# & 2 & 2 & 2 & 2 & \# & \# & \# & 0 & 2 \\ 4 & 2 & 2 & \# & 2 & \# & \# & \# & 0 & 7 \end{array} \right\}$$

If m_e is greater, the high-dimensional coverage space of the chromosomes is greater. On time the forecast generalization for frigid weather is greater.

152.3.3 Forecast Result Output

Forecast Rule

Choose a sample c from C . Judge every real value of c whether they are all covered by the corresponding sub-ranges of every allele in a chromosome. If there is a chromosome like this in ND , therefore the decision attribute of this chromosome is the forecast result of the sample c .

Illustration by Example

Suppose a sample and a chromosome as follow:

$$c_0 = (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = (8.7, 34.8, 17.3, 184, 310.4, 177.2, 592.1, 69.8), e_0 = (1, *, *, 1, *, *, *, 1, 1).$$

Get frigid forecast expression from e_0 . If some condition attributes satisfy the following expression $X_1 \leq 8.7 \wedge X_4 > 183.5 \wedge X_8 < 74.5$, then frigid weather can appear this winter. Because real values of c_0 all satisfy the forecast expression above, so the frigid weather can appear this winter.

Forecast Result

According to the matrix $Z_{5 \times 10}^1$ and forecast rule, get the forecast result of test samples C (Table 152.3).

The only frigid sample is forecasted correctly. The correct rate of non-frigid samples is 80%. Conventional statistical methods are difficult to solve forecast problems. But make a more accurate forecast with classification learning system based on multi-objective GA in this chapter. Thus, the result verifies the validity and superiority of new algorithm. Moreover, it provides an important reference and a new approach for other disastrous weather forecasting.

152.4 Conclusion

A new classification learning system based on the multi-objective GA is presented in this chapter. It shows unique effectiveness and intelligence and practicality through forecast result. Because disastrous weather has some common features, so it also has a great help and guide for other regions and other disastrous weather forecast, this algorithm possesses a good application prospect and practical significance.

Generalization and intelligibility are all of importance for machine learning. Analyze and find dominant factors, sensitive factors, restrain factors in many meteorological factors. Increase forecast lead-time and forecast accuracy as soon as possible. These works will be researched next.

References

1. Ming C, Shu-Ping H, Fan-Zhang L (2009) Research on linear classifier algorithm in lie group machine learning. *Microelectron Comput* 26(10):64–70
2. Zhong W, Ling Z (2007) A two-layer classifier design method based on covering algorithm. *Comput Technol Dev* 17(1):65–72
3. Yan-Fang J, Qiang-Li Z (2009) *Machine learning techniques*. Electron Industry Press, Beijing
4. Michalski RS et al (2004) *Machine learning and data mining*. Publishing House Electronics Industry, Beijing
5. Hongwei Z, Xiaoke C, Shurong Z (2010) Multi-objective transportation optimization based on fmca. *IEEE Intern Conf Inf Manag Eng* 4(2):426–430
6. Gong M, Jiao C et al (2008) Multi-objective immune algorithm with non-dominated neighbor-based selection. *Evolut Comput* 16(2):225–255
7. Hua-Sheng Y et al (2005) *Analysis and forecasting of climate change*. Meteorology Press, Beijing
8. Long J (2005) *Research on theoretical methods and application of meteorological forecast based on neural network models*. Meteorology Press, Beijing
9. Hong-Xiang Z (2005) *Forecast methods for disastrous weather forecast*. Meteorology Press, Beijing

Chapter 153

Empirical Analysis on Choice of Payment Terms in Foreign Capital Acquiring State-Owned Enterprise

Ling Liu and Jingru Wu

Abstract With the development of economic globalization foreign acquisition has become the important method of international direct investment. In order to adapt international tendency of multinational corporation investment China adjusts foreign capital policy gradually and the form of using foreign capital is changing from traditional ways of joint venture and cooperation to using multinational acquisition. Combining the further reform of state-owned enterprise China established laws and regulations as to foreign capital acquisition, in which includes the regulations of foreign capital acquiring state-owned enterprise. So on the basis of development progress of Chinese foreign capital acquiring state-owned enterprise studying on change of legal surrounding and choice of payment terms in foreign capital acquiring state-owned enterprise has the vital meaning. By studying on legal surrounding in foreign capital acquiring state-owned enterprise the article analyzes the payment terms and give some workable suggestions to faults in payment terms in splitting off stock in foreign capital acquiring state-owned enterprise in Corporation Law and Security Law.

Keywords Foreign capital acquisition · State-owned enterprise · Payment terms · Empirical analysis

L. Liu (✉) · J. Wu
Hebei University of Science and Technology, No. 70 Yuhua Road,
Shijiazhuang, Hebei Province, China
e-mail: Liulingpriyu@126.com

J. Wu
e-mail: wujingru.teacher@163.com

153.1 The Development of Foreign Capital Acquiring State-Owned Enterprise

153.1.1 The First Period

After South Tour speeches of Mr. Deng Xiaoping in 1992 China began to reform deeply and enlarge opening up. At this time trends of economic globalization has developed quickly and multinational corporation began to step to global management. China is the important circle in global strategy. When multinational corporations have the global strategy adjustment it must bring China into the total schedule of it. With the development of multinational corporation investment in China, multinational corporation acquisition in China is spreading quietly. According to the statistics of Business Development Conference of U.N. from 1991 to 2001 multinational acquisition in China is between US\$ 1 and 2 billion, which covers about 5% of total volume of foreign direct investment. Objectively foreign capital acquisition in China is still in the first period from 1980 to 2001.

153.1.2 Developing Period

After China entered WTO in December 2001 the step of multinational corporations acquiring Chinese enterprises is improving obviously. In 2001 the proportion of joint-venture covering total foreign investment decreases to less than 3% and the proportion of sole foreign projects covering foreign projects is 70%, which the proportion of actual using amounts of foreign investment is over 50%. In change of joint-ventures and sole foreign capital foreign capital acquisition is improving gradually. In 2003 China becomes the most one to attract foreign direct investment in the world, which amounts is US\$ 53.5 billion. Although in 2003 the proportion of selling amounts of acquisition in China covering total foreign capital amounts improves from 5% to 13.6%, the famous multinational groups chose to use acquisition to carry out its investment strategy in China. So multinational acquiring state-owned enterprises enters into the gradual developing period.

153.2 Legal Surrounding in Foreign Capital Acquiring State-Owned Enterprises

153.2.1 Legislation of the First Period

During the first period of foreign capital acquiring state-owned enterprise China did not have legislation about foreign capital acquiring state-owned enterprises and only has minor department regulation. In 1989 State Commission for Restructuring

the Economic System, Treasury Department and State Administration of State Property issued <Provisional Regulations about Selling Property Rights of Small Size State-owned Enterprises> in which regulates “When foreign businessman, overseas Chinese businessman and businessman in Hon Kong, Macau and Taiwan purchase small size state-owned enterprises, except business enterprises and other national regulation, it can carry out referring to this regulation”. It affirms that foreign capital can acquire the assets of small size state-owned enterprises directly and does not regulate whether other state-owned enterprise can be acquired directly. On 23 September 1995 the State Council issued <Opinions to Related Issues on Listed Companies referring to Foreign Investment>, in which transferring state shares and corporate shares of listed companies to foreign businessmen is forbidden. In 1999 State Economic and Trade Commission issued <Provisional Regulation of Foreign Investment Acquiring State-owned Enterprises >, in which it regulates definitely that foreign can be engaged in acquiring state-owned enterprises. In July 2000 Ministry of Foreign Trade and Economics and State Administration of Industry and Commerce issued <Provisional Regulation of Inner Investment of Foreign Investment Enterprises>, in which it allows eligible foreign investment enterprise acquires inner state-owned companies. In 2001 Ministry of Foreign Trade and Economic Cooperation and State Administration of Industry and Commerce issued <Modification to Regulations about Combination and split of Foreign Investment Enterprises>, in which it allows foreign investment combines with Chinese enterprises and it opens the door to foreign capital acquisition. In 2001 Ministry of Foreign Trade and Economics and Securities and Futures Commission issued <Opinions to related issues on Listed Companies referring to Foreign Investment>, in which it regulates that foreign investment enterprises (exclusive professional investment companies) can be transferred and hold non-tradable shares in inner listed companies according to <Provisional Regulations about Inner Investment of Foreign Investment Enterprises>.

About the permits of foreign acquiring state-owned enterprises limits and progress of different departments lack operable regulations and there are many barriers in terms of payment terms, preferential policy, debt disposing, foreign exchanges management and the price of state-owned stock rights. So it is difficult for foreigners to have operable judge and estimate the economic, political and legal risks.

153.2.2 Legislation in the Development Period

The Fourth Plenary Session of the Fifteenth of the Central Committee of the Communist Party of China issued <Decision to Many Vital Problems about State-owned Enterprises Reform and Development> and put forward to continue having strategic reorganization to state-owned enterprises and should “seize the big and free the small”. When training strong big enterprises and groups we should adopt many ways of reorganization, union, merger, lease, contractual operation, stock

cooperation and sale to enliven state-owned enterprises. The report of 16th CPC National Congress <Building well-off society in an all-around way, opening up new situation of socialism with Chinese characteristics> put forward to utilizing middle or long term foreign investment by many ways and connecting utilizing foreign investment with domestic economic structure adjustment and state-owned enterprises reform.

Entering to WTO symbols the great adjustment of Chinese foreign economic development strategy. Before it Chinese government has revised many law and regulations. In terms of absorbing foreign capital the limits to foreign capital acquisition were cancelled gradually and issued regulation of foreign acquisition. On 1st, January 2003 four ministries and commissions issued <Provisional Regulations of Utilizing Foreign Capital to Reorganize State-owned Enterprises>, which lists property forms of payment in foreign acquisition in detail. In April 2003 four ministries and commissions issued <Provisional Regulation of Foreign Investors Acquiring Domestic Enterprises> which regulates definitely that foreign investors regard his own stock as payment term. Otherwise in Chinese stock market a series of policy of encouraging foreigners direct investment and acquisition were carried out. For example in 2002 China Securities Regulatory Commission, Treasury Department and National Economics and Trade Commission issued <Notification about Transferring State-owned Shares and Corporate Shares in Listed Company to Foreigners> which defines that foreign capital can adopt merger and purchase to acquire domestic enterprises (inclusive state-owned enterprises). These laws and regulation brings about conditions of foreign investors entering to China according to market economics rules and law of enterprise self-development.

153.3 Payment Terms of Foreign Capital Acquiring State-Owned Enterprises

153.3.1 Cash Payment

In foreign capital acquiring state-owned enterprises cash purchase is the main payment term. Especially in the first period of foreign capital acquiring state-owned enterprises almost in all cases foreign capital adopted the term. Cash purchase refers to purchaser uses cash to purchase enterprises. Cash purchase includes currency, deposits in banks and other notes which can be exchanged cash. According to the objects of purchasing aimed companies it can be divided into cash purchasing assets, cash purchasing shares right, cash purchasing stocks, cash purchasing financial claim and cash purchasing debt.

153.3.1.1 Cash Purchasing Assets

Foreign capital usually adopts cash purchasing assets form to acquire state-owned non-listed company. For example, in May 1992 China Strategic Holdings Limited purchased the assets of Taiyuan Rubber Plant of Shanxi Province; Kodak Company acquired all industry of Chinese sensitive material market; in February 2001 Huawei Group sold Emerson Electric to Avansys Electric Group at the price of US\$ 0.75 billion; in March 2001 the joint venture holding by Michelin paid US\$ 0.32 billion to purchase core business and assets of Chinese tyre and rubber enterprises.

153.3.1.2 Cash Purchasing Shares Right

Foreign capital usually adopts cash purchasing shares right form to acquire state-owned listed company. For example, in October 2001 France Alcatel Company purchased 51% shares of Shanghai Bell Company at the basis of cash US\$ 0.3 billion; in 2003 Citibank invested RMB ¥ 0.6 billion to purchase 5% shares of Shanghai Pudong Development Bank; in October 2003 Eastman Kodak company invested US\$ 0.1 billion and other assets and supplied technical support to purchase 20% state-owned corporate shares of Lucky Film Co., Ltd. Controlled by Lucky Group. In December 2003 Greencool Group got RMB ¥ 0.56 billion from Shunde Greencool Enterprise Development Company to purchase 20.6% corporate shares of Kelon Electrical on treaty.

153.3.1.3 Cash Purchasing Stocks

Foreign capital usually adopts cash purchasing stocks form to acquire state-owned listed company. For example, on 2nd August 1995 Ford Motor Company purchased US\$40 million B share of Jianglin; in March 1999, Huaxin Cement issued additional stock 77 million B shares to HOLCHIN B. V. Company; in September 2000 Royal Dutch Shell Group invested US\$0.43 billion to purchase stocks of Sinopec and cooperated with Sinopec and on 28th October purchased US\$0.4 billion shares of China National Offshore Oil Corporation; in April 2001 US IDT Company invested US\$850 million to purchase Xintao Science and Technology Company (Shanghai) by the form of cash purchasing stocks.

153.3.1.4 Cash Purchasing Financial Claim

Foreign capital usually adopts cash purchasing financial claim form to acquire state-owned non-listed company because state-owned assets is financial claim. For example in the fourth season 2001 China Huarong Asset Management Corp had the assets over RMB ¥ 22 billion involving 18 provinces and cities international

bidding and constituted 5 assets bags for sale. Bidding groups at the head of Morgan Stanley purchased four assets bags in the form of cash adding cooperative management.

153.3.1.5 Cash Purchasing Debt

When foreign capital purchased stocks of state-owned enterprises because this state-owned enterprise owed to the third one foreign capital took the place of the state-owned enterprise to repay a debt and got stocks of state-owned enterprises and so they adopted cash purchasing debt which is the transformation of cash purchasing debt. For example, in March 2002 Greencool Enterprise Development Company repaid a debt of RMB ¥ 0.348 billion in the place of Rongsheng Group and he became the first shareholder owning 20.6% ST Kelon.

153.3.2 Comprehensive Payment Terms

Chinese existing laws and regulations broaden payment terms of foreign capital acquisition which allowed foreign investors use stocks and other securities as payment terms. For example in acquisition case of Qingdao Beer Group and the famous Anheuser-Busch Company Qingdao Beer Group issued directional convertible securities valued US\$0.182 billion to AB Company for three times and in the agreement all securities can be transformed shares right in seven years. Although in the case financial innovative way convertible securities is applied because Qingdao Beer who issued convertible securities was the target enterprise AB company still paid at the basis of cash. In the case of Eastman Kodak Company purchasing 20% state-owned corporate shares of Lucky Film Co., Ltd Eastman Kodak Company used cash and assets technical support as the payment terms, which is called comprehensive payment. Although it did not escape cash payment it makes the step of using cash in full.

From above although exchange payment is the main payment of foreign capital acquiring state-owned enterprises and there is no obstacle on law up to now no case of foreign capital acquiring state-owned enterprise uses the international popular payment terms. The reason is that there are many legal faults concerning exchange payment in Corporate Law and Securities Law which needs to perfect.

153.4 Suggestions

Corporation Law should be revised and regulates definitely that stocks can be payment terms and introduces mandatory procedure of stocks exchange in American corporation law and establishes relative rights for opposed shareholders

in order to protect. Corporation Law should be revised and introduce the system of stocks buy-back and offers legal protection barrier to oppose hostile takeover to Chinese state-owned enterprises. Securities Law should be revised and must adjust that purchase must be by the term of cash to that purchase can be by the term of exchange stocks and comprehensive securities besides on the basis of cash and should add the regulation that purchase offer of exchange payment should be registered. Securities Law should be revised to especially protect the rights of middle or small shareholders and change state-owned shares to preferred stocks and establishes circulation system in stock market to state-owned stocks and perfect the price style of state-owned stocks.

Acknowledgments The work is supported by supported by the Funds Project of Hebei University of Science and Technology (XL200778).

References

1. Du X, Zhang R (2003) Focus on international acquisition and reorganization aroused by state-owned shares transfer. *Int Finance* 12:087–092
2. Jin Z, Jiang R (2003) Study on state-owned share buy-back to Chinese listed company. *J Beijing Univ Chem Technol (Soc Sci Edn)* 43:3–6
3. Zhang L (2003) Systemic obstacle of foreign capital acquisition from <notice of relative problems about state-owned shares and corporate shares of limited company transferred to foreign investors>. *J Tongling Coll* 2:142–148
4. Chen C, WTO and development of multinational corporate acquisition in china. *J Hefei Univ Technol (Social Science Edition)*, 9:22–26
5. Wang Z (2004) Report on multinational corporate in China. Chinese Economics Press, p 19
6. Wang Q (2005) Study on multinational corporate acquiring Chinese enterprise. PHD dissertations of Chinese Academy of Social Science, pp. 39–40
7. Yang ZG (2003) Foreign capital acquisition focusing on China. *Int Finance* 1(2):66

Chapter 154

An Integrated Application of Tourism Planning Based on Virtual Reality Technology and Indicator Assessment

Yan Zhang and Xun Zhang

Abstract This chapter tries to manage the indicator analysis based on Virtual Reality (VR) technology to present an integrated method for rational tourism development planning. It makes the application of the integrated method with indicators and sub-indicators of ecological and VR assessment. It develops a VR model by presenting the attraction's images and a prototype of an interactive visual simulation system in the planning area. It analyses that VR technology is one of the multi-indicator ways that provide useful support in the choice alternatives for tourism planning. It makes suggestions for the future planning of the study area, and implementing a vision of future studies based on this system.

Keywords Tourism planning · Virtual reality technology · Indicator assessment · Nansong imperial palace

Y. Zhang (✉)

Zhejiang Gongshang University, No.18, Xuezheng Str, Xiasha
University Town, 310018 Hangzhou, China,
e-mail: isabelle@126.com

X. Zhang

Zhejiang University of Finance and Economics, No.18, Xueyuan Str,
Xiasha University Town, 310018 Hangzhou, china,
e-mail: zhangx1974@126.com

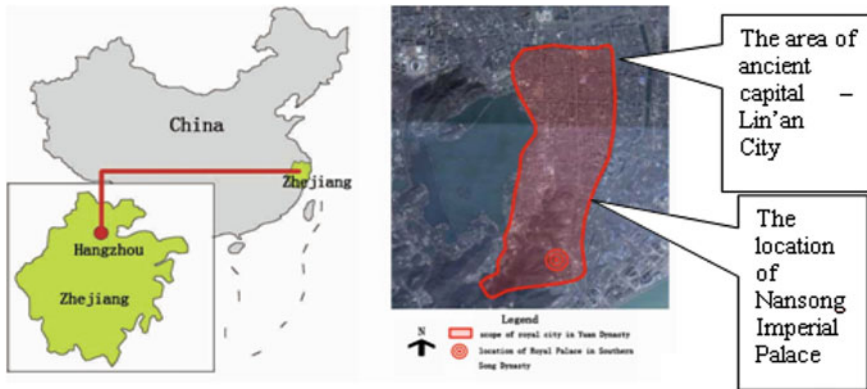


Fig. 154.1 The study area

154.1 Introduction

Due to the complexity in the decision process, many conventional methods are able to consider limited indicator and may be generally deficient [1–5]. Therefore, it is clearly seen that assessing of all indicators connected to the method selection by combining the decision making process is extremely significant (Hartman and Mutmansky 2002).

This chapter tries to manage the matrix analysis between ecological analysis and Virtual Reality (VR) environment assessment to present an integrated method for rational tourism development planning and management by the case study application of the integrated method with indicators and sub-indicators of ecological and VR assessment for rational planning. It enables to develop a VR model by presenting its images and a prototype of an interactive visual simulation system in the heritage area. The analysis is one of the multi-indicator technologies that provide useful support in the choice among several alternatives with different objectives and indicators [6–9].

154.2 Study Area

Figure 154.1 shows the location of Hangzhou city and Zhejiang Province, and the location of Ancient capital, Lin'an City, and Nansong imperial Palace. The area of ancient capital—Lin'an City touched the foot of Mantou hill in east, while linking to Fenghuang hill in west, Wansongling road in north and Songcheng road in south, covered the land area about 500,000 square meters. Nansong Imperial Palace located at the south part of ancient capital, which had its own unique urban fabric and a number of historical landmarks.

Table 154.1 Coding of indicators for VR environment information collection

Main code	Main indicator	Sub-code	Sub-indicator
1.	Environment	1.1	Urban environmental factors
		1.2	Inner environment
2.	Leisure	2.1	Attractiveness
		2.2	convenience
3.	Basic infrastructure	3.1	Accessibility
		3.2	Inter traffic
		3.3	Congestion and Safety
4.	Services	4.1	Lodging facilities
		4.2	Food and beverage
		4.3	Shopping
		4.4	Nightlife
5.	Attractions	5.1	Natural resources
		5.2	Cultural resources

154.3 Method

154.3.1 Indicators Description

In order to determine the ecological suitability in a certain scope [10, 11], it is necessary to collect the basic information of existence, environmental and infrastructure values. VR technology is helpful for well-understanding about travel experience. Heidegger's phenomenon existence theory point shows that the display of VR is realized through computer program [12]. Definitely, VR environment is helpful to get information of travel feeling, assisting to the elimination of the gaps between tourism environment existence and the awareness of tourists. Table 154.1 below depicts the coding of indicator and sub-indicator of travel experience for VR environment information collection.

154.3.2 VR Environment Assessment

The VR model of Nansong Imperial Palace can present a part of heritage area, presenting the circumvallation, building, roads, greenbelt, gates in this area. Participants can enjoy a VR tour to help them know well about the heritage construction design. And then let them to decide the level of their preference of different indicators.

We build the model in Multigen Creator and achieve its roaming in Vega. Techniques for the three-dimensional modeling of virtual environment in Multigen Creator are as described. We used the software like AutoCAD, 3D Max, Vega, and trying to show the model by the graphical interfaces, e.g. Lynx. In addition, Visual

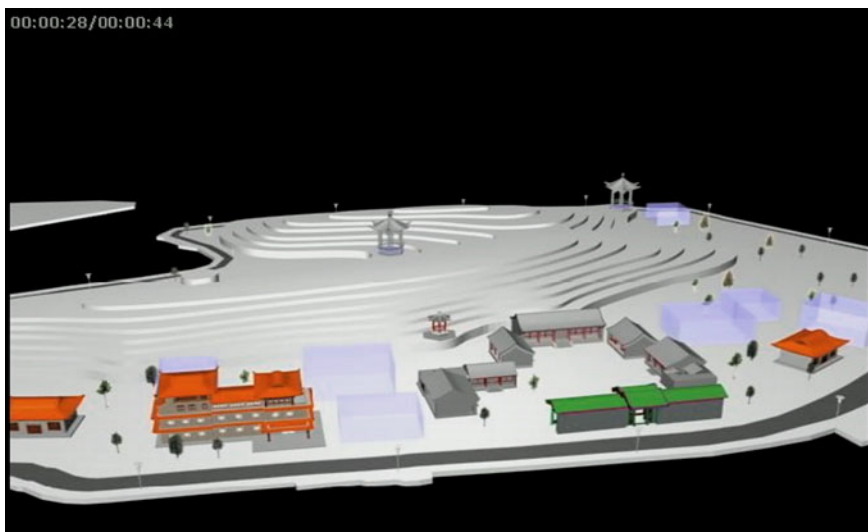


Fig. 154.2 The VR Model of Part of Nansong Imperial Palace

Studio 2005 has improved significantly in the helping information and tools for visiting information. It is very suitable for the creation of efficient, highly interactive terminal equipment applications rapidly. As the choice of program language is very important for the realization of design patterns, Visual Studio 2005 platform developed a good design pattern to achieve the required mechanisms: such as inheritance, package and polymorphism. Figure 154.2 is a screenshot picture of the VR model of Nansong Imperial Palace.

154.3.3 Experience Information of Virtual Reality Platform

Therefore, we invited 64 participants to have imitated tour in VR environment of Nansong Imperial Palace. Inside of these 64 participants, there are 14 young and mid-aged men, 20 young and mid-aged women, 10 students and 20 elder men (Fig. 154.3).

All the participants will take a simulated tour in VR environment of Nansong Imperial Palace. We simulated the records between audio and video information recorded from Participants' VR tour, additional with the questionnaires. As the same judgment of ecological indicator assessment, there are several assessment levels for each sub-indicator in VR environment experience of participants: level A is the best amount to 8–10, level B is normal amount to 5–7, level C is the worst amount to 1–4. When there is more than 2/3 level A in the group of sub-factors, the level of this group is Level A; when there are more than 2/3 level C in

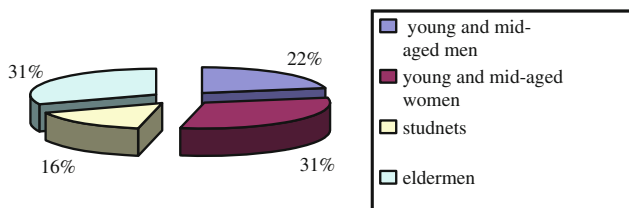


Fig. 154.3 Amount distribution of each participants flock

Table 154.2 Three major indicators assessment by all the participants

Competitive forces	Evaluation	Attractiveness
Leisure factors by different level of ages	C	L
Landscape factors by different level of ages	B	L
Building factors by different level of ages	A	M
Overall assessment on meeting key success factors		L
Overall assessment on gap matching capabilities		M

the group, the level of this group is Level C; the rest is Level B. When the gap between requirements and current performance and experience is more than 3 grades, it comes into +2/-2. When the gap between requirements and current performance and experience is more than 2 grades, it comes into +1/-1. The gap less than 1 grade is omitted.

154.4 Conclusion

154.4.1 VR Assessment

After analyzed all 4 flocks' assessment for the experience in the VR model, we can get the general conclusion of people's requirements, and make the conclusion for STR planning. Table 154.2 shows the final attractiveness of participants in VR model.

Figure 154.4 shows the final evaluation of participants in VR model. The final experience of the Nansong Imperial Palace VR tour assessment shows that the buildings owning a high assessment, landscape is medium level, and leisure is low.

154.4.2 Matrix Comparison

Merging the assessments contained in the two matrixes of ecological and VR evaluation, every sector is placed in its corresponding location in the planning

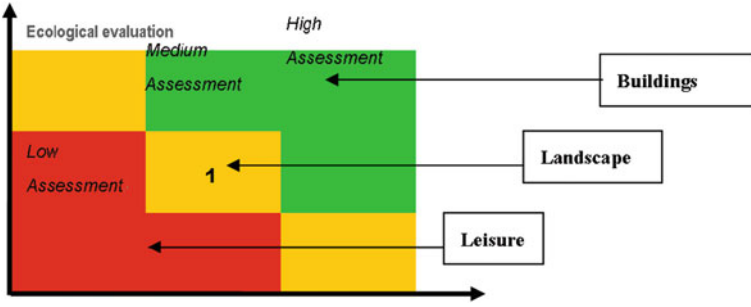


Fig. 154.4 Final matrix value of VR experience assessment

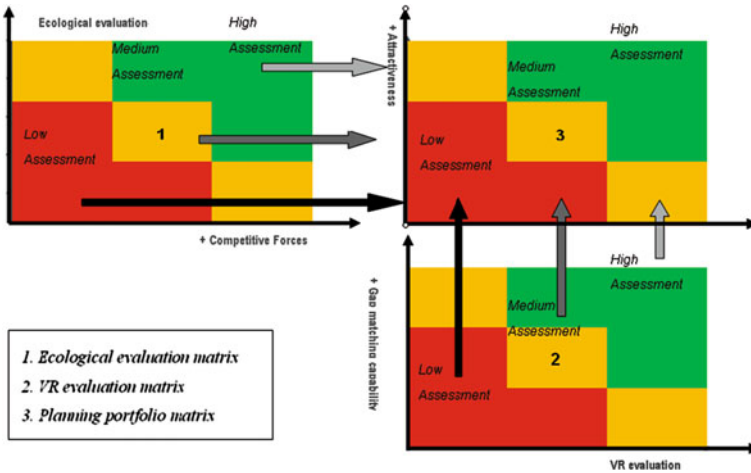


Fig. 154.5 Matrix comparison illustration between ecological and VR assessment

portfolio matrix that defines the final priorities according to the competitiveness and attractiveness of STR planning. Figure 154.5 shows the matrixes comparison illustration between ecological and VR assessment by different indicators.

154.5 Promoted STR Planning

According to the result of matrixes comparison in four different flocks, we make the improved STR planning to satisfy all the potential market’s need.

The protection of the main buildings in Nansong Imperial Palace should be based on the imperial city system according to the time of Gao Emperor and Xiao

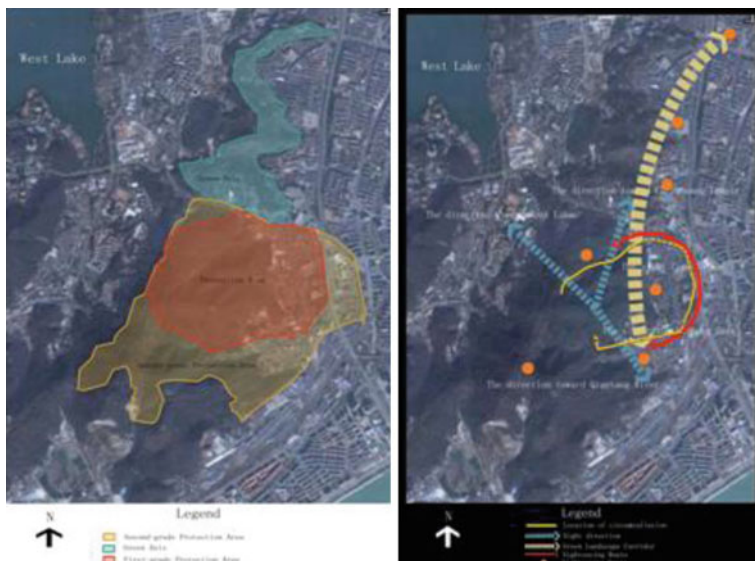


Fig. 154.6 Ecological protection planning focuses area

Emperor. The system sets up two palaces symmetrically at the north and south part of the ancient imperial city with multi-axes. In order to satisfy all the tourists' requirement about the culture attractions, we suggest that the main buildings' reconstruction can be combined with the royal garden (Fig. 154.6). The combination with the city structure with Hangzhou urban city closely, representing the archaeological sections selectively based on the history. By this way, it can improve environmental standards according to the gap of the traffic congestion and accessibility. The establishment of greenbelt based on the terrain and cultural points to avoid the flaky development of the city, and to make contributions to the development of the landscape design, which can remedy the gap of graceful scenery and natural landscape between the heritage area and urban city. According to the conclusion above, there should have three ecological protection planning focuses as following:

Taking care of the relationship between the city axis and the city layout in order to protect the whole landscape style. To combine the heritage and hill area, replenishing graceful landscape new landscape for the West Lake Scenic Spot. To make sure and protect the basic pattern of Nansong Imperial Palace through the road excavations. And then protect and manage the relative sites reasonably. The combination points are the city axis, cultural points, the landscape area and tourism routes. In the future city development level, make it a green axis along the Qiantang River into the city based on the mountain green.

References

1. Naghadehi M, Mikaeil R, Ataei M (2008) The application of fuzzy analytic hierarchy process (FAHP) approach to selection of optimum underground mining method for Jajarm Bauxite Mine, Iran. *Expert Syst Appl* 36(2009):8218–8226
2. Borja A, Bricker SB, Dauer DM, Demetriades NT, Ferreira JG, Forbes AT, Hutchings P (2008) Ecological integrity assessment, ecosystem-based approach, and integrative methodologies: Are these concepts equivalent? *Mar Pollut Bull* 58(3):457–458
3. Zhang Y, Iki K, Homma R (2008) The study on Nansong Dynasty heritage landscape reversion and planning by 3D digital model technique. In: *Proceedings of the conference on Chinese ancient architecture research, Shanghai Vol 1*, pp 86–90
4. Tommy F (2009) The continuity and change in mega-urbanization in Asian cities: a survey of Jakarta Bandung Region (JBR) development. *Habit Int* 33(4):327–339
5. Long H, Zou J, Liu Y (2007) Differentiation of rural development driven by industrialization and urbanization in eastern coastal China. *Habit Int* 33(4):454–462
6. Vouligny É, Domon G, Ruiz J (2006) An assessment of ordinary landscapes by an expert and by its residents: landscape values in areas of intensive agricultural use. *Land Use Policy* 26(4):890–900, October 2009, Pages 947-953
7. Hamblin A (2009) Policy directions for agricultural land use in Australia and other post-industrial economies. *Land Use Policy* 26(4):1195–1204
8. Desideri U, Proietti S, Sdringola P (2008) Solar-powered cooling systems: technical and economic analysis on industrial refrigeration and air-conditioning applications. *Appl Energy* 86(9):1376–1386
9. Denong Z (1997) *Illustrated Chinese architecture history*. China Architecture Press, Beijing
10. Xujie L (1989) *Ancient Chinese architecture history*. History & Culture Press, Tianjing
11. Zhang Y, Iki K, Homma R (2008) The Study on the Urban Landscape planning based on 3D model linked with QTVR technique. In: *Proceeding of the conference on 2008 symposium of academic committee of foreign studyies, Xiamen*, pp 438-446
12. Zhang Y, Iki K, Homma R, Yu W (2008) Study of the evaluation analysis of ningbo tianyi square through apace-time technology 24:87–91

Chapter 155

A Context Information Management Model for Tour Mobile E-Commerce

Szemin Wong and Dongsheng Liu

Abstract With the development of E-commerce, it has become a study focus as how to get mobile service timely and accurately, as well as how to realize the scenario information management of users. In this paper, it takes mobile tourism for example to classify and describe scene information firstly, then extract scene features to construct a quality restriction system and eventually build up a self-guided ontology-based mobile information management model for the scene. It also cites West Lake Self-help guide as an example to verify the validity of above model in this chapter.

Keywords Context-aware · Context modeling · Fuzzy rough sets · Hierarchy ontology

155.1 Introduction

With rapid development of gaining information for service demand anytime and anywhere in E-commerce [1], it has become a study focus as how to satisfy the

S. Wong (✉) · D. Liu
College of Computer Science & Information Engineering,
Zhejiang Gongshang University, Hangzhou, China
e-mail: szemin.w@gmail.com

D. Liu
e-mail: lds1118@zjgsu.edu.cn

D. Liu
Center for Studies of Modern Business, Zhejiang
Gongshang University, Hangzhou, China

dynamic and uncertain demand of users, who'd like to enjoy scene perception in common environment and express themselves when they're moving.

At present, many extensive researches are involved in the perception, modeling, reasoning and application of context information. B. Schlit and Themer [2] discussed the definition of the context, Dey [3] have done a system research related to the definition of context, which laid a basis on its application. On the context modeling, Strang.T and Linnhoff-Popien [4] carried out a comparative study of five context modeling methods, and Razzaquem [5] discussed classification and quality control problems of context information. Henricksen and Indulska [6] studied the molding problem in context management structure, and the modeling method based on location of mobile environment was studied in details by Chen and Kotz [7]. While Tao Gu [8] raised the specific context models after studying the perception and modeling in smart home environment. In these studies, the context modeling which combine with industry-specific is relatively less, and classification and the corresponding quality control is not sufficient consideration to the redundant attribute in certain industry [5]. There is also indistinctness in classification and description in the literature 6, 7 [9]. While the domestic current research focuses on access and reasoning algorithms of context and context modeling are mainly for a single procedure [10–12].

Therefore, we raise a context information management model for tour mobile E-commerce combined with the feature of tour mobile E-commerce. In Sect. 155.2, we introduce readers to the definition about context-aware, ontology, and the theory of fuzzy rough sets. Section 155.3 explains the model design and construction. Finally, in Sect. 155.4, we provide an applied example for model.

155.2 Theory Review

155.2.1 Context-Aware and Ontology

Context-aware, perceiving the status information of a group of entities associated with the user and tasks [13], the aim is to use context to provide user with suitable place, time and task information or services, which is to enable the center of attention to return in the task itself. Since sources of information diversity and complexity, the use of ontology can effectively implement context information sharing and reuse [14, 15], which will formalize the concept of sharing clearly, and implement analysis and reasoning in pervasive environment.

As the core language of ontology, OWL is easy to express meaning and semantic [16]. It's convenient to achieve semantic interoperate, exchange and sharing between different systems, through the entity type (the class) to describe the concept, and the constraints (such as hierarchy, attribute, etc.) to clearly describe the meaning and semantics. Therefore, combining the characteristics and needs of tour mobile E-commerce, we use OWL DL as a model description language.

155.2.2 Fuzzy Rough Sets

Rough set theory can dig out the information hidden in the knowledge in the form of rules [17]. Suppose U and W as limited field of non-doctrinaire, R is the fuzzy relationship from W to U , the triples (U, W, R) is called general fuzzy approximation space. Arbitrary $X \in 2^W$, the lower approximation $\underline{R}(X)$ and upper approximation $\overline{R}(X)$ of X on the approximation space (U, W, R) is the fuzzy sets of U , the functions which they are subject to are defined as:

for arbitrary $x \in U$,

$$\underline{R}(X)(x) = \bigwedge_{y \in W} [(1 - R(x, y)) \vee X(y)] = \min\{1 - R(x, y) | y \notin X\}, x \in U$$

$$\overline{R}(X)(x) = \bigvee_{y \in W} [(1 - R(x, y)) \wedge X(y)] = \max\{R(x, y) | y \notin X\}, x \in U$$

$(\underline{R}(X), \overline{R}(X))$ is general fuzzy rough sets [18].

The sources of context information in tour mobile E-commerce are complex and varied; while properties of the context information are equally important, and even some are redundancies. To maintain the context information classification ability staying the same with deleting unimportant or redundant properties, and retain the important properties related to user preferences, for which we adopt the fuzzy rough set theory to realize the context information classification.

155.3 Building the Context Information Management Model Based on Ontology

The context information management model for tour mobile E-commerce get the context information from context providers and context reason engines. Then on the one hand we determine the natural properties of information based on the acquired way; the other hand, it construct the quality constraint model based on soft constraint in satisfactory optimization theory, which can find out the basic information and useful information. Finally, the fuzzy rough set theory was proposed to achieve ontology classification of information, where the type of information category was identified by mutual information values, and ultimately we build the multilevel model of context sub-ontology. The specific steps are shown in Fig. 155.1:

155.3.1 Classification and Description of Context Information

After obtaining context information, we determined the natural properties of information based on the acquired way for acquiring the systems representation and unified naming convention of information, which can make the context model describing the different entities will have semantic interoperability.

We divided the context information into two main categories, direct context and indirect context, shown in Fig. 155.2. Direct context can be obtained from the

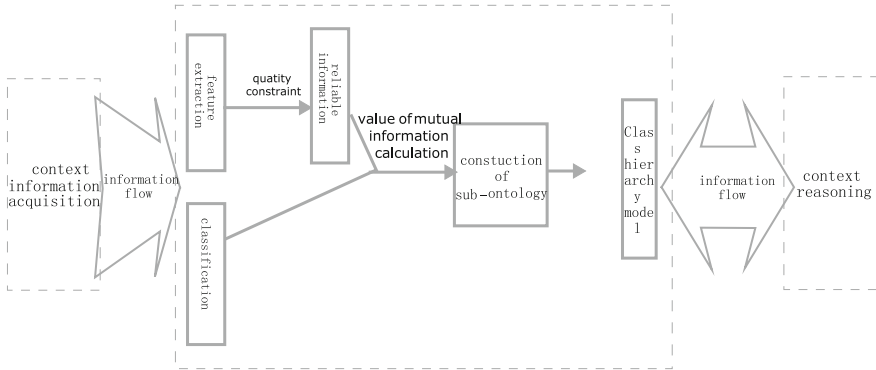
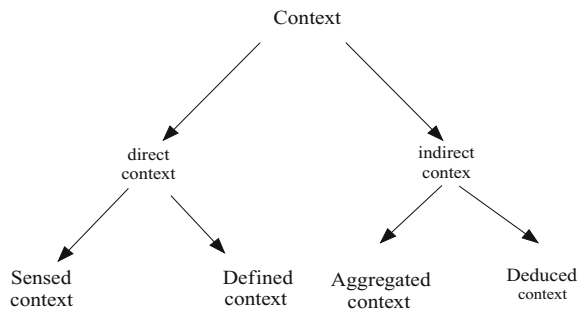


Fig. 155.1 Flow chart of modeling

Fig. 155.2 Classification of context information

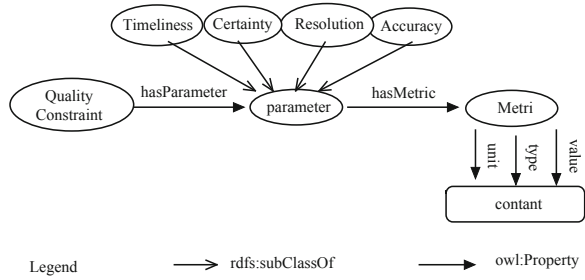


source provider directly, and it can be can be further classified into sensed context from physical sensors and defined context which is typically defined by a user. Indirect context is obtained by integration and reasoning of direct context.

To describe context classification information in our context, we introduced property element—owl: classified as in property restrictions. This element can capture related properties about data types and object in the context classification. The values of the property element are Sensed, Defined, Aggregated or Deduced. As follows, we described the classified information—Defined on the Object Property—M transportation, which is the limited information about means of transport based on the spot position.

```
<owl:Class rdf:ID="sspots">
<owl:Restriction>
<owl:onProperty rdf:resource=" Mtransportation"/>
<owl:to Class rdf:resource="# Scenery-spot">
<owl:classifiedAsrdf:resource="http://www.zjgsu.edu/owl/
classification# Deduced"/>
</owl:Restriction>
</owl:Class>
```

Fig. 155.3 Ontology-orient model of quality constraint



155.3.2 Quality Constraint of Context Information

Considering highly dynamic nature of pervasive computing systems and imperfect sensing technology led to information inconsistency at different levels, we constructed quality constraint, an extensible ontology for quality of information, as indicators describing authenticity of context information.

Quality constraint is associated with a number of quality parameters, which capture the dimensions of quality relevant to the attributes of entities and relationships between entities. Each parameter is described by one or more appropriate quality metrics, which defines how to measure or compute context quality with respect to the parameter. Besides a value, a metric contains a type and a unit.

For different type information focused on the difference, and fuzzyness can not be ignored in the defined information, we have defined four types of quality parameters, accuracy, resolution, certainty and timeliness of the four types of parameters as the basic quality criteria of context information .

Shown in Fig. 155.3, we have defined four types of quality parameters: accuracy-range in terms of a measurement, which is different calculation indoor and outdoor; resolution-smallest perceivable element; certainty—the probability to describe the state of being certain and timeliness—production time and average lifetime of a measurement. The resolution is positive correlated with accuracy; resolution, accuracy is to some extent negative correlated with timeliness; certainty is associated with the classification property of the information.

Quality constraint is a set of parameters which is used to judge the credibility of the information. Considering ambiguity information and possibility of the parameters boundary adjustment, using soft constraint build the constraint condition of the quality of information [19], the steps:

Step 1: The number of performance index of information quality is n, and P_j^{\min} is the lower limit of j th index, and P_j^{\max} is the upper limit, and the information quality constraint is:

$$P_j^{\min} \leq f_j(x_1, x_2, \dots, x_k) \leq P_j^{\max} \quad , j = 1, 2, 3, \dots, n \quad (155.1)$$

$f_j(x_1, x_2, \dots, x_k)$ is the linear function which describes the relationship between j th indicator and other indicators, it is signifying the value of proportion, a certain index in overall indexes

Step 2: Introducing n logical variables $\delta_{(m)}^{\min}, \delta_{(m)}^{\max}$ and intermediate variables $\mu_{(m)}^{\min}, \mu_{(m)}^{\max}$:

$$P_{(m)}^{\min}(1 - \delta_{(m)}^{\min}) + \delta_{(m)}^{\min} \cdot \mu_{(m)}^{\min} \leq f_{(m)}(x) \leq P_{(m)}^{\max}(1 - \delta_{(m)}^{\max}) + \delta_{(m)}^{\max} \cdot \mu_{(m)}^{\max} \quad (155.2)$$

Step 3: Seeking minimum and maximum of adjusted constraint:

$$\begin{aligned} \min f_{(d)}(X) &= f_{(d)}(x_1, x_2, \dots, x_k) \\ s.t. \begin{cases} P_{(M)}^{\min}(1 - \delta_{(M)}^{\min}) + \delta_{(M)}^{\min} \cdot \mu_{(M)}^{\min} \leq f_{(M)}(x) \\ f_{(M)}(x) \frac{\max}{(M)}(1 - \delta_{(M)}^{\max}) + \delta_{(M)}^{\max} \cdot \mu_{(M)}^{\max} \\ M = \{m | m < d\} \end{cases} \end{aligned} \quad (155.3)$$

Step 4: Initializing parameters and computing after assigning values to $\delta_{(d)}^{\min}, \delta_{(d)}^{\max}$.

155.3.3 Building the Multilevel Ontology of Model

With the complexity of context information and context ontology considered, we divided the context ontology into Person sub-ontology, Scenery-spot sub-ontology, and S-content sub-ontology based on trigonal structure, then delaminate category feature of sub-ontology and use the reliable information constructing it. The key is determining value of mutual information, and we calculated it by formula [20]:

$$M(t_i) = \sum_{j=1}^s K_{t_i j} \cdot p(c_j) \left| \log \frac{p(t_i | c_j)}{p(t_i)} \right| = \sum_{j=1}^s \frac{\max\{n_{t_k(i)}\}}{n_{t_k(j)}} \cdot p(c_j) \left| \log \frac{p(t_i | c_j)}{p(t_i)} \right| \quad (155.4)$$

t_i is the feature item of information, and the occurrence frequency of t_i and category c_j is $p(t_i, c_j)$. $p(t_i)$ signify the frequency which t_i occurs in the information, and $p(c_j)$ is the probability that the information belongs to category c_j . $K_{t_k j}$ is modifying factor, which means the tendency that t_k vary with the numbers of information in category c_j .

As determined sub-ontology, we built the sub-ontology by vocabularies provided by ontology-based domain. Context ontology defines a common vocabulary to share context information in a pervasive computing domain; and include machine-interpretable definitions of basic concepts in the domain and relations among them.

The context ontology should be able to capture all the characteristics of context information. As the pervasive computing domain can be divided into a collection of subdomains which would be easy to specify the context in one domain in which

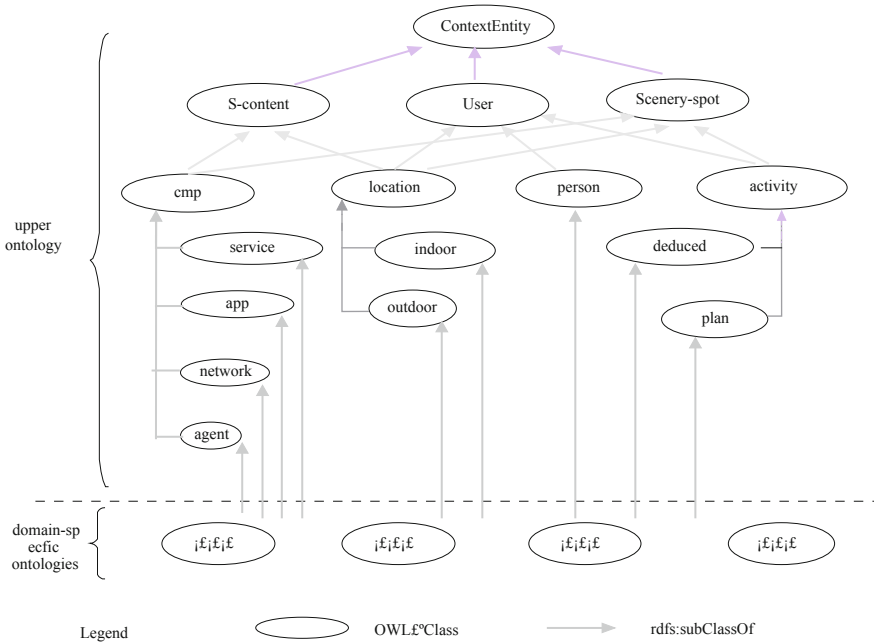


Fig. 155.4 Class hierarchy model for context ontologies

a specific range of context is of interest. Our context ontologies are divided into upper ontology and domain-specific ontologies. The upper ontology is a high-level ontology which captures general context knowledge about the physical world in pervasive computing environments. The domain-specific ontologies are a collection of low-level ontologies which define the details of general concepts and their properties in each subdomain. The low-level ontology in each sub-domain can be dynamically plugged into and unplugged from the upper ontology when the environment is changed. Fig. 155.4

The upper ontology defines the user, location, computing and activity the four basic concepts of entities as shown in Fig. 155.5, and it uses the four basic concepts to build the three sub-ontology. Context ontology provides a reference starting point to the upper ontology and one instance of Context Entity exists for each distinct user, each instance of Context Entity presents User, S-content and Scenery-spot. The details of these basic concepts are defined in the domain specific ontologies which may vary from one domain to another.

155.4 Applied Example and Analysis

We have taken the West Lake of Hangzhou for example to illustrate the application of the model. User M shares information with guide system as well as

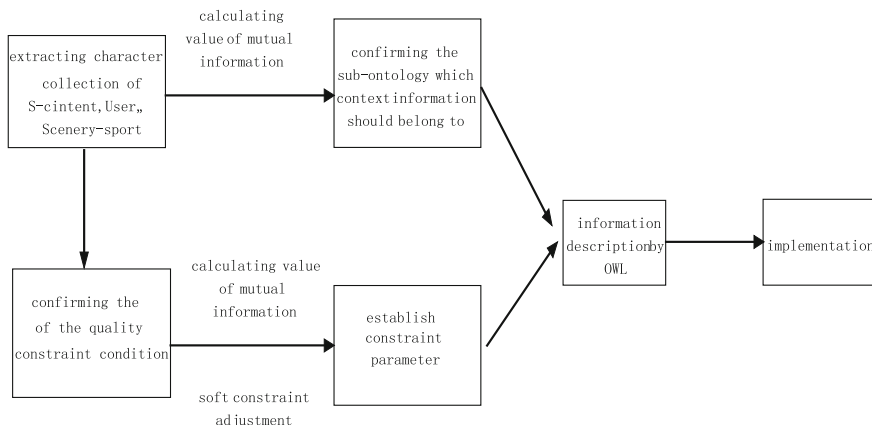


Fig. 155.5 Procedure for implementation

RFID, GPS by PDA. GPS and RFID get the location of users in the outdoor and indoor, and use the wireless network to return the information to the user’s PDA and guide system. Then provide personalized services to users by context reasoning through integrating information in guide system, requirements and preference of user. The implementation process is as follows:

Discernibility matrix extracts knowledge feature characteristic set, which contains time, place, activity as the basic dimension, and different range of dimension weighted value corresponding to different sub-ontology. Finally, the value of mutual information calculated by formula (5) can correspond to relevant context information. The figure below shows the correspondence of three sub-ontology and context information: Table 155.1

By setting the highest priority of different context information, using adjustment methods in Sect. 155.3.2 assigned $\min_{(d)}(X)$ and $\max_{(d)}(X)$, then calculated the range of context information quality constrains.

After determining the relevant information in the mode, we described ontology information by OWL. Then we recommended the best route to users by reasoning based on context KB and RCB. When users reach a scenery spot, the information about surrounding bus, shop, and public facilities will be displayed on their PDA, so does the recommendation based on the plan and preference. Combining with users’ dynamic location information, the recommendation will have timely adjustments.

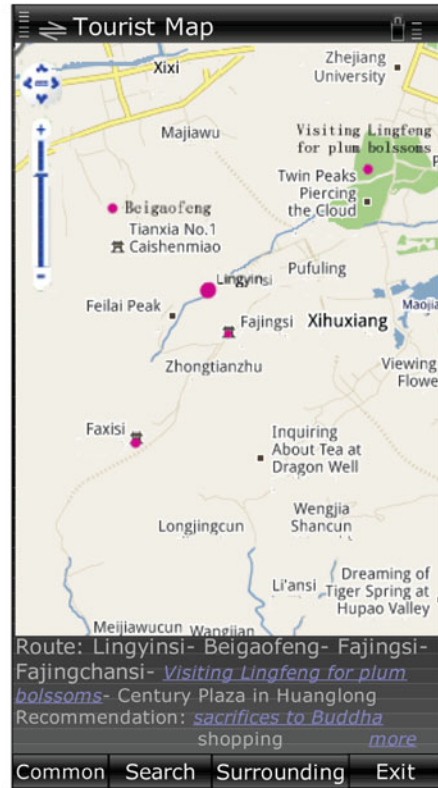
There are Jiuxiyanshu, LongjingCun and LinyinSi in user’s plan, and the integration of information saved in PDA found the user hiking frequently, so put the climbing routes which include Lion Ridge, the north peak as the first recommendation when he was in Jiuxiyanshu, and it also recommended participating in the surrounding activity which hikers like, as shown in Fig. 155.6 (a). But when the user M choose to take bus down the hill in LongjingCun, according to the

Table 155.1 Three sub-ontology and corresponding context information

Name of subclass	Classify of info	Context
Person	Deduced, sensed, defined	Personal information, location, plan, preference
Scenery-spot	Sensed, aggregated	Geographic information, landscape, ancillary facility
S-content	Deduced, sensed, aggregated	Device, network, service, application



(a)



(b)

Fig. 155.6 Mobile navigation

change of user's position, the first recommended route was changed into travel route, and the surrounding recommended activities were also changed into sight-seeing theme, as shown in Fig. 155.6 (b).

References

1. Zimmermann A, Specht M, Lorenz A (2005) Personalization and context management [J]. *User Model User-Adap Inter* (15):275–302
2. Schilit B, Theimer M (1994) Disseminating active map information to mobile hosts [J]. *IEEE Netw* 8(5):222–232
3. Dey A (2000) K. Providing architectural support for building context-aware applications[D]. Georgia Institute of Technology, Atlanta
4. Strang T, Linnhoff2popien C (2008) A context modeling survey[EB/OL]. [2008210202]. <http://media.cs.tsinghua.edu.cn/~qinwj/readings/paper/strang-acmrm04.pdf>
5. Razzaquem A, Dobson S, Nixon P (2008) Categorization and modelling of quality in context information. [EB/OL].[2008210202]. <http://www.csi.ucd.ie/UserFiles/publications/1124274826156.pdf>
6. Henriksen K, Indulska J, Rakotonirainy A (2002) Modeling context information in pervasive computing systems [M]. *Pervas Comput* 2414:79–117
7. Held A, Buchholz S, Schilla. Modeling of context information for pervasive computing applications [C]. In: *Proceedings of SCI 2002/ISAS 2002*
8. Chen GL, Kotzd A (2000) Survey of context-aware mobile computing research [R]. Dartmouth Computer Science Technical Report TR20002381
9. Gu T, Wang XH, Pung HK (2004) An ontology-based context model in intelligent environments [C]. In: *Proceedings of communication networks and distributed systems modeling and simulation conference, 2004*
10. Gong Ruinan, Ning K, Li Q (2009) Business Collaboration Oriented Context Modeling [J]. *Comput Integr Mfg Syst* 9(15):1731–1737
11. Zhang Q, Yong Qi, Hou Y (2006) The research of context awareness activities calculation based on Hidden Markov Model. *J Xi'an Jiaotong Univ* 4(40):398–401
12. Yin L, Miao K, Li Z (2010) The study of dynamic configuration algorithm in context aware computing system [J]. *Huazhong Univ Sci Technol (Nature Science)* 1(38):35–38
13. Gu J (2009) Context-aware Computing. *J Huazhong Univ Sci Technol (Nature Science)* 9(5):1–20
14. Studer R, Benjamins VR, Fensel D (1998) Knowledge engineering: principles and methods [J]. *Data Knowl Eng* 3(25):161–197
15. Thomas R, Grub ER (1993) A translation approach to portable ontology specifications [J]. *Knowl Acquis* 6(5):199–220
16. WN Borst (1997) Construction of Engineering Ontologies [D]. University of Twente, Enschede
17. Pawlak Z (1982) Rough Sets [J]. *Int J Comput Inform Sci* (11):341–356
18. Bodjanova S (1997) Approximation of a fuzzy concepts in decision making. *Fuzzy Sets Syst* 85:23–29
19. Shaoyuan Li, Xi Yugeng (2002) Satisfactory optimization control of complex systems in fuzzy dynamic environment [J]. *Acta Automatica Sinica* 28(3):408–412
20. Yang S, Gu J (2004) Feature selection based on mutual information and redundancy-synergy coefficient [J]. *Zhejiang Univ Sci A* 5(11):1382–1391

Chapter 156

Quantitative Quality Evaluation and Improvement in Incremental Financial Software Development

Bin Xu, Meng Chen, Cun Liu, Juefeng Li, Qiwei Zhu and Aleksander J. Kavs

Abstract Software quality is extremely important in financial software development. Though incremental software development enables high quality releases, efficient quality management is essential to make the best tradeoff between schedule, effort, cost and quality in order to reduce the potential risks in financial systems. In this chapter the authors suggest quantitative quality framework with a set of evaluation, analysis and improvement approaches. Related practice in a global IT corporation shows that the approaches have significant business value in avoiding decision issues.

Keywords Quantitative quality management · Incremental financial software development · Quality evaluation and improvement

B. Xu (✉)
College of Computer Science & Technology, Zhejiang University,
Hangzhou 310027, Zhejiang, China
e-mail: xubin@zjgsu.edu.cn

M. Chen · C. Liu · J. Li · Q. Zhu
State Street Technology, Hangzhou, 310030 Zhejiang, People's Republic of China
e-mail: MChen2@statestreet.com

C. Liu
e-mail: CLiu4@statestreet.com

J. Li
e-mail: JuefengLi@statestreet.com

Q. Zhu
e-mail: Qiwei.Zhu@statestreet.com

A. J. Kavs
State Street Corporation, Boston, MA 02111, USA
e-mail: ajkavs@statestreet.com

156.1 Introduction

Software applications are becoming increasingly large and complex due to the improvement of software development techniques and people more and more rely on them. Software quality is extremely important in financial software in order to handle transactions accurately and reliably, to protect the business privacy and to successfully deal with all kinds of exceptions. Though the delivery time, budget, and resource effort can be negotiated, software quality is the most important criteria for the customer to accept the product and must not be jeopardized. Poor quality in software applications results in loss of revenue, goodwill and user confidence, failure to provide competitive advantage to business and creates additional workload [1].

Incremental development model [2] is supposed to cover requirements gradually and develop the system in steps that accumulate functionality. It allows partial utilization of product, shortens the development time and helps ease the traumatic effect of introducing completely new system all at once [3]. While it takes advantage of flexibility of resource utilization [4], incremental software development model also introduces some problems. First, because every increment is developed and integrated into the application, great effort is required if the previous increments were poorly designed, hard to understand or difficult to modify [5]. Defects undiscovered in previous increments may be mixed and expanded, which may introduce new problems to the following increments and thus increase the effort to fix them [6]. Second, resource turnover requires knowledge sharing effort and further increases the effort in subsequent increments development [7]. Third, more requirement changes are expected in incremental development and if not properly managed or controlled, development will be led into chaos [5, 8]. Quality management [9] after each increment or even during one increment is necessary to solve these problems.

Quality management can be considered to have four main components: quality planning, quality control, quality assurance and quality improvement [9]. It is not only a principle that ensures quality in software products and services but also the mean to control processes. Quality management therefore uses quality assurance and control of processes as well as products to achieve more consistent quality. Quality management is important to companies for high product quality, better customer satisfaction, increased revenues and smaller waste [10]. A survey conducted by original software in 2010 revealed that the importance of quality management has risen. On the other hand, the survey also revealed that most managers are dissatisfied with their current quality management solutions [1].

There are few works that focus on quality management in incremental software development. A method is proposed on how to order requirement in incremental development to improve quality management [11] and Unified Process is chosen as the embedded framework to perform the quality management [12–14] and [15].

Unfortunately, quality is often been ignored when the project is short of schedule or budget. The requirement on a specific release date often overrules

quality objectives, for example reliability. A software product may be delivered on time, but not thoroughly tested or verified before release due to delays. In other words, testing effort is cut off due to schedule and thus poor quality software is released to client [1]. In order to ensure the high quality in the incremental software development, quantitative quality management is essential to measure the quality and the risk of low quality. However, the quality characteristics defined in the ISO/IEC 9126-1 standard [16] is not practical to be measured directly, lower-abstraction attributes of the product should be accessed [17].

In this chapter the authors argue that some smart techniques in project management should be adopted in software quality management domain and suggest a series of practical solutions to evaluate, analyze and improve the software quality in incremental financial software development.

The rest of the chapter is organized as follows. Section 156.2 presents a framework of quantitative quality management and introduces the quantitative quality evaluation approaches related to the framework. The quantitative defects analysis and quality improvement are proposed in Sect. 156.3. Experiences in a global IT software development company are demonstrated in Sect. 156.4. Conclusion and discussion are in Sect. 156.5.

156.2 Quantitative Software Quality Evaluation

156.2.1 Quantitative Quality Management Framework

Quantitative quality management framework is a dynamic framework, which could be improved during the incremental software development practice. It starts at evaluation phase, within which the quality and the quality management will be evaluated with some defined rules and criteria. After the evaluation, some further analysis could be done on increments with potential issues. The rules, criteria and the factors could be calibrated and improved to reduce the total effort in quality management. The whole framework generates a loop which in practice will be overlapped.

156.2.2 Indicate the Possible Quality Issues after Increment

Bug density can be a good metric to measure the quality of an increment by comparing the subsequent increments in an incremental software development project. Generally, a lower bug density means “High” quality, while a higher one means “Low” quality, which will be the subject of later analysis.

In this research, bug density is viewed in another view and used to indicate the possible quality issues after increments. Bug density will be estimated according to

some historic project from the benchmark repository. The estimated bug density will be calibrated with the average value during the progress of the incremental software development. We assume that there are some problems in those extremely “High” or extremely “Low” quality increments. For example, the testing team had much more time to test the increments and found more defects than the average for some “Low” quality increments. For the “High” quality increments, maybe the testing team did not have sufficient time to test the increments and found far too little defects than the average.

In this research we focus on the extremely “High” quality increments. Several criteria are set to indicate the possible quality issues:

- (1) Scale is large but complexity is NOT lower. Line of source code (ESLOC) is used to evaluate the scale and scorecard is used to measure the complexity.
- (2) Test duration is short.
- (3) Product management is highly efficient and does an excellent job for resource allocation and schedule planning.

156.2.3 Determine the Quality Status within Increment

S-Curve is another commonly used metrics to track the daily total defects found and total defects remaining within release. In this research, we introduce this metrics to measure the risk once the increment must be released to clients before the testing period is finished. We use the formula (156.1) to evaluate the risk.

$$\text{Exposure Defects} = \text{Remaining Defects} + \text{Non-found Defects} \quad (156.1)$$

Where $\text{Non-found Defects} = \text{average bug density} * \text{KLOC}$. If the data of exposure defects is closed to zero, the release risk is low, which means the testing can be finished during that period. Otherwise, numbers of Exposure Defects will be exposed to clients so that they could make the decision and tradeoff accordingly.

156.3 Quantitative Software Defect Analysis and Quality Improvement

156.3.1 Quantitative Software Defect Analysis Based on the Summarized Report of Root Cause Analysis

Root cause analysis is very important for the team to find out the weakness of the development and enables them to improve the quality of the increments in the future. A summarized report of root cause analysis could provide some further information for us to improve the quality level.

Definition 1 (*Summarized Report of Root Cause Analysis*) Summarized report of root cause analysis can be defined as a matrix $SRCA :: = \langle RC, N, P \rangle$, where RC is the root cause category, N is the number of the related defects, and P is the price to fix a defect in this category.

Possible root cause category can be “Requirement misunderstanding”, “Requirement unclear”, “Requirement faulty”, “Detail design failure”, and etc.

Definition 2 (*Rule based Improvement Solution*) Some rules can be defined for the quality improvement according to different defect category. $RIS :: = \langle RC, IS, Benefit, Cost \rangle$, where RC is the root cause category, IS is the improvement solution, $Benefit$ is the possible benefit with this solution, and $Cost$ is the investment for this solution.

156.3.2 Quantitative Quality Improvement

The benefit is measured in the unit of defect removal efficiency and the cost refers to the effort to avoid such defects through review or similar techniques. The values of benefit and cost should be benchmarked and calibrated for individual IT projects according to its local historic experience. The possible improvement solution can be as follows.

To reduce defect of Coding:

- Enhance development test
- Involve code review
- New technology training

To reduce defect of Requirement Misunderstood/Unclear:

- Requirement review meeting
- Requirement question list

To reduce defect of Faulty Requirement:

- Involve more clients interactive
- Enhance business analysts' role

To reduce defect of Detail Design:

- Architecture review

To reduce defect of Configuration and Infrastructure

- Early deployment to real environment

Generally, Pareto analysis will be performed to find out the possible defect category to be improved. For each category to be improved, a suitable solution should be decided according to the number, price, related benefit and cost.

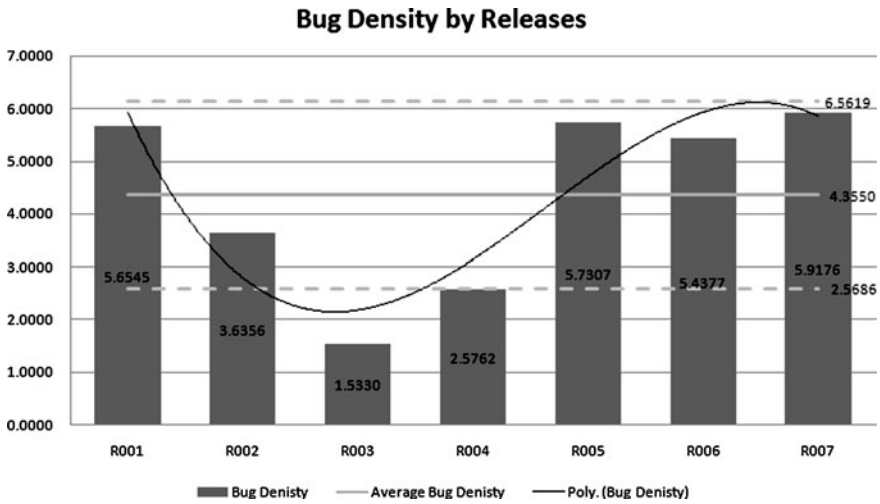


Fig. 156.1 Bug density analysis on the increment (released to client). Since the increments have been released to the clients, the increments can be recognized as releases as well

156.4 Case Study in Global IT Corporation

We applied our approach in a financial software company which provides advanced IT solutions to a top financial organization.

As usual, we collected bug density increment by increment as shown in Fig. 156.1. We found that the increment R003's bug density was fairly lower than the average bug density, even lower than the average minus one sigma. Therefore, we checked the criteria built in Sect. 156.2.2.

- (1) The scale is LARGE but the complexity is NOT lower. The features in the increment R003 were more complicated than the average based on the scorecard from development and testing teams.
- (2) Test duration is very short comparing with other increments.
- (3) As a surprise there was no sufficient product management for R003!

As a result, the increment R003 was indicated to have quality issues: test cycle was not sufficient based on the stable testing resources across the multi-increments and the short testing duration.

S-Curve which has been introduced in Sect. 156.2.3 was used to validate our assumption. S-Curve was generated as in Fig. 156.2 to demonstrate the expected and exposed defects number. There was a big gap between expected and exposed defects number for the increment R003: only half of the expected defects had been found. This was an important evidence to support our assumption.

We verified this assumption with the project team and got the confirmation that the testing for the increment R003 was insufficient. Due to the tight schedule of the

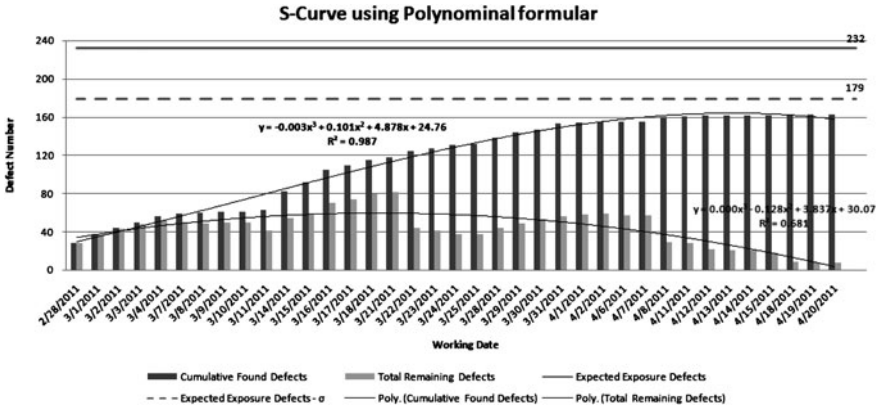


Fig. 156.2 S-Curve analysis on the increment R003. There is large gap between expected and exposed defect in number

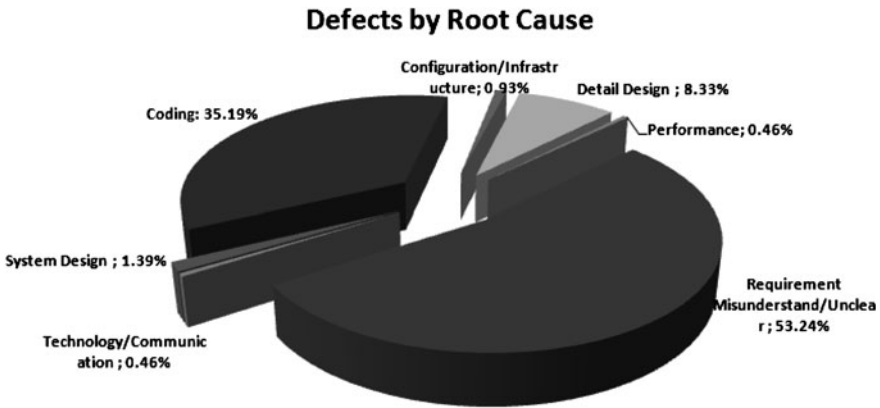


Fig. 156.3 Defects by root cause. Requirement related defects account for more than half in the number

increment R003, testing duration was cut short, which caused more after-release defects. It was really a good example of improper cutting testing efforts under tight schedule as many defects were found on clients’ site.

After we had indicated the increment with high quality issue (though it looked like an extremely “High” quality increment) Root cause analysis was used to clarify the defects in the distribution of root cause, as is shown in Fig. 156.3.

Pareto chart was used to identify the most important root causes that must be eliminated or diminished to improve the quality to an acceptable level. As shown in Fig. 156.4, it is clear that Requirement Misunderstand/Unclear and Coding were the main two causes for the increment R003.

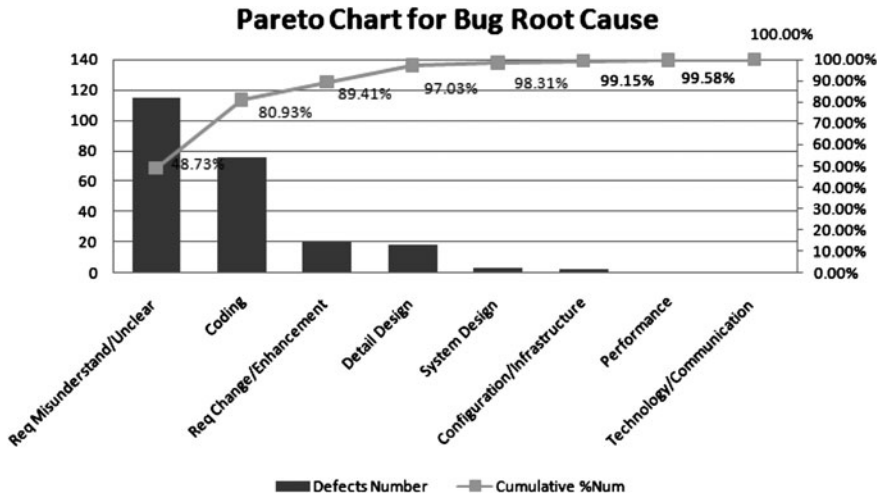


Fig. 156.4 Pareto chart for the root cause report. The most important factors which brought the quality issues were indentified

According to our experience, the possible improvement solutions to reduce the Requirement Misunderstand/Unclear related defects are enhancing the requirement review and establish requirement issue list. We suggested to the project team to enhance the requirement review in the following increments and host additional requirement review meetings for the previous increments to reduce the impact of the quality issue from R003. The decision making for the quality improvement should be made according to the feature of the project and the capability of the project team.

156.5 Conclusion and Discussion

In this chapter we suggested a quantitative quality management framework with a series of analysis and evaluation approaches. It has been used in a real project and the experience showed its value in identifying the quality issue which may not have been easily found manually with some traditional approaches.

The framework and approaches suggested in this chapter can also be used in some other quality-critical incremental software development. The historic project data should be prepared and be used to calibrate the benchmark. This will be the future work of the authors.

Acknowledgments This work is part of “Global Collaborative Software Development” research project, which is an attempt to improve the dual-shore software development with integrated best practice, software engineering technology and project management methodology. The research

project is funded by State Street Corporation, USA. The project is collaboration between Zhejiang University, China, and State Street Corporation, USA. All the company and product names are trademarks or registered trademarks of their respective owners. The contents of this chapter are the opinions and conclusions of the authors only and do not necessarily represent the position of State Street Corporation or its subsidiaries, affiliates, officers, directors or employees.

References

1. Original Software, Application Quality Management Survey Result (2010) http://www.origsoft.com/products/qualify/docs/aqm_survey_results.pdf. Accessed 25 April 2011
2. Mills HD, Dyer M, Linger RC (1987) Cleanroom software engineering. *IEEE Softw* 4(5):19–24
3. Sommerville I (2001) *Software engineering*, 6th edn. Addison-Wesley, Boston
4. Ruhe G (2005) Software release planning. *Handbook software engineering and knowledge engineering*, vol. 3. World Scientific, Singapore, pp 365–394
5. Xu B (2005) Extreme programming for distributed legacy system reengineering. *COMPSAC* 2:160–165
6. Xu B (2010) Cost efficient software review in an E-Business software development project. *ICEE* :2680–2683
7. Xu B, Pan XP (2006) Optimizing dual-shore SQA resource and activities in offshore outsourced software projects. *CCECE* :2405–2409
8. Xu B, Yang X, He Z, Srinivasa RM (2004) Achieving high quality in outsourcing reengineering projects throughout extreme programming. *SMC* 3:2131–2136
9. Rose KH (2005) *Project quality management: why, what and how*. J.Ross Publishing, Fort Lauderdale
10. Ahire SL (1997) Management science—total quality management interfaces: an integrative framework. *Interfaces* 27(6):91–105
11. Wohlin C (1994) Managing software quality through incremental development and certification. *Building quality into software*. Computational Mechanics Publications, Southampton, pp 187–202
12. Booch G, Jacobson I, Rumbaugh J (1999) *The unified software development process*. Addison-Wesley, Reading
13. Nørbjerg J (2002) Managing incremental development: combining flexibility and control. *ECIS*, Gdansk, pp 229–239
14. Kroll P, Kruchten P (2003) *The rational unified process made easy: a practitioner’s guide to the RUP*. Addison-Wesley, Boston
15. Shuja Ahmad, Krebs Jochen (2007) *IBM rational unified process reference and certification guide: solution designer*. IBM Press, New York
16. ISO/IEC 9126-1:2001 (2001) *Software engineering—product quality—part 1: quality model*
17. ISO/IEC 14598-5:1998 (1998) *International standard, information technology—software product evaluation—part 5: process for evaluators*

Chapter 157

Study on Internet Drug Market Access Management

Ying Lu, Bin Dun, Feng Chen and Ming Lu

Abstract As the Internet drug market develops, strengthening the Internet drug market access management is very important in order to effectively protect people's health and safety. This paper analyzes and discusses the current research status of Internet drug market access management, then submits an effective access management solution, and provides the general framework and specific work flow. Meanwhile, this paper makes the discusses on the further improvement on Internet drugs market access conditions, registration for the basic information of market body, security review, and implement of continuous monitoring strategy. The paper realizes effective Internet drug market access management, eliminates the possibility of illegal market body' access to the Internet drug market and establishes strong security foundation for the drug market.

Keywords Internet drug market body · Access management · Security review · Interface adapter

157.1 Introduction

As the rapid development of Internet and information technology, the method of using Internet operation and management mode to operate drug market has become increasingly significant to the pharmaceutical industry, and has brought

Y. Lu (✉) · B. Dun · F. Chen
Information Center of State Food and Medicine Administration, 100053 Beijing, China
e-mail: ying-lu2010@163.com

M. Lu
Beijing ITOWNET Cyber Technology Ltd, 100070 Beijing, China

great revolutionary change. On one hand, it realizes the information sharing of all types of drug market body, greatly improves the overall drug market efficiency, and reduces the operational cost of drug enterprises which leads to greater economic and social benefits. On the other hand, it can also provide the scientific and effective decision making for the Governments to achieve strong supervision, thus fake drugs and illegal drug activity can be effectively prevented.

As a special commodity, drugs are directly related to people's health and safety; therefore effective Internet drug market access management is particularly necessary and urgent, and also has played a positive role for guiding the Internet drug market standardization run. However, The Internet drug market is met with many outstanding issues in the development process as follows: the Internet drug market access system is not perfect, market behaviors are disorder, market monitoring methods falls behind, and public services are not in place, and so on. For the above problems, this paper makes further study on the Internet drug market access management, further improves the market access conditions (licensing, digital certificates, etc.), explores new access review technology, establishes the market body basic information system, and realizes the identity registration application and automatic confirmation for the market body who are trying to access to the Internet drug market, so as to eliminate the possibility of illegal market body' access to the Internet drug markets and establish strong security foundation for the drug market [1, 2].

157.2 Status Analysis of Internet Drug Market Body Access

After many years of development, China's drug market access legal system has been established basic drug market access legal institutional framework which is adaptation with the social market economy, but there still exists some outstanding issues such as confusing regulatory system, too strict market access conditions, complex programs, a confusion of access order, unreasonable function deployment of supervision situations, a loss for supervision regulation and backward supervision means and so on [3].

The pharmaceutical industry is vigorously pursuing medical e-business model which has brought a great deal of revolution, has been becoming a strong force of the pharmaceutical enterprises marketing transformation and industrial restructuring. Currently a series of laws and regulations have been issued to regulate and supervise medical e-commerce, and also the medical e-commerce business access system was established. As a present more systematic and complete legal regulation for medical e-commerce, "Internet drug transaction services and approval of the provisional regulations" proposes some regulations and requirements about hardware facilities and market access for the Internet drug deal service providers building the medical e-commerce supervision system together with "Internet drug information services management approach" and "medical e-commerce supervision measures". In order to develop the more perfect legal supporting environment

and keep the health and regulation of Internet drug market, we need to establish the related complete regulations and maintain scientific and effective legal system strengthening the governments' efforts to control medical e-commerce [4–6].

At present, China's medical e-commerce market access management has implemented a strict system of dual-license. The enterprises who conducting medical e-commerce business must obtain "Internet drug information service qualification certificate" and "Internet drug dealers service qualification certificate" for drugs online transactions.

As a most important aspect in China's Internet drug market supervisory system, Internet drug market access management has been playing a positive role in guiding the standardization run of Internet drug market. In order to further enhance access management, it is necessary to make more efforts to further improve the relative legal regulations and market access conditions and establish the platform of scientific and effective market body basic information management.

157.3 Solution Designing of Internet Drug Market Body

General Framework Internet drug market body includes information service operator (third-party information service platform), transaction service operator (third-party transactions and service platform), network operation enterprise (drug production and operation enterprise web site), production-oriented enterprise users (drug manufacture enterprises), business-oriented enterprise users (drug operation enterprises), medical institutions (hospitals, clinics, etc.) and individual consumers. Study of Internet drug market access management needs to further improve the Internet drug market access system, perfect the market access conditions of various types of Internet drug market body based on the existing access system, and carry out the basic information management system of the market body. The design of this management program mainly includes the following several aspects: the improvement of market access conditions, registration for the market body basic information, record information security review, building market body basic information management platform. The general framework of the program is as shown in Fig. 157.1.

Specific Work Flow The specific management work flow of the Internet drug market access management program which is proposed above is as follows:

- (1) Register for the market body basic information and submit them to the platform;
- (2) Market body basic information management platform requests online the license database of commercial department, the ID card database of public security department, CA digital certificate database, pharmaceutical license database of drug administration departments to conduct the automatic review for the commercial licenses information and digital certificate information submitted by the market body;

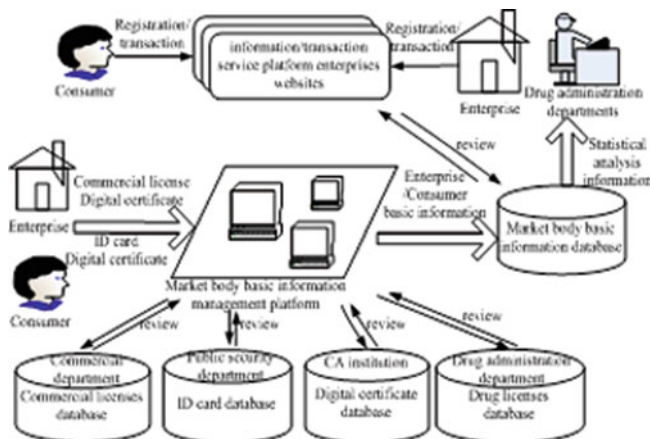


Fig. 157.1 Internet drug market access management general framework

- (3) The review results are returned to the market body basic information management platform, and the system will assign the registration number and change the status of registered information according to all the review results;
- (4) The registration information after certificated is stored into the market body basic information database providing strong data support and decision support for the information sharing of drug market body and statistical analysis of drug administration departments.

The specific work flow is shown as in Fig. 157.2.

Internet Drug Market Body Access Conditions Improvement Despite of achieving some development to some extent in recent years, China's Internet drug market still exists some issues such as not perfect market access regulations and the market disorder just because of inadequate management. One important reason is that the drug market body access conditions are not been strictly restricted, thus there need some further improvement to form the system. To achieve more effective Internet drug market access management, various market body need to meet the following access conditions before entering Internet drug market and then make the drug production, operations and transactions.

For the drug information and transaction services provider, they must have business licenses, ICP license, Internet drug information and transaction services license and digital certificates; For the drug manufacturers, they must have commercial licenses, ICP license, pharmaceutical production license, the Internet drug information and transaction services license and digital certificates; For the drug operators, they must have business licenses, ICP license, medicine business license, Internet drug information and transaction services license and digital certificates; For the drug operation and production enterprises and medical institutions, they must have digital certificates.

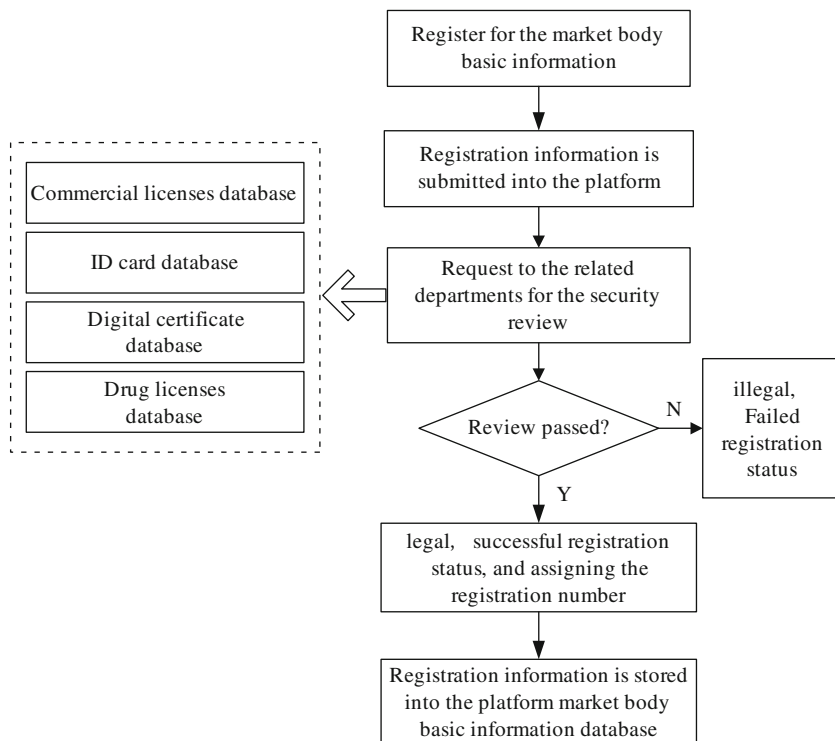


Fig. 157.2 Internet drug market access management work flow

Registration and Security Review In order to achieve more scientific and efficient management of market access, it needs to register for the market body basic information, and carry out the security review, then keep the records within the related market access legal regulations system realizing the safe and effective management for the various market body basic information and further promoting the healthy development of Internet drug market.

The drug registration must be strictly conducted and for the new drugs to market, generics and imported drugs, a rigorous technical review and administrative approval should also be made. In China, only the drugs which have achieved drug approval number or import drug registration certificate (medical product registration card) are allowed to manufacture or sell. All the enterprises which have applied for production and operation must be audited strictly, and the key auditing is focused on staff qualifications, factory environment, equipment and facilities, business place, storage conditions, quality management agencies, etc. Only those who meet the conditions and guarantee the drugs quality strictly can acquire the production and operation license and then carry out the business.

Market body security audit can be conducted automatically for the registered and recorded information which is submitted by the market body. The commercial

registration information and personally identifiable information in the market body registration information are sent to the commercial and public security departments for verification through the interface adapter tools. After all the information is approved, the review status will be changed and sent to the market body registration and recording system, then the registration information will be stored into the unified data warehouse of the management platform.

For the realization of interoperability, information verification and security access control among the commercial registration database, drug production and operation license database, Internet drug information and transaction services license database and the CA database, it is need to apply the visual interface adapter technologies. Through the development of related message format, transmission protocol and some other specifications, we need to establish the visual interface adapter to connect and access for the systems and databases of the government departments (police departments, commercial departments and internal departments of drug administration), third-party trading departments (drug manufacturers web site, drug operators websites, Internet drug information service providers and Internet drug dealers, etc.), third-party agencies (notarial bodies and CA, etc.).

The use of adapter technology can realize the data interaction among the heterogeneous systems or databases, so as to complete the task of data extraction, analysis and transformation. In order to further improve the reusability and loose coupling of adapter systems, Adapters can be quickly built based on the existing developed components. Each data interactive task is completed by specific application adapters. Application adapter is running in the environment of client adapters. Each task is composed by different independent functions, and different components can realize each function respectively, so these components can be combined into a specific application adapter [7, 8].

157.4 Continuous Dynamic Monitoring

For a long time, Internet drug market has been lack of continuous dynamic monitoring mechanisms, and thus some drug production and operation enterprises that have entered the drug market takes this advantage it to produce and operate the defective products and successfully entered the market seriously damaging to people's safety and health. The drug market access should not only include the registration and approval before accessing to the market, but also include the continuous assessment after accessing to the market, and this process should be continued dynamic.

Therefore, for China's drug market access system establishment, we should establish the mechanism which makes a certification and license system as the core and is supported by continuous dynamic supervision, not only focus on the supervision before access, but should be more focused on sustained supervision after access. Once the drug production or operation market body no longer have

the market access conditions, their entry qualifications should be canceled, and shall be ordered to withdraw from the market. Similarly, for the drugs which have already passed the registration and approval, they should accept the continued re-evaluation. Once the drugs appear any adverse reaction or some other condition which is detrimental to the patients, they should be recalled immediately.

157.5 Conclusions

As a special commodity, drugs are directly related with people's health and safety, and with the growing development and maturing of the Internet drug market, it is a top priority to further strengthen the Internet drug market access management. This chapter studies and presents a set of Internet drug market access management solutions, and gives the implementation general framework structure and specific work flow. Meanwhile the paper makes further improvement for the Internet drug market access conditions, and also analyzes and discusses the market body information registration for the records, security review technology and continuous dynamic supervision strategy. However, there are little researches on the Internet drug market access management and not mature enough market information security review technologies, thus it still needs the Government, the community pharmaceutical industry and the public's positive cooperation and efforts to further strengthen the Internet drug market safety and promote its rapid and healthy development.

Acknowledgment This work is supported by the Key Projects in the National Science & Technology Pillar Program during the Eleventh Five-year Plan Period under Grant No.2009BAH48B01.

References

1. Shen M, Wang J, Wu H (2002) The status problems and solutions of the pharmacy e-commerce development in China [J]. *Chin Pharm* 13(7):56–57
2. Shaoran L (2009) Strengthen the supervision of online pharmacies countermeasures [D]. J Shenyang Pharm Univ
3. Weimin Z (2008) Study on drug market access legal system in China [D]. Southwest University of Political Science and Law, Chongqing
4. Yanli W, Suxiang W, Shanshan Z (2009) The status analysis of China's pharmacy e-commerce development [J]. *Jilin Med Colg* 30(4):211–212
5. Huogen C (2009) China's pharmaceutical e-business models—Jane Cheng model of [J]. *Commerce* p 12
6. Bojin W (2005) Government regulation and standardized operation of medical e-commerce [D]. Henan University, Henan
7. Zhai T (2007) Study and implementation of adapter-based heterogeneous system integration technology [D]. Northwestern Polytechnical University, Xi'an
8. Zhong Y, shaowei C, Zhao Y (2007) Research and implementation on a visual developing tool of adapter [J]. *Sci Technol Eng* 7(5):777–781

Chapter 158

Study on the Real-Name System Technology for the Ontology of Internet Medicine Market

Bin Dun, Ying Lu, Feng Chen and Ming Lu

Abstract With the vigorous development of the internet medicine market, the real-name system for market ontology is increasingly becoming the main factor to maintain market stability and development. This chapter describes the current status of the real-name system for internet medicine market and the goals need to achieve, analyses the key technologies for the market ontology real-name system, proposes a security system for market ontology real-name system, at last, designs an implementation process based on the real-name system security system.

Keywords Real-name system · Network security infrastructure · User authentication · Security system

158.1 Introduction

At present the development of the internet medicine market system is imperfect for the problems that market access, regulations and standards is not perfect, market behavior disorders, ways and means of market monitoring backwards and other prominent problems. The market ontology real-name system is increasingly becoming the main method to resolve these issues and maintain the stable development of internet medicine market. Although the internet pharmaceutical

B. Dun (✉) · Y. Lu · F. Chen
Information Center of State Food and Medicine Administration,
Beijing, China
e-mail: dunbin@sfd.gov.cn

M. Lu
Beijing ITOWNET Cyber Technology Ltd., Beijing, China

market is a virtual market, consumers have been closely related with the majority of medicines. Pharmaceutical market on internet not only brings convenience to people's lives, but also brings people of the medicine trade a lot of hidden dangers [1]. So the market real-name system is made subject to the pharmaceutical market development and stability. Market ontology to implement real-name system would help to improve the supervision of the relevant departments of the market [2]. Market ontology to implement real-name system can also guarantee the market's growth. Internet real-name system for the pharmaceutical market will promote the market ontology integrity, law-abiding awareness and promote development of the market into the healthy track. This is necessary for us to build a successful internet pharmaceutical market [3, 4].

158.2 Situation and the Main Goal of Internet Real-Name System for the Pharmaceutical Market

Currently the vast majority of internet authentication is based on the user name and password. Username is set up by the user in the registration. Registration information could be provided false or not necessarily associated with the user. So one person can register multiple accounts and the system could not confirm whether these accounts actually owned by the same person. In this way, network users can use a variety of false identities for various activities. Even if he had a network of criminal activities, the public security organs is difficult to trace [5]. Therefore, it is only used to obtain permission to publish information in cyberspace for network authentication. Users do not need to be responsible for their own basic words and deeds on the network. In this way, users can virtually "Speak out freely" and "Go their own way". This leads to confusion in the network society. If the internet medicine market has such a situation, there would be a serious threat to the interests of all patients, and even lead to serious social problem [6].

The real-name system for the ontology of internet medicine market supports a medicine dealer that before the publishing of information, online trading platform must verify and record their identity information and then allow the user to conduct transactions in cyberspace. With this provision it is practice to regulate trading activities of the network order, crack down on illegal online transactions and to protect the legitimate rights of market participants. This is a mechanism that requires internet medicine dealer must verify the true identity before publishing information in the trading platform. Its purpose is effective in limiting the dissemination of false and harmful medicines information and to combat internet crimes. The real-name system for the ontology of internet medicine market allows law-abiding and reasonable medicine dealers use the internet and medicine transaction platform. But the justice open against unscrupulous traders has long arms. Crime facts can provide accurate evidence.

158.3 Related Technology for Real-Name System

The implementation of internet real-name system for the pharmaceutical market of ontology technical requirements of no more than three elements and the key point is network security infrastructure. Firewall, intrusion detection, virtual private network, security gateways and other safety measures, in combination, in order for the implementation to provide a solid foundation for the real-name system for the ontology of internet medicine market, constitute the whole system security basic safety measures and at different levels to prevent unauthorized access of illegal users of the system, while safeguarding the legitimate authorized users to access. In addition, a good user authentication system and the basis and means for later accountability are indispensable.

Network Security Infrastructure Firewall, Intrusion Detection (ID), Virtual Private Network (VPN), security gateways (SG) and other security measures used in combination will be able to implement real-name system from the general lifting of the security risks. These are the security of the entire system composed of basic security measures, they are at different levels prevents unauthorized users from accessing the system. These objects are authorized user's digital identity.

User Authentication System Commonly used e-commerce authentication solution is: to use cryptography, PKI/CA and ActiveX technology in B/S system architecture, design a set of e-commerce platform based on digital certificate authentication solutions. The real-name system for the ontology of internet medicine market user authentication system can learn from the solution below.

Cryptography in real-name system technology for the ontology Cryptography is the study of the password and code-breaking establishment of technical science. The objective law of password changes, applied to the preparation of a secret password to the conservative communications disciplines, known as coding theory. Used to decipher the password to access to communications intelligence discipline, known as decipher theory. General term for these two disciplines of cryptography. Cryptographic algorithms and security protocols are the two basic parts of cryptography. According to the function to complete the division algorithm, encryption algorithm can be divided into symmetric encryption, public key, digital signature and information summary of the five algorithms. Agreement means the two participants to accomplish a specific task to take a series of orderly steps. The two participants may be complete mutual trust and they also could be the attacker and not trust people. Use of cryptography to complete a specific task and meet the security needs of the protocol is the security protocol, such as message authentication protocols, digital signature protocols.

PKI/CA and digital certificate. Public Key Infrastructure (PKI) refers to the public key infrastructure. Certificate Authority (CA) refers to the certification center. PKI, in terms of technology, solves the network traffic from the security of the obstacles. CA, in terms of the operation, management, norms, laws, officers and other perspective, solves the network of trust. Thus, it is referred to as "PKI/CA". From the overall framework of view, PKI/CA is formed mainly by

end-users, certification and registration center. PKI/CA's working principle is to establish a trust network. The network is a digital certificate by sending and maintenance to achieve. Different user in the same trusted network has a unique number that corresponds with their certificates. The trust network is done through the digital certificate authentication and safe handling. This can be understood as the digital version of a digital certificate ID card, driver's license. In e-commerce activities, when the system requires customers to show identification, the customers must show his digital certificate.

ActiveX controls. Internet trading system in the pharmaceutical market for browser security control development, you can use ActiveX component technology. The main components to complete these functions: 1. It uses digital envelope to package the data that the client sends. 2. Des symmetric encryption algorithm used to encrypt messages sent and digital signatures. 3. Verify the legitimacy of the server digital certificate. ActiveX controls in internet trading system in the medicine market, realization process is as follows: The client makes a request to the server, the server returned to the client a page with embedded ActiveX controls, ActiveX controls from the browser to interpret. According to the ID value in the ActiveX control, browser queries in the registry page of the client computer.

Security measures. As a user authentication system, the basic security measures, through the following methods to ensure system security: 1. Content security. Through the application of a high security level encryption algorithm stored in the user login information, including user login name, login ID number, e-encrypted digital certificate, ensure that even the system administrator could not access or change the information. 2. Transmission security. Information in the transmission process, there may be eavesdropping and tampering.

158.4 Market Ontology Real-name System Certification System

System Architecture. The real-name system for the ontology of internet medicine market security architecture is as shown in Fig. 158.1. Users of the system are divided into monitors and market users. Monitors' permission: 1. Verify the true identity of the market users, and electronic distribution for the market as the main user number, private key and public key market players user authentication system. 2. Verify the true identity of medicines dealers and personal identification related procedures for corporate registration information online pharmacies, and its distribution dealer identification number, private key, public key authentication system dealers and monitoring system public key. 3. To monitor the operation of a dealer name system and impersonate individual users and illegal traders, and to obtain direct evidence. Medicine consumer rights: 1. Get a virtual identity number. 2. Register in the trading platform and speak (Fig. 158.2).

The Specific Implementation Plan. We design a real name system based on real-name system technical requirements, and draw relatively mature mode of operation. Entire operation of the system process is not complicated. Simply put, it can

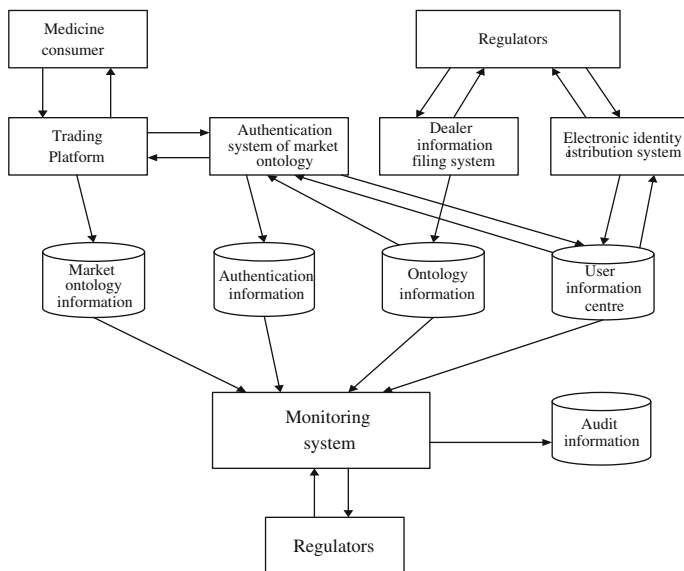


Fig. 158.1 Market ontology real-name system architecture diagram

be roughly divided into the main four blocks: release to market consumer electronic identification number, issuing transaction number as indicated by the individual user to speak website and network monitoring.

Issuing electronic ID number. After regulators get the verification of the true identity of market ontology, they will submit to the electronic distribution system the user ID number and related personal details. Electronic identity system will be distributed the user a unique electronic identification number according to certain rules of operation, followed by the individual user’s electronic identification number in a safe spot. Then the user’s electronic identification number, real personal details and other information will be saved into the personal information database.

Evaluation of medicine consumers. Medicine consumers can register in the traditional way, registration information will be kept in pharmaceuticals trading platform information database. When medicine consumers log into the trading platform, their registered user account, password, authentication code along with other information will be submitted to the trading platform. When medicines consumers want to speak, comment, rate and do other online activities, the trading system will automatically jump to the user authentication system, suggesting that consumers enter their e-pharmaceutical identification number, and answer system about their own information (to verify whether it is the user himself), if all information and personal information consistent with information in the database, it will be medicines and consumer electronic identification number registered

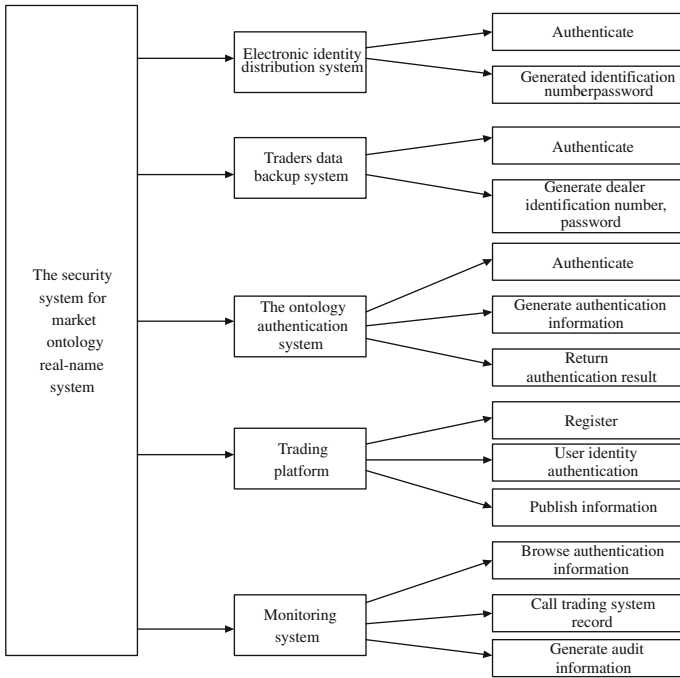


Fig. 158.2 market ontology real-name system security system implementation program

member ID number, and traders into the surveillance system. Each message sent by the consumer, with their corresponding data records contain evidence for the safety of the two fields: a field used by the electronic storage of the user identification number, identification number of speech where the traders, the user’s User ID, login time, and the other field, the true identity of the consumer storage, electronic body copies of the real numbers and other personal information.

Network monitoring. With rigorous authentication process, regulatory authorities can get authentication information by monitoring the system from the database access to records and put them in a variety of data analysis and statistics. If necessary, regulatory authorities can get direct evidence from trading information databases and personal information database and set up the formation of audit information database. The audit information deposited in the database can be used for further processing. Such a real-name system for the ontology of internet medicine market both to keep track of the market in the background, words and deeds of the parties involved, in time to capture illegal acts, and the user can quickly search out personal details in order to deal with public security organs, while also maximizing hide the true identity of the individual user to protect the legitimate users the right to privacy in the network.

158.5 Conclusion

The real-name system for the ontology of internet medicine market can ensure the market in health development. The real-name system can promote the establishment of market integrity and sound development of the internet medicine market. Of course, the real-name system of the target is only through technical means to make the internet medicine market as a natural extension of the pharmaceutical market. The implementation of real-name system also needs to establish the relevant legal system in China. And it needs to make full information to consumers and businesses. Internet medicine market to achieve the real-name system of the market ontology will help speed up the process of building the internet medicine market, it is to achieve the orderly development of the market the only way and it is an effective way to establish credit trading environment. In the long run, China's pharmaceutical market in the internet real-name system does not limit the growth of trade, market development. Rather, it helps to online banking, online trading, online market integrity and other aspects of the building and healthy development. With the real-name system, market ontology will continue to improve. The ontology of the internet medicine market to implement real-name system will no longer be difficult, let us wait and see.

Acknowledgment This work is supported by the Key Projects in the National Science and Technology Pillar Program during the Eleventh Five-year Plan Period under Grant No.2009BAH48B01.

References

1. Da L, Ziqi D (2008) Analysis of the feasible implementation of the internet real-name system [J]. *Yan An Colg Edu* (01):14–18
2. Meng L, Zhou Y, Tong X (2008) Analysis of restrict problems for the development of online pharmacies [J]. *Chin Pharm* (16):095–099
3. Yang J, Li B (2009) China's consumer confidence online pharmacies influencing factors [J]. *Med Soc* (02):113–116
4. Ge L (2008) Pattern of China's pharmaceutical e-commerce applications [D]. Henan University
5. Chen Y, Zou J (2009) NRS: the only way to regulate Internet information dissemination [J]. *Shandong Social Science* (01):232–237
6. Fan D, Lu Q (2010) On the development of online pharmacies and recommendations [J]. *Pharm Affair* (04):151–155

Chapter 159

The Combined Stock Price Prediction Model based on BP Neural Network and Grey Theory

Yulian Fei, Junjun Xiao, Ying Chen and Weiwei Cao

Abstract Stock market is a complex dynamic system, so it is difficult to reflect market with the trait of mole factors, non-linear and time variety by using the traditional time series prediction technology. In view of the failure to a comprehensive analysis and prediction of the problem by a single prediction method, this chapter puts forward the combination of the BP neural network and grey theory to establish a combined stock price prediction model. The Empirical analysis shows that the new model can provide better prediction than the traditional prediction methods in closing price prediction of stock market.

Keywords BP neural network · Grey theory · Combined prediction · Stock price

159.1 Introduction

In the prediction of the financial system, the research of the stock prediction is a hot topic. The stock system of internal properties and prediction research has the important theoretical significance and application value. The stock market has the

Y. Fei (✉) · J. Xiao · Y. Chen · W. Cao
School of Computer Science and Information Engineering, Zhejiang Gongshang
University, HangZhou, Zhejiang, China
e-mail: fyl@mail.zjgsu.edu.cn

J. Xiao
e-mail: cafe0926@126.com

Y. Chen
e-mail: cy860915@126.com

W. Cao
e-mail: caoweimei1987@yahoo.com.cn

characteristic of high income and high-risk, so for maximum interests, the internal rules of stock market have been explored in order to find effective forecasting methods and tools. The research of Stock Market Analysis and Forecast has been an issue, the traditional statistical methods can predict the general trend of stock changes in a period of time, but stock investors are often more interested in short-term ups and downs of stock index. In recent years, the development of computer technology and artificial intelligence technology provides many new methods and opportunities for modeling and forecasting of the stock market [1]. Currently, the gray theory, neural network algorithm was used in a wide range of stock market forecasts. Document [2] established the BP nerve network model, which carried on the short-term forecast to the stock price based on the training of the network by taking some stock actual closing price as the primary data sample, but this model convergence rate is slow and the forecast precision is low. Tan [3] established GM(1, 1) predicting model which can better predict short-term stock price. Ye [4] put forward a new scheme of knowledge encoding in a fuzzy neural networks using rough set theoretical concepts, which can simply the neural network topologies and shorten the time, but its forecast prediction accuracy is general. Deng [5] used prediction method combined by global optimization of GA and local optimization of BP neural network to improve the calculation accuracy and convergence rate of the traditional BP neural network. That method had a high accuracy, but its model was relatively more complex.

In view of the above questions, this chapter adopts a combined prediction method composed of BP neural network and grey theory to determine the weight of each prediction method through the combined prediction model and uses the optimal utilization of the weight as the importance of the various prediction methods to combined prediction. This method can overcome many of the limitations and difficulties of a traditional single prediction method, but also can avoid the influences of many human factors. It has its own unique advantages in terms of rationality and applicability in the stock prediction model.

159.2 The Combined Prediction Model based on BP–GM (1, N)

159.2.1 Combined prediction

In the actual forecast, the same problem can have many different forecasting methods. The different forecasting methods can provide useful information in different aspects; the prediction accuracy is often different. If only according to the prediction precision, some useful information will be lost by giving up some certain methods. So this chapter puts forward the concept of combined forecast [6, 7], which is also called a comprehensive prediction or a combination of forecasts. The combined forecast will give different weights to a number of prediction methods to form a comprehensive predictive model. The combined prediction

theory thinks that the same target can be predicted from different ways systematically. Based on the maximum use of the information provided by various methods, prediction accuracy should be improved as far as possible. Combined prediction model is more comprehensive, more scientific and therefore more effective than a single model system, so it can reduce the impact of some environmental factors caused by a single model prediction.

The key of combined prediction is to how to determine weights of various forecasting methods in the combined forecasting model. Based on different combinations of the determination of weight, the combined prediction can be divided into two categories: constant weight combination forecasting method and the variable weight combination forecasting method.

The basic idea of constant the weight change is: uses the past-time forecasting error as the objective function to find the weight of the various forecasting methods, and then finds the optimal use of the weights as the importance of the various forecasting methods to combine forecasts. Variable weight combination forecasting method is changing over time and the prediction model weights are changing constantly. Every moment of the weight of each method is determined by the error size from the recent individual forecast model. There are many Constant weight combination forecasting methods, such as the optimal combination forecasting method, such as the weighted average method, the variance-covariance method, among which the optimal combination forecasting method is widely used. The model is as follows:

Set the real observation in financial forecast problems during a period as $y(t)(t = 1, 2 \cdots m)$. This problem has P kind of forecasting methods. The i kind of method for prediction is $y_i(t)(i = 1, 2 \cdots p, t = 1, 2 \cdots m)$, and then combined prediction model can be expressed as:

$$y'(t) = \sum_{i=1}^p w_i y_i(t) \quad (159.1)$$

which $y'(t)$ is predictive value of combination forecasting model; w_i is the coefficient of the i kind of prediction methods, $\sum_{i=1}^p w_i = 1$

The i kind of prediction error is:

$$e_i(t) = y_i(t) - y(t)(i = 1, 2 \cdots p, t = 1, 2 \cdots m) \quad (159.2)$$

Combined prediction model error is:

$$\begin{aligned} e(t) = y'(t) - y(t) &= \sum_{i=1}^p w_i y_i(t) - y(t) = \sum_{i=1}^p w_i (y_i(t) - y(t)) = \sum_{i=1}^p w_i e_i(t) \\ &= (w_1, w_2, \dots, w_p) (e_1(t), e_2(t), \dots, e_p(t))^T \end{aligned} \quad (159.3)$$

Then:

$$\begin{aligned}
 e^2(t) &= \left((w_1, w_2 \dots w_p) (e_1(t), e_2(t) \dots e_p(t))^T \right)^2 \\
 &= \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \end{pmatrix}^T \begin{pmatrix} e_1(t) \\ e_2(t) \\ \vdots \\ e_p(t) \end{pmatrix} \begin{pmatrix} e_1(t) \\ e_2(t) \\ \vdots \\ e_p(t) \end{pmatrix}^T \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \end{pmatrix} = W^T E(t) W \tag{159.4}
 \end{aligned}$$

which $W = (w_1, w_2 \dots w_p)^T, E(t) = \begin{pmatrix} e_1^2(t) & e_1(t)e_2(t) & \dots & e_1(t)e_p(t) \\ e_2(t)e_1(t) & e_2^2(t) & \dots & e_2(t)e_p(t) \\ \vdots & \vdots & \ddots & \vdots \\ e_p(t)e_1(t) & e_p(t)e_2(t) & \dots & e_p^2(t) \end{pmatrix}$

Make combined prediction model for the error squares was:

$$J = \sum_{t=1}^m e^2(t) = W^T \sum_{t=1}^m E(t) W = W^T E W \tag{159.5}$$

which $E = \begin{pmatrix} \sum_{t=1}^m e_1^2(t) & \sum_{t=1}^m e_1(t)e_2(t) & \dots & \sum_{t=1}^m e_1(t)e_p(t) \\ \sum_{t=1}^m e_2(t)e_1(t) & \sum_{t=1}^m e_2^2(t) & \dots & \sum_{t=1}^m e_2(t)e_p(t) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{t=1}^m e_p(t)e_1(t) & \sum_{t=1}^m e_p(t)e_2(t) & \dots & \sum_{t=1}^m e_p^2(t) \end{pmatrix}$.

Optimal weighting solution model for:

$$\min J = W^T E W \tag{159.6}$$

$$W^T R_p = 1$$

which $R_p = (1, 1 \dots 1)^T$ is a p-dimensional column vector, and the elements are assigned 1. Optimal solution of the model is:

$$W = \frac{E^{-1} R_p}{R_p^T E^{-1} R_p} \tag{159.7}$$

159.2.2 The Combined Stock Price Prediction Model

Closing price of the stock is a comprehensive index, which is affected by many factors, such as yesterday’s closing price, K index, D indicators, RSI and so on. Under certain conditions, it has correlation between stock closing price and stock

parameters significance. Accurate prediction of the price of the stock has important implications for the development of enterprises. Some scholars have already studied in this field, but they all only use a method to predict the price, and cannot give a comprehensive, in-depth analysis to the relationship between the parameters and the stock prices. BP neural network has a strong nonlinear mapping capability, and GM (1, N) model prediction accuracy is high, so with the concept of the combined prediction, this chapter puts forward a stock price prediction model based on BP-GM (1, N).

Set a stock has n main parameters: $X_i(i = 1, 2 \cdots n)$. Stock closing price is Y . There are m groups of observations: $x_i(t), (i = 1, 2 \cdots n, t = 1, 2 \cdots m)$, Prediction model of the stock price process based on BP-GM (1, N):

- (1) m groups of observations constitute matrix:

$$E = \begin{pmatrix} y(1) & y(2) & \cdots & y(m) \\ x_1(1) & x_1(2) & \cdots & x_1(m) \\ \vdots & \vdots & \ddots & \vdots \\ x_n(1) & x_n(2) & \cdots & x_n(m) \end{pmatrix} \tag{159.8}$$

- (2) Followed by the column vectors $(y(t), x_1(t), x_2(t) \cdots x_n(t))^T$ matrix E , the input of each BP network training, get output value $y_1(t)$.
- (3) GM(1,N) model constructs the prediction equation of matrix E , and obtains the predictive value $y(t)$, set $y_2(t)$.
- (4) Building the combination forecasting model with $y_1(t)$ and $y_2(t)$.

$$y'(t) = w_1y_1(t) + w_2y_2(t) \tag{159.9}$$

With the squared error and minimum for target function, solving w_1, w_2 .

If the existing set of observations $(x_1(m + 1), x_2(m + 1) \cdots x_n(m + 1))$ to predict the corresponding stock prices. Input $(x_1(m + 1), x_2(m + 1) \cdots x_n(m + 1))$ into the trained BP network, get output value $y_1(m + 1)$; use GM equation to predict the stock price, the value is $y_2(m + 1)$; Input $y_1(m + 1)$ and $y_2(m + 1)$ into the combined prediction model, get the final predictive value: $y'(m + 1)$.

159.3 Example Analyses

The method is now applied to forecast future trends of stock closing price of listed company Jiuzhou Electric by its stock data and related economic, technical indexes of historical data. Stock market forecast is essentially a nonlinear function value estimation and extrapolation. Here we select the Jiuzhou Electric yesterday's closing price, the day before yesterday's closing price, K index, D indicators, RSI, DIF, DEA samples of the seven indicators as the original input. Gathering the time interval from 11.04 to 12.31 in 2010, there are 40 sets of data. On this basis, the

Table 159.1 Original data of each index variable

Motor no.	Closing price	Yesterday's closing price	The day before yesterday's closing price	K index	D indicators	RSI	DIF	DEA
1	20.72	20.68	20.42	36.35	52.18	47.48	0.32	0.31
2	22.5	20.72	20.68	54.07	52.81	67.42	0.43	0.34
3	21.94	22.5	20.72	58.9	54.84	58.97	0.46	0.36
4	22.1	21.97	22.5	60.74	56.8	60.67	0.5	0.39
5	21.2	22.1	21.94	52.25	55.28	47.46	0.45	0.4
6	23.32	21.2	22.1	68.16	59.58	67.47	0.57	0.43
7	22.64	23.32	21.2	64.48	61.21	58.84	0.61	0.47
8	22.5	22.64	23.32	60.97	61.13	57.04	0.62	0.5
9	23.08	22.5	22.64	62.22	61.5	62.72	0.67	0.53
10	25.1	23.08	22.5	73.26	65.42	75.98	0.86	0.6
...
36	20.05	21.2	22.25	18.9	35.9	18.76	-0.22	0.05
37	19.4	20.05	21.2	14.32	28.71	15.54	-0.41	-0.04
38	19.56	19.4	20.05	12.36	23.26	19.61	-0.55	-0.14
39	19.33	19.56	19.4	10.53	19.02	18.11	-0.67	-0.25
40	19.86	19.33	19.56	13.78	17.27	32.45	-0.71	-0.34

Table 159.2 The result of the closing price prediction by BP-GM (1, N) and BP-GM (1, N)

Motor no.	Closing price	BP prediction	GM(1,N) prediction	BP-GM(1,N) prediction	BP relative error (%)	GM(1,N) relative error (%)	BP-GM(1,N) relative error (%)
1	23.08	23.0972	22.9146	22.9956	0.07	0.71	0.37
2	22.9	22.9443	22.8422	22.8875	0.19	0.25	0.05
3	22.4	21.9431	22.2785	22.1297	2.04	0.54	1.21
4	22.25	21.3502	22.2710	21.8625	4.04	0.09	1.74
5	21.2	20.4481	21.4783	21.0213	3.55	1.31	0.84
6	20.05	20.0395	20.7040	20.4092	0.05	3.26	1.79
7	19.4	19.7647	19.9864	19.5880	1.88	3.02	0.97
8	19.56	19.6561	19.6068	19.6287	0.49	0.23	0.35
9	19.33	19.5843	18.8466	19.1739	1.32	2.5	0.81
10	19.86	19.7268	19.3033	19.4912	0.67	2.8	1.85
Average change	/	/	/	/	1.43	1.47	0.998

first 30 sets of data as training samples to predict the remaining 10 groups of data. The original input data in Table 159.1.

Through the BP-GM(1, N) establish the closing price and index parameters prediction model, meanwhile individual BP prediction and GM(1, N) prediction. Operation by Matlab software, the results is shown in Table 159.2.

Fig. 159.1 Compare between actual results and predicted results

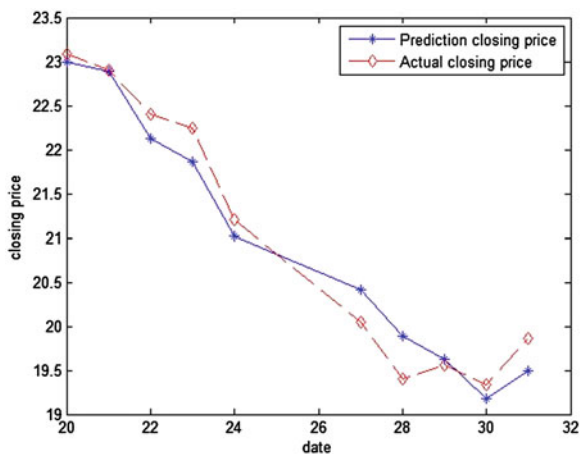


Table 159.2 gives the comparison results, the results of stock price prediction by BP–GM (1, N) are better than BP or GM (1, N) alone. Actual results and predicted results compare in Fig. 159.1.

159.4 Conclusions

The chapter establishes up a BP–GM (1, N) model for stock price prediction. In view of the failure to a comprehensive analysis and prediction of the problem by a single prediction method, this chapter puts forward the combination of the BP neural network and grey theory to establish a combined stock price prediction model. According to the relationship between previous closing stock price and the major technical indicators, the model predicts the closing price of existing stock.. Empirical analysis shows that the new BP–GM(1, N) model of stock price prediction can provide better prediction than the traditional prediction methods. The model can improve the prediction accuracy and low computational complexity and set up all effective analytical systems for vast investors with a practical value.

Acknowledgment This chapter is supported by The 10th College Student “Hope Cup” Extracurricular Work of Technology Competition, Zhejiang Gongshang University (No. B44). The authors thank them heartedly for supporting the paper funds.

References

1. Zhang X, Xu L (2003) The stock market forecast model based on neural network ensemble.[J] Syst Eng Theory Pract 23(9):67–70
2. Gao Q (2007) The application of artificial nerve network in the stock market forecast model [J]. Microelectron Comput 24(11):147–151

3. Tan S, (2006) The GM(1,1) prediction model of stock market [J]. *Stat Decis* (6):21–23
4. Ye D, Ma Z, Li G et al (2008) Rough set based on fuzzy neural network for the application in prediction of stock market [J]. *Comput Sci* 35(4):168–170
5. Deng K, Zhao Z (2009) Research and simulation of stock price prediction model based on genetic algorithm BP neural network [J]. *Comput Simul* 26(5):316–319
6. Shen C, Zhou K, Wang H (2010) Combined model for forecasting based on positive weight and Its application in the economy [J]. *Math Econ* 27(1):85–91
7. Ju J, Wang J, Wang J (2010) Combined prediction method of total power of agricultural machinery based on BP neural network [J]. *Trans Chin Soc Agric Eng* 41(6):87–92

Part XV
Social and Economical Systems

Chapter 160

Multivariate Curve Resolution with Elastic Net Regularization

Zhu Xin-feng, Wang Jian-dong and Li Bin

Abstract The purpose of Multivariate Curve Resolution (MCR) is to recover the concentration profile and the source spectra without any prior knowledge. We hypothesis that each source is characterized by a linear superposition of Gaussian peaks of fixed spread. Multivariate curve resolution–alternating least squares (MCR-ALS) is a Conventional MCR method. MCR-ALS has some disadvantages. We proposed a solver with L1 regularizer and L2 regularizer to obtain a sparse solution within MCR-ALS. L1-norm involves a sparse but non-smooth solution, L2-norm will keep all the information and bring the smoothness, but it will lead non-sparse solutions. So we combined the L1-norm and the squared L2-norm to seek the optimal solutions. This is accomplished via Elastic Net Regularization algorithm which is LARS (least-angle regression). We named this method MCR-LARS. This paper applies MCR-LARS to resolve the hard overlapped spectroscopic signals belonging to the three aromatic amino acids (phenylalanine, tyrosine and tryptophan) in their mixtures. MCR-LARS was compared with MCR-ALS. The results show the effectiveness and efficiency of MCR-LARS and the results show that MCR-LARS provides more nicely resolved concentration profiles and spectra than pure MCR-ALS solution.

Keywords MCR · MCR-ALS · Elastic Net · L1-norm · L2-norm · MCR-LARS

Z. Xin-feng · L. Bin
College of Information Technology, Yangzhou University,
Yangzhou, China
e-mail: zxfeng168@163.com

L. Bin
e-mail: Yangzhou lb@yzu.edu.cn

W. Jian-dong (✉)
College of Computer Science and Technology, Nanjing
University of Aeronautics and Astronautics, Nanjing, China
e-mail: aics@nuaa.edu.cn

160.1 Introduction

The goal of Multivariate Curve Resolution (MCR) is to help resolve mixtures by determining the number of constituents, their response profiles (spectra, pH profiles, time profiles, elution profiles) and their estimated concentrations, when no prior information is available about the nature and composition of these mixtures [1]. Imposing additional knowledge by using constraints, such as such as positivity, local rank, closure, etc., using MCR can lead to better solutions and easier interpretation of the results [2].

Spectroscopic data can be mathematically modeled basically by using two different approaches. One is analyzing the data by using hard-modeling methods, which require a reaction model to be postulated. The other one is using soft-modeling methods, which do not need to know the kinetic model linked to the reaction studied [3]. The most widely used soft-modeling method may be multivariate curve resolution-alternating least squares (MCR-ALS). Several authors have proposed hard modeling versions of MCR to get the parameters of the process [4].

Spectra are characterized by the presence of a series of peaks for some spectroscopic measurements. However, the number of peaks and their position can not be known in advance in many cases. This point seriously hinders the application of peak shape models within the MCR-ALS.

In this chapter, we assume the source spectra as a dense superposition of Gaussian peaks and then want to obtain a sparse solution in the model obtaining the proper number of peaks and their location automatically without imposing a priori information either the location or the number. To do so, we proposed to use the Elastic Net Regularization technique within the MCR-ALS loop. We name the new algorithm MCR-LARS. Elastic Net Regularization was proposed by Zou [5] and it is known as basis pursuit in the field of variable selection [5]. To test this concept we applied MCR-ALS and MCR-LARS to a three-component mixture's spectra signals which include three kind of aromatic amino acids (phenylalanine, tyrosine and tryptophan). Amino acids play central roles both as building blocks of proteins and as intermediates in metabolism. So the determination of amino acids is useful for biochemical research and for commercial product analysis.

160.2 Multivariate Curve Resolution

160.2.1 MCR Introduction

MCR is based on the assumption that the observed signals are linear superpositions of underlying hidden source signals. MCR techniques may provide 'pure' spectra and concentration profiles under the assumption of linearity. The MCR decomposition of matrix D is carried out according to Eq. 160.1:

$$D = C \cdot S^T + E \quad (160.1)$$

D is the data matrix containing the observed signals of experiment, with the dimension of $M \times N$. C ($M \times K$) is the matrix that contains the concentration profiles (concentration) related to the K pure components. S^T has dimensions $K \times N$ and contains the spectra associated with each pure component. E ($M \times N$) is a matrix of residuals.

We are interested in obtaining the underlying ‘pure’ components present in the sample from the data matrix D . MCR technique provides C without any prior knowledge about S . That means C is obtained without any prior knowledge about the composition of the mixture. MCR is an unsupervised technique that gives us C and S . Eq. 160.1 can be solved by performing either non-iterative methods such as window factor analysis or iterative approaches. Iterative methods are used more frequently because their flexibility, not requiring assumptions in the model.

MCR and PCA [6] techniques model the data matrix as a bilinear decomposition Eq. 160.1. PCA generates a new basis of eigenvectors mutually orthogonal and looks for directions with the maximum variance, while MCR techniques look for vectors maximally decorrelated and maximally pure. A pure variable is defined as that which has only contributions from one of the components [7].

In order to improve the quality of the generalization and the interpretation of the results, several constraints can be applied in MCR: (1) number of pure components expected to be found in the mixture, (2) non-negativity, (3) unimodality, (4) local chemical rank, and (5) closure. In the experiments presented in this paper, only the constraints 1 and 2 were adopted.

160.2.2 Curve Resolution Techniques: MCR Techniques

In MCR methods in the signal processing domain, the source signals are assumed to be statistically independent. This is the case for Independent Component Analysis (ICA) [7. A]. ICA is a popular technique that has been successfully applied in a large number of applications [8 – [13]. However, it cannot be applied to spectrum data since the assumption of statistical independence is not satisfied in general. In many occasions, reactions taking place among the analytes, which violate the assumption of statistical independence. Therefore, alternative MCR techniques should be applied to Spectrum data.

A number of proposals have been put in the chemometrics field. SIMPLISMA was proposed by Windig [39] and MCR–ALS was proposed by Tauler [14]. SIMPLISMA is a single step algorithm, while MCR–ALS is an iterative algorithm initialized using first estimations from SIMPLISMA or Evolving Factor Analysis (EFA [15]). These iterative algorithms calculate C and S at every iteration using alternating least squares (ALS), until convergence condition is reached. Constraints can be applied in the ALS solution during every iteration step [14].

160.3 Elastic Net Regularization

Regularization to MCR–ALS solution has been previously proposed by Wang [18] using Ridge Regression to perform a proportional shrinkage to all the coefficients. But Ridge Regression uses a L2-norm ($\sum a_j^2$) penalty in the coefficients, it is not sparse to tackle high dimensional problem. LASSO has been proposed by Víctor Pomareda to bring sparse solution [25]. The elastic net is similar to the lasso. Elastic net is a way to works as well as the lasso whenever the lasso does the best, and can fix the problems that lassp can not cope with, such as for this kind of $p \gg n$ and grouped variables situation. The elastic net can be expressed by Eq. 160.2.

$$\min_{\beta \in \mathbb{R}^p} \|Y - \beta X\|^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \quad (160.2)$$

In Eq. 160.2, λ is the regularization parameter and α controls the amount of sparsity and smoothness. The L_1 term $\|\beta\|_1$ promotes sparsity and the L_2 term $\|\beta\|_2^2$ promotes smoothness. The elastic net solution is strictly convex and the solution is unique. The main motivation to propose elastic net regularization to MCR technique is to improve the quality of the estimation and the generalization ability of the result. We used Elastic Net Regulizatio methods in the process of MCR–ALS. We named this MCR–LARS.

LARS is an iterative least squares solver using both L_1 -norm and L_2 -norm penalty. The LARS shrinks some fitting a_{ij} coefficients and sets many others to zero. It is accepted that LARS produces more sparse models providing better prediction accuracy and more interpretable models [19].

160.4 MCR–LARS Algorithm

160.4.1 Algorithms Judgement Criterion

MCR–LARS is a technique used within MCR–ALS. A flexible mathematical model for the spectrum shape is introduced in the algorithm and the complexity of this model is controlled by a regularization parameter. In our model, spectrum S_j is modeled as a linear superposition of Gaussians of variable width (regressors):

$$S_j(t) = \sum_{i=1}^N a_{ij} \phi_i(t) \quad (160.3)$$

$$\phi_i(t) = \exp\left[\frac{-(t - t_i)}{2 \cdot \sigma_i^2}\right] \quad (160.4)$$

Equation 160.2 forms the basis of regressors which is used for the linear regression problem; t is the drift time, t_i is the centre of each Gaussian in the x -axis, N is the number of eigenvectors to reconstruct a spectrum, and σ_i is the standard deviation of each one of these Gaussians located on each drift time point. It is obvious that the only fitting parameters are the set of weights $\{a_i\}$ Eq. 160.2. In this way, the model becomes linear in the parameters. We can extend this dense Gaussian model to fit wider non-Gaussian peaks.

The regression parameters $\{a_i\}$ can be estimated by Ordinary Least Squares, but this estimation is ill conditioned when the correlation among the regressors is large [16, 17]. In our spectrum model, each Gaussian is highly correlated with a number of neighboring Gaussians and the data matrix is nearly singular. This fact can lead to singularities and instabilities producing a poor estimation for the a_{ij} coefficients, which will exhibit high variance. The problem can be regularized by imposing a constraint on the coefficients.

We use LARS within Multivariate Curve Resolution in the estimation of pure spectra in each ALS iterative step. An estimation can be derived from a linear approximation in such a way that the vector of weights a_j for the j -th spectrum can be written as the Ridge Regression estimator and the solution can be found by iteratively computing the Ridge Regression [20]:

$$\vec{a}_{j,new} = \left[\phi^T \phi + \lambda_1 \cdot \text{diag} \left(\|\vec{a}_{j,old}\|_1 \right)^{-1} + \lambda_2 \cdot \text{diag} \left(\|\vec{a}_{j,old}\|_2^2 \right)^{-1} \right]^{-1} \phi^T S_j \quad (160.5)$$

In Eq. 160.5 ϕ is the regression matrix of Eq. 160.3 at each time; λ_1 and λ_2 are the regularization parameters which is adjusted by cross-validation [19]. The tuning parameter λ_1 ($\lambda_1 \geq 0$) controls the amount of shrinkage on the coefficients. The larger value of λ_1 , the greater the amount of shrinkage. The λ_1 regularizes the estimation of the a_{ij} coefficients by adding a positive constant to the diagonal of $\phi^T \phi$ before inversion Eq. 160.5. This makes the problem non-singular, even if $\phi^T \phi$ is not of full rank, which is popular in Fluorescence Spectrum data.

MCR-LARS algorithm starts filtering the data matrix with PCA or EFA to get the number of pure components and using the first estimations for spectra from SIMPLISMA. Then C is obtained by means of fast non-negative least squares (FNNLS [21]) from Eq. 160.1 and S is estimated using C and FNNLS (which impose non-negativity in a least squares sense). S is normalized to unit area before being used by LARS, which, given a penalty parameter λ , estimates the best a_{ij} coefficients for each spectrum using Eq. 160.5 iteratively. For the sparsity of LARS, most of the a_{ij} coefficients are zero, generating a sparse solution producing a small and stable subset of Gaussians even with the presence of noise. Using these a_{ij} coefficients, spectra are reconstructed and consequently filtered, removing most of the high frequency noise, and normalized to unit area. The algorithm enters in an iterative procedure recalculating C , S and a_{ij} and goes on until it converges to a given condition. There are two judgement criterions used in this chapter. This first is (RMSE (root mean squared errors) is small enough) or the maximum number of iterations is achieved. The RMSE can be defined as Eq. 160.6:

$$\text{RMSE} = \sqrt{\frac{\sum_{m=1}^M \sum_{n=1}^N (D_{m,n} - \hat{D}_{m,n})^2}{M \cdot N}} \quad (160.6)$$

In Eq. 160.6, D is the original data matrix filtered with PCA and \hat{D} is the reconstructed data matrix using C and S for the current iteration Eq. 160.1. The other judgement criterion is the angular distance between profiles, which provides an estimation of the difficulty of the problem, can be calculated as:

$$\theta_{ij} = \arccos \left[\frac{C_i \cdot C_j}{\|C_i\| \|C_j\|} \right] \quad (160.7)$$

In Eq. 160.7 C_x is the vector containing the evolution in time for the selected spectrum (concentration profile) and $\|C_x\|$ is the Euclidean norm. An angular distance of 0° would indicate that the concentration profiles are identical.

160.4.2 Algorithms Block Diagram

The MCR–LARS algorithm’s block diagram is given in Fig. 160.1, which is almost the same as the MCR–ALS’ block diagram, but for the additional LARS step. When λ_1 and λ_2 tend to zero, the MCR–LARS solution tends to the MCR–ALS solution with slight differences due to the rigidity of the Gaussian model.

160.4.3 Algorithms Implementation Details

MCR–LARS was implemented using MATLAB R2007b and PLS toolbox 5.2 (EigenVector Research, Seattle, USA). The MATLAB software for MCR–ALS is available in [22]. The algorithm used for SIMPLISMA corresponds to the MATLAB function `purity`, which can be found in PLS Toolbox 5.2. SIMPLISMA was applied using the pure variable approach in all the cases [23].

In our experiments, only a fixed number of components and non-negativity constraints were applied in order to interference algorithms’ performance as less as possible. Non-negativity was applied through FNNLS in both techniques (MCR–ALS and MCR–LARS). Spectra were normalized to unit area in each iterative step within the algorithms. The stop criterion (the same for both algorithms) was based on the RMSE between the data matrix filtered with PCA and the reconstructed matrix using C and S Eq. 160.6. At each iteration step, the RMSE is calculated and the algorithms stop if the relative difference between iterations is less than 0.1%.

While some approaches to find the optimal number of pure components (K) have been reported [23, 24], in our case the choice has been based on the EFA (evolving factor analysis) .

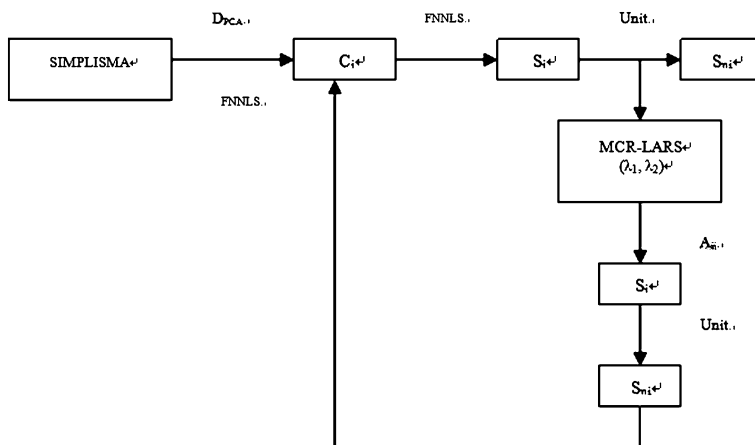


Fig. 1 MCR-LARS algorithm’s block diagram

In the case of LARS, the tuning parameters are determined by cross-validation method between the training and validation datasets. Given a certain value for λ_2 , λ_2 and the a_{ij} coefficients are estimated by ten-fold cross-validation method. The LARS method is applied to the training dataset; the recovered spectra S_{train} are used to estimate the concentration profiles C_{val} using a validation dataset and FNNLS in a single step. The RMSE Eq. 160.6 is evaluated between the validation dataset and the reconstructed dataset using C_{val} and S_{train} . An averaged RMSE is obtained for each λ_1, λ_2 value. High values of λ_2 constricts too much the model; low values of λ_2 allow noise. High values of α constrict too much the sparsity; low values of α constrict to model smoothness. Therefore, the best value for λ_1 and λ_2 is that around a change in the slope in the graph RMSE vs. λ_1, λ_2 . After selecting the tuning parameters, they are applied to the test dataset to compare the MCR solutions.

160.5 Results and Discussion

Phenylalanine, tyrosine and tryptophan are three kinds of essential amino acids containing aromatic ligands. They have been used as drugs or food additives. We use Fluorescence spectra measurements about mixture of Phenylalanine, tyrosine and tryptophan as our datasets to test the algorithms under noisy conditions and these spectra were high overlapped. The training and validation datasets were used for the determination of the parameters (λ_1, λ_2) , and the test set was used to evaluate the performance of algorithm. For comparison, the MCR-LARS will be compared with the MCR-ALS in exactly the same conditions regarding initialization and convergence criteria.

For the synthetic dataset, since the underlying solutions for spectra and concentration profiles are known, the RMSE Eq. 160.6 can be calculated. Eq. 160.6 is used using the matrix of real spectra (profiles) and the matrix of recovered spectra (profiles) instead of the data matrices (D and \hat{D}).

Fig. 160.2 shows a comparison for the results provided by MCR-ALS and MCR-LARS, after selecting the first estimations from SIMPLISMA. Fig. 160.2a, Fig. 160.2b and Fig. 160.2c describe the resolved spectral information of the three amino acids, Fig. 160.2d, Fig. 160.2e and Fig. 2f describe the resolved concentration information of these amino acids.

Due to the challenging conditions, the first estimations from SIMPLISMA were very noisy in the concentration profiles and showing bad resolution in the spectra (not shown). From Fig. 160.2 we can see that MCR-ALS is not able to recover the peaks in the correct position and the peak's shapes are distorted due to noise. Moreover, the estimations for the concentration profiles are very noisy. On the contrary, for a certain range of λ_1 and λ_2 values, although the peak's shapes are a little bit distorted due to the rigidity of the Gaussian model, MCR-LARS is able to recover properly the peak position for the spectra and provides stable, regularized and much less noisy estimations for the profiles. For suitable values of λ_1 and λ_2 , MCR-LARS is performing as a filter at the same time that regularizes the solution.

Regarding the RMSE Eq. 160.6 between recovered solutions and real solutions, MCR-LARS provides much more less errors for spectra and for profiles than those gotten by MCR-ALS. Here we only give the computational results of Phenylalanine for show, MCR-ALS provides the following errors: $2.6 \cdot 10^4$ for spectra and $3.4 \cdot 10^2$ for profiles; MCR-LARS provides the following errors: $3.7 \cdot 10^3$ for spectra and $1.2 \cdot 10^2$ for profiles

Concerning the orthogonal angle Eq. 160.7 for Phenylalanine, MCR-ALS provides: 10° for spectra; 14° for profiles. MCR-LARS provides: 4° for spectra; 3° for profiles. MCR-LARS provides better solutions in terms of orthogonality and RMSE. MCR-LARS results contain less noise, not only in the spectra but also in the concentration evolution, and their interpretation is easier than for MCR-ALS or SIMPLISMA.

Although our work has been based Gaussian peaks, the dense superposition of Gaussian is able to model wider asymmetric peaks usually found in spectroscopy. MCR-LARSO can be applied to other analytical settings provided that spectra models can be based on dense linear superposition of regressors. It is also important to mention that the LARS approach can be used with other peak models such as non-Gaussian peak.

Experiments have shown that in challenging conditions (high noise, very similar concentration profiles, overlapped spectra, and asymmetric peaks) MCR-LARS provides better estimation of the time evolution and spectra of the underlying components than MCR-ALS.

In conclusion, the LARS technique is introduced within the ALS loop for MCR in this chapter. Model complexity is limited by the use of L_1 -norm regularization and smoothness can be kept by the use of L_2 -norm regularization. It can bring a

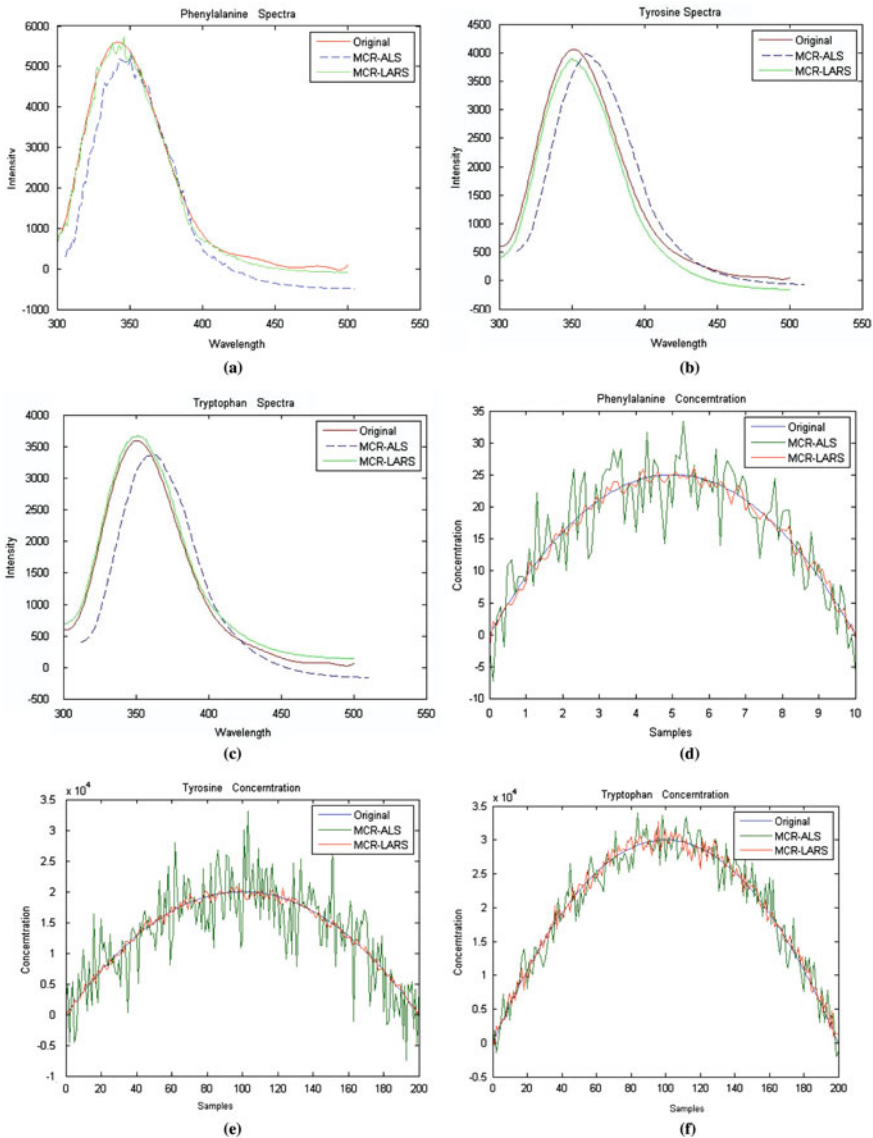


Fig. 2 Results for Amino mixture dataset with MCR-ALS and MCR-LARS (a, b, c) Spectra recovered by MCR-ALS and MCR-LARS (d, e, f) Concentration profiles by MCR-ALS and MCR-LARS

flexible model to tackle ill-conditioned least square problem, and find a sparse solution of limited complexity. So MCR-LARS can be widely used to resolve overlapped signals in many application domains.

Acknowledgements This work is supported by the Jiangsu Key Laboratory of Environmental Material and Engineering (K08021), National Natural Science Foundation of China (No.61070133), the Natural Science and Technology Foundation of Jiangsu Province of China (No.BK210311) and the Natural Science Foundation of Jiangsu Province of China (No. BK2009697).

References

1. Cichoki A, Zdunek R, Phan A, Amari S (2009) Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation. Wiley, Chichester
2. Juan AD, VanderHeyden Y, Tauler R, Massart DL (1997) Assessment of new constraints applied to the alternating least squares method. *Anal Chim Acta* 346:307–318
3. Juan A, Casassas E, Tauler R (2000) Soft-modelling of analytical data. In: *Encyclopedia of analytical chemistry: instrumentation and applications*. Wiley, New York
4. Juan Ad, Maeder M, Martinez M, Tauler R (2000) Combining hard and soft modelling to solve kinetic problems. *Chemometr Intell Lab Sys* 54:123–141
5. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J Royal Stat Soc: Ser B*:301–320
6. Jolliffe IT (1991) *Principal component analysis*. Springer series in statistics New York 2002. Interactive self-modeling mixture analysis. *Anal Chem* 63:1425–1432
7. Windig W, Guilment J (1991) Interactive self-modeling mixture analysis. *Anal Chem* 63:1425–1432
8. Hyvärinen E, Oja (2000) Independent component analysis: algorithms and applications. *Neural Netw* 13:411–430
9. Young G (2005) Independent component analysis, *Encyclopedia of statistics in behavioral science*. Wiley, Chichester, pp 907–912
10. Vigário R, Jousmäki V, Hämäläinen M, Hari R, Oja E (1998) Independent component analysis for identification of artifacts in magnetoencephalographic recordings. *Adv Neural Inf Process Sys* 10:229–235
11. Vigário R (1997) Extraction of ocular artifacts from EEG using independent component analysis. *Electroencephalogr Clin Neurophysiol* 103:395–404
12. Hyvärinen A (1999) Sparse code shrinkage: denoising of nongaussian data by maximum likelihood estimation. *Neural Comput* 11:1739–1768
13. Back D, Weigend AS (1997) A first application of independent component analysis to extracting structure from stock returns. *Int J Neural Syst* 8:473–484
14. Tauler R (1995) Multivariate curve resolution applied to second order data. *Chemometr Intell Lab Sys* 30:133–146
15. Maeder M (1987) Evolving factor-analysis for the resolution of overlapping chromatographic peaks. *Anal Chem* 59:527–530
16. Eldén L (1977) Algorithms for the regularization of ill-conditioned least squares problems. *BIT Numerical Math* 17:134–145
17. Golub G (1965) Numerical methods for solving linear least squares problems. *Numer Math* 7:206–216
18. Wang J-H, Hopke PK, Hancewicz TM, Zhang SL (2003) Application of modified alternating least squares regression to spectroscopic image analysis. *Anal Chim Acta* 476:93–109
19. Hastie T, Tibshirani R, Friedman J (2009) *Linear methods for regression*, *The elements of statistical learning: Data mining, inference, and prediction*, springer series in statistics. pp 43–100
20. Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 96:1348–1360

21. Bro R, Jong S (1997) A fast non-negativity-constrained least squares algorithm. *J Chemometr* 11:393–401
22. Tauler R, Juan Ad (2005) <http://www.ub.edu/mcr/ndownload.htm>, In: multivariate curve resolution homepage
23. Windig W, Gallagher NB, Shaver J, Wise B (2005) A new approach for interactive self-modeling mixture analysis. *Chemometr Intell Lab Sys* 77:85–96
24. Gourvénec S, Massart DL, Rutledge DN (2002) Determination of the number of components during mixture analysis using the Durbin–Watson criterion in the orthogonal projection approach and in the SIMPLE-to-use interactive selfmodelling mixture analysis approach. *Chemometr Intell Lab Sys* 61:
25. Pomareda V, Calvo D, Marco S (2010) Hard modeling multivariate curve resolution using LASSO: application to ion mobility spectra. *Chemometr Intell Lab Sys* 104:318–332

Chapter 161

Research about Exoskeleton's Reference Trajectory Generation Based on RBF Neural Network

Shaowu Dai, Shuangming Li and Zhiyong Yang

Abstract Multidimensional force sensor, introduced into the system, can overcome with inconvenience of wearing the exoskeleton. The estimated trajectory, by the nonlinear mapping between the body's path and human-machine force through RBF neural net, was taken as the input of the impedance controller. This can resolve difficulty in body's path measurement, which is caused by system's closed coupling and misalignment. At the end, Simulation test was did without the control and with the control. The Simulation result indicated that the RBF can accurately estimate human body's path and make exoskeleton track human body's path.

Keywords Exoskeleton · Impedance control · Multidimensional force sensor · RBF net

Fund: Natural Science Foundation of Shandong Province (ZR2010FM009).

161.1 Introduction

The exoskeleton system is one kind of close coupling man-machine intelligent system, and also called the flexible exoskeleton system [1], which can be divided into the remote exoskeleton system [2, 3], the power assist exoskeleton system [4], the movement exoskeleton system [5], the medicine exoskeleton system [6] and so on. What this chapter studies is the power assist lower limb exoskeleton [7].

S. Dai · S. Li (✉) · Z. Yang
Naval Aeronautical and Astronautical University, 264001 Yantai, China
e-mail: aminglishuang@126.com

Although the research about it is quite early, there still are some imperfect aspects in the man–machine integration control [8], human body information and information fusion. The BP neural network sensitivity control method has been used in the literature [9], making up the SAC controller’s flaw to establish the precise mathematical model. But, in the approximation of the function, this method has many problems, such as convergence rate shortcoming and local minimum. The literature [10] has used passive control policy to help patient rehabilitation training. Machine’s gait was defined by procedure in advance, not reflecting person’s movement situation. In the literature [11], angle sensors, EMG sensors and ground sensors were used to get state information, which largely reduced the comfort of the exoskeleton system.

To solve the above problems, this chapter proposes a method-exoskeleton’s reference trajectory generation based on RBF neural network. Approaching the nonlinear mapping relation between the man–machine force [12] and body’s path, it can estimate the impedance controller’s reference trajectory and reduce operator’s energy consumption. When different people wear the exoskeleton, the interpersonal model parameter is also different. Thus, different man–machine model were selected to verify its influence.

161.2 Impedance Control of Exoskeleton System

Exoskeleton system can be viewed as more links’ mechanical structure. Referring to the D–H (Denavit Hartenberg model) model used commonly in the multi-body kinematics and [13], the system’s dynamic model can be built by Lagrange equation:

$$H(q)\ddot{q} + C(q, \dot{q})\dot{q} + G(q) = T_a + T_{hm} \quad (161.1)$$

Where, q, \dot{q}, \ddot{q} are separately joint angle, joint angular velocity and joint angular acceleration, $H(q)$ is inertia matrix, $C(q, \dot{q})$ is Coriolis force, $G(q)$ is Gravity item, T_a is the joint torque imposed by driver in the joint space, T_{hm} is human–machine interactive force.

Relation between the T_{hm} and generalized force f is:

$$T_{hm} = J^T(q)f \quad (161.2)$$

Let x, \dot{x}, \ddot{x} denote the position, velocity, acceleration of the trunk, then

$$\dot{x} = J(q)\dot{q} \quad (161.3)$$

$$\ddot{x} = J(q)\ddot{q} + \dot{J}(q)\dot{q} \quad (161.4)$$

By Eqs. 161.2–161.4, the system dynamic model can be built as:

$$A(q)\ddot{x} + B(q, \dot{q})\dot{x} + D(q) = u + f \quad (161.5)$$

Where

$$A(q) = J^{-T}(q)H(q)J^{-1}(q), B(q) = J^{-T}(q)C(q, \dot{q})J^{-1}(q) - A(q)J(q)J^{-1}(q),$$

$$D(q) = J^{-T}(q)G(q), A(q) = J^{-T}(q)H(q)J^{-1}(q)$$

With the inverse dynamic method, controller is designed as;

$$T_a = \hat{H}(q)\alpha + \hat{C}(q, \dot{q})\dot{q} + \hat{G}(q) - J^T(q)f \quad (161.6)$$

$$\alpha = J^{-1}(q)(\tau - \hat{J}(q, \dot{q})\dot{q}) \quad (161.7)$$

$$\tau = \dot{x}_c + K_{Mp}^{-1}(K_{Dp}\Delta\dot{x} + K_{Pp}\Delta x + f) \quad (161.8)$$

Where, $\hat{H}(q)$, $\hat{C}(q, \dot{q})$, $\hat{G}(q)$ are separately the estimated values of $H(q)$, $C(q, \dot{q})$ and $G(q)$. $x_c, \dot{x}_c, \ddot{x}_c$ denote separately the position, velocity, acceleration of the estimated trajectory, and $\Delta x = x_c - x$, $\Delta\dot{x} = \dot{x}_c - \dot{x}$, $\Delta\ddot{x} = \ddot{x}_c - \ddot{x}$. K_{Mp} , K_{Dp} , K_{Pp} are separately ideal definite matrix, damping matrix and stiffness gain matrix.

It follows from the Eqs. 161.5–161.7 that:

$$\Delta\ddot{x} + K_{Mp}^{-1}(K_{Dp}\Delta\dot{x} + K_{Pp}\Delta x + f) = \hat{A}(q)^{-1}(\Delta A(q)\ddot{x} + \Delta B(q, \dot{q})\dot{x} + \Delta D(q)) \quad (161.9)$$

Where

$\Delta A(q) = A(q) - \hat{A}(q)$, $\Delta B(q) = B(q) - \hat{B}(q)$, $\Delta D(q) = D(q) - \hat{D}(q)$. $\hat{A}(q)$, $\hat{B}(q)$, $\hat{D}(q)$ are separately the estimated values of $A(q)$, $B(q)$, $D(q)$, and

$$\hat{A}(q) = J^{-T}(q)\hat{H}(q)J^{-1}(q) \quad (161.10)$$

$$\hat{B}(q) = J^{-T}(q)\hat{C}(q, \dot{q})J^{-1}(q) - \hat{A}(q)J(q)J^{-1}(q) \quad (161.11)$$

$$\hat{D}(q) = J^{-T}(q)\hat{G}(q) \quad (161.12)$$

Under ideal condition, $\Delta A(q) = \Delta B(q) = \Delta D(q) = 0$, impedance relation between the man–machine force and exoskeleton is:

$$-f = K_{Mp}\Delta\ddot{x} + K_{Dp}\Delta\dot{x} + K_{Pp}\Delta x \quad (161.13)$$

Selecting proper K_{Mp} , K_{Dp} and K_{Pp} can enable operator easily to swing exoskeleton for the aim of impedance control.

However, due to the simple man–machine interaction model and the high order item (nonlinear item) existing in the actual model, so a wrong result may be got with this solution, which will increase human body's energy consumption instead. Thus, research about estimation of the exoskeleton trajectory based on RBF is proposed, as presented in Fig. 161.1.

Fig. 161.1 Block diagram of exoskeleton's impedance control based on RBF neural network

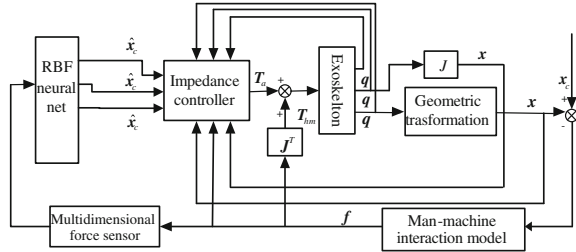
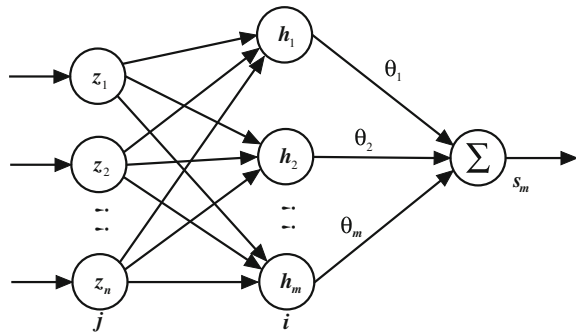


Fig. 161.2 Structure of RBF neural network



161.3 RBF Neural Network

Radial Basis Function was proposed by J. Moody and C. Darken in the later 1980 s, which is a kind of three forward network. Identification structure is presented in Fig. 161.2.

$H = [h_1, h_2, \dots, h_m]^T$ is Radial Basis Vector, where h_i is radbas [14].

$$h_i = f(\|\mathbf{z} - \varphi_i\|/\sigma_i^2), i = 1, 2, \dots, m \tag{161.14}$$

Where $\mathbf{z} = [z_1, z_2, \dots, z_n]^T$ is the input of network, $\mathbf{h}(\mathbf{z}) = [h_1, h_2, \dots, h_m]^T$ is the output of radbas, m denotes the number of neurons, and $\varphi_i = [\varphi_{i1}, \varphi_{i2}, \dots, \varphi_{in}]^T$ denotes the center vector of the i th mode.

Suppose that base wide vector is:

$$\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \dots, \sigma_m]^T \tag{161.15}$$

where $\sigma_i(\sigma_i > 0)$ is the base wide parameter of mode i .

Net weight vector is:

$$\boldsymbol{\theta} = [\theta_1, \theta_2, \dots, \theta_m]^T \tag{161.16}$$

And output of network is:

$$s_m(t) = \boldsymbol{\theta}^T \mathbf{h}(\mathbf{z}) \quad (161.17)$$

Then, performance index function is:

$$J = \frac{1}{2} (s_m(t) - s(t) - \frac{f}{\hat{K}_{pf}})^2 \quad (161.18)$$

Where \hat{K}_{pf} is the stiffness coefficient estimation of man-machine interaction model, f is man-machine force. With the gradient decent method, iterative algorithm of output weight is presented as follow:

$$\theta_i(t) = \theta_i(t-1) + \eta (s_m(t) - s(t) - \frac{f}{\hat{K}_{pf}}) h_i + \alpha (\theta_i(t-1) - \theta_i(t-2)) \quad (161.19)$$

161.4 Model Simulation

For simplicity, the single degree of freedom exoskeleton model is selected as the control object, which can be regarded as a pendulum. It simultaneously receives the driving of hip joint driver and man-machine force. In the simulation, its mass is 10 kg, and length is 0.5 meter. Linear combination of the stiffness, spring and damping model is selected as the man-machine interactive model [15], as follows:

$$\mathbf{f} = \mathbf{K}_{Pf}(\mathbf{x}_h - \mathbf{x}_e) + \mathbf{K}_{Df}(\dot{\mathbf{x}}_h - \dot{\mathbf{x}}_e) + \mathbf{K}_{Mf}(\ddot{\mathbf{x}}_h - \ddot{\mathbf{x}}_e) \quad (161.20)$$

161.4.1 Simulation with Control

Different man-machine model parameters have different influence on the simulation. that means different people wear the exoskeleton. To verify the influence caused man-machine model parameters, three simulation tests are done with different parameters,

- a.: $K_{Pf} = 500, K_{Df} = 0, K_{Mf} = 0$, denotes the spring model.
- b.: $K_{Pf} = 500, K_{Df} = 10, K_{Mf} = 0$, denotes the spring and damping model.

Results are shown in the Figs. 161.3–161.6.

The simulation results of test a are shown in Figs. 161.3–161.4. The curve indicates that exoskeleton can not effectively track the motion of the body, and operator need impose much torque.

The simulation results of test b are shown in Figs. 161.5–161.6. The curve indicates that exoskeleton can not exactly track the motion of the body, and operator need impose little torque compared with Fig. 161.4.

Fig. 161.3 The curve of angle trace (a)

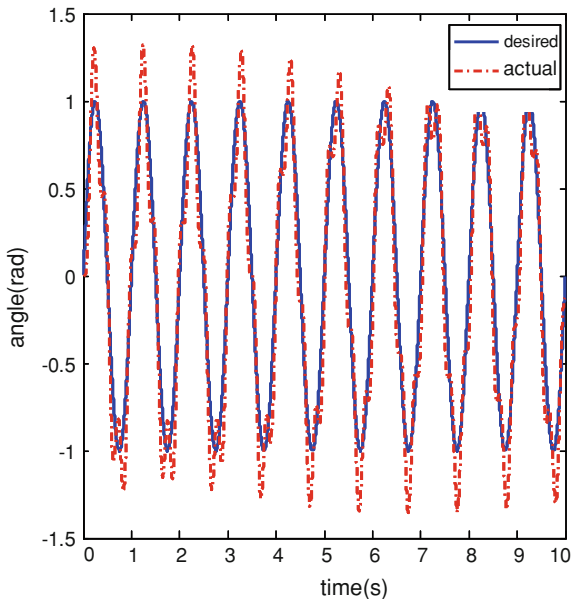
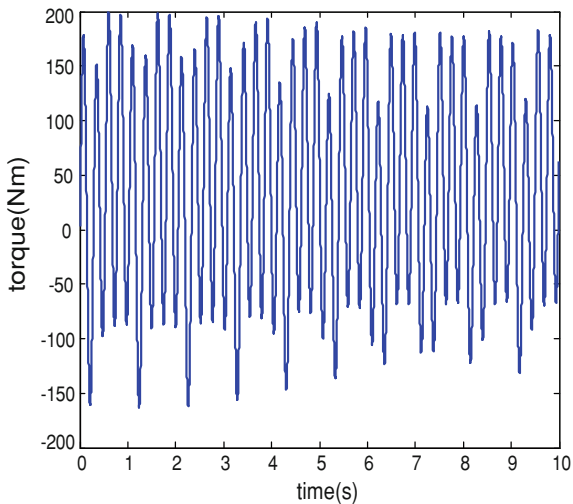


Fig. 161.4 Force exerted by actuator (a)



Without control, people need apply all the force, driving the exoskeleton to walk. Besides, the stiffness, spring and damping model has the better control (Figures 161.7–161.8).

Fig. 161.5 The curve of angle trace (b)

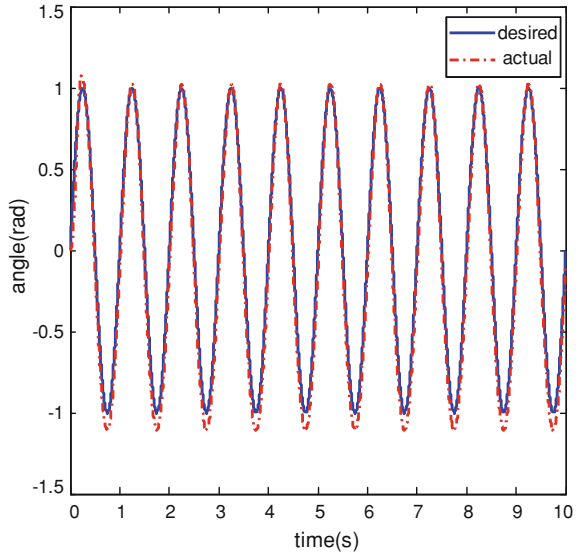
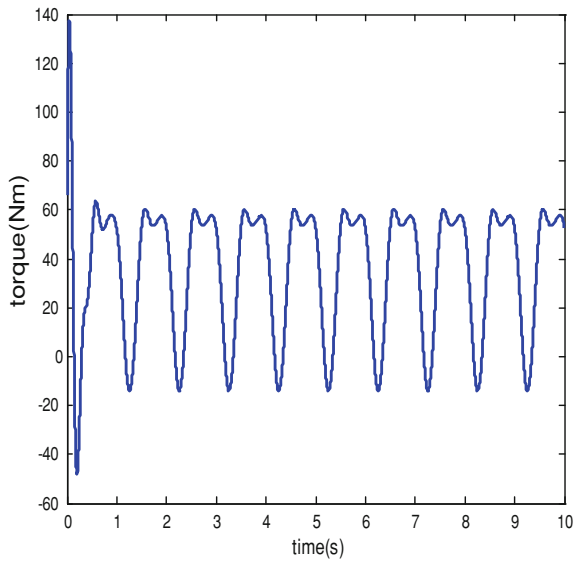


Fig. 161.6 Force exerted by actuator (b)



161.4.2 Simulation with Control Based on RBF

We have known the effect of man-machine parameters on the simulation, so it's not necessary to choose different parameters. In this section, one set of data is chosen: $K_{Pf} = 500, K_{Df} = 2, K_{Mf} = 0.006$.

Fig. 161.7 Angle trace with control

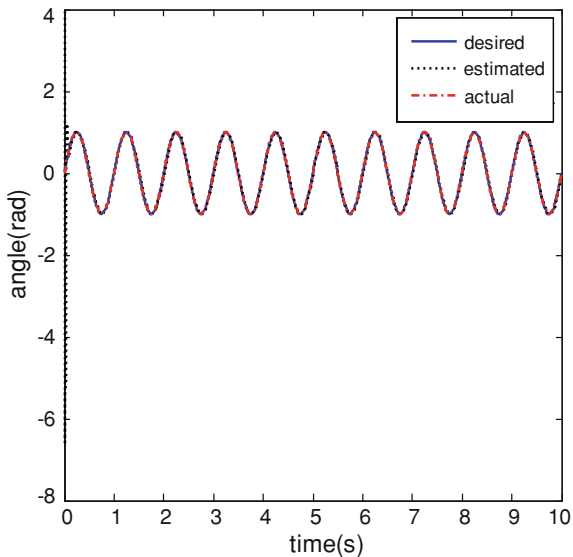
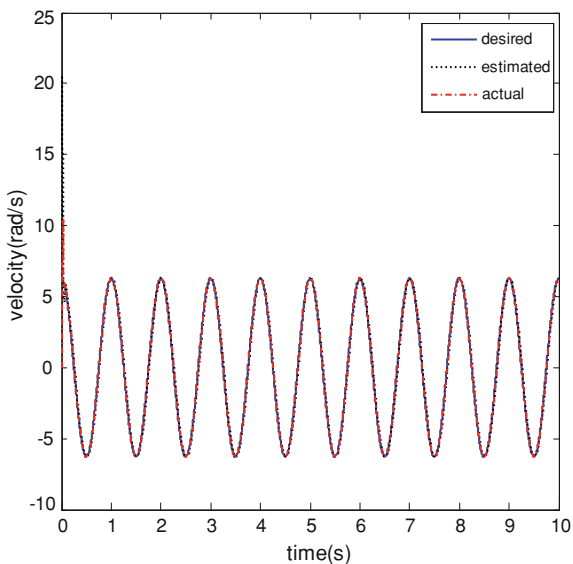


Fig. 161.8 Angle velocity trace with control



Number of neurons is ten, base function is radbas, mode center parameter is 0.8, mode base wide parameter is 20, rate of net learning is 0.9. The expected impedance coefficient is $K_{Pp} = 1, K_{Dp} = 3, K_{Mp} = 0.01$, and the estimated parameters of man-machine model are $K_{Pp} = 1, K_{Dp} = 3, K_{Mp} = 0.01$

Fig. 161.9 Angle acceleration trace with control

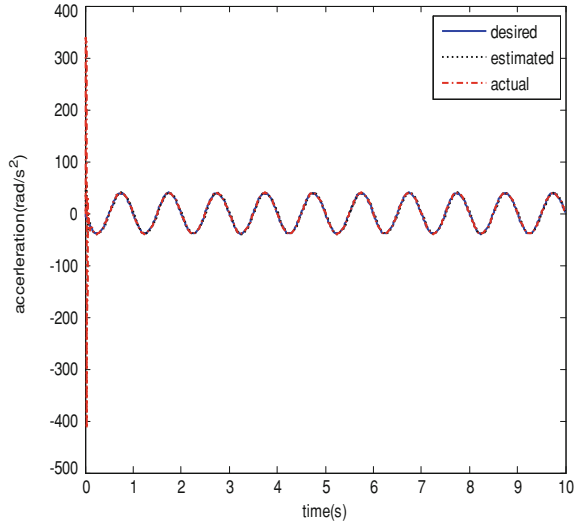
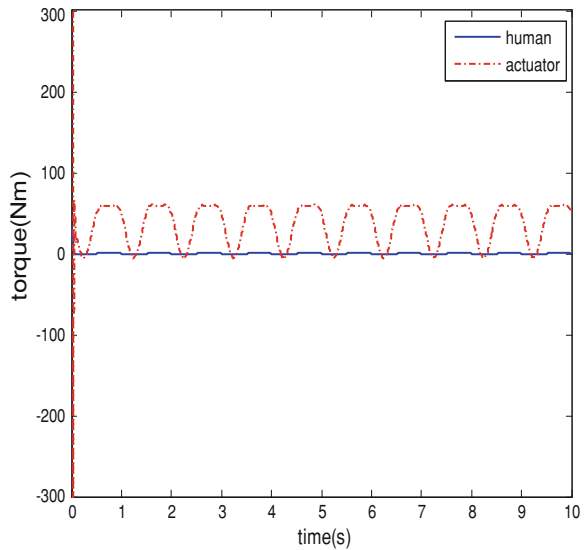


Fig. 161.10 Torque comparison between driver and man



The simulation results are shown in the Figs. 161.9–161.10.

The desired curve, estimated curve, actual curve of the angle, angel velocity and angle acceleration are shown in the Figs. 161.9–161.10. Through the comparison among the desired value, estimated value and actual value, we can find that RBF neural network can estimate body's movement information, and impedance controller can exactly drive exoskeleton to track body's path. Figure 12 indicates

that, in the joint space, exoskeleton driver exerts the majority of the torque, while operator only need little force to exert. It means that man-machine interactive force is very small. Compared with no control, this way obviously reduced body energy consumption to realize power assist.

161.5 Conclusion

Human body's trajectory is required as the reference input for the impedance controller of system, which is difficult to directly measure in the system, compared with the human-machine interactive force. Generally speaking, if the man-machine interaction model has been known in advance, human body's path may be estimated. However, with the man-machine interaction model parameter's uncertainty, there is big error in the estimated reference path. In order to solve this problem, a method about BRN net's application in the exoskeleton's reference trajectory was proposed. It can estimate exoskeleton's reference trajectory on real-time line and has the very good robustness to the man-machine model parameter variation. The simulation results verify the accuracy and validity of this method.

References

1. Sheridan TB (2002) Human and automation: system design and research issues. John Wiley, New York
2. Yang CJ, Zhang JF, Chen Y et al (2008) A review of exoskeleton type systems and their key technologies. In: Proceeding of the Institution of Mechanical Engineers. Part C; J Mech Eng Sci 222(8):1599–1612
3. Sheridan TB (1989) Telerobotics. *Automatica* 25(4):487–507
4. Huang GT (2004) Wearable robots. *Technology Review* 70–73
5. Riener R, Colombo G, Lunenburger L (2006) Overview of robot-aided gait biofeedback and assessment. In: The first IEEE/RAS-EMBS international conference on biomedical robotics and biomechatronics, Feb. 2002, pp 965–970
6. Nessler JA, Reinkensmeyer DJ, Timoszyk WK et al (2003) The use of a robotic body weight support mechanism to improve outcome assessment in the spinal cord injured rodent. In: Proceeding of the 25th annual international conference of the IEEE EMBS. Cancun, Mexico, September 17–21, pp 1629–1632
7. Riener R, Lunenburger L, Jezernik S et al (2005) Patient-cooperative strategies for robot-aided treadmill training: first experimental results. *IEEE Trans Neural Syst Rehabil Eng* 13(3):380–394
8. Yang Z, Gui L, Yang X, Gu W, Zhang Y (2007) Simulation research of exoskeleton suit based on sensitivity amplification control [C]. In: Proceeding of the IEEE international conference on automatic and logistics, Jinan, China, pp 1353–1357
9. Yang Z, Gui L, Yang X, Gu W (2008) Simulation research of exoskeleton suit based on neural network sensitivity amplification control. In: Chinese control and decision conference, Yantai, China, July 2–4, pp 3256–3260
10. Young C, Rusk HA (2007) Rehabilitation Center. <http://www.ruskrehab.com/Web%20DocumentsSpinal50.pdf>. Accessed 23 Nov 2007

11. Kawamoto H, Sankai Y (2002) Power assist system HAL-3 for gait disorder person. In: Proceedings of 8th international conference helping people with special needs, pp 196–203
12. Siciliano B, Luigi V (1999) Robot force control [M]. Kluwer, Boston/Dordrecht/London
13. Racine JLC (2003) Control of a lower extremity exoskeleton for human performance amplification [D], University of California
14. Liu J (2008) Design of the robot control system and Matlab simulation [M]. Publishing House of QingHua University, Beijing
15. Wang H, Low KH, Wang MY (2006) Reference trajectory generation for force tracking impedance control by using neural network-based environment estimation. In: 2006 IEEE conference on robotics automation and mechatronics. Bangkok, Thailand, pp 1–6

Chapter 162

The Application of Digital Archives Classification with Progressive M-SVM to Wisdom School Building

Xiaobo Tao

Abstract When using the SVM algorithm, the training set is so large that the traditional classification methods can't satisfy the real-time requirements, how to design a more efficient SVM algorithm is one of the important study problems. We improve the method of the building about the digital archive's corpus and also improve the course of Chinese participle and the multiprocessing of text feature selection with TF, IDF and Information Gain. The experiment shows that this improved method about M-SVM has obtained a better result.

Keywords Wisdom school · Digital archive's corpus · Feature selection · Progressive M-SVM

162.1 Introduction

Text classification method which used to judge the category of pre-defined text according to their content automatically by computer with classification algorithm. In foreign countries, the research on english text has been relatively maturely. In China, the text classification method is a still more foreland research. K-nearest Neighbor [1] is a widely used classification method which is more simple and efficiently; Decision Tree [2] is an induction learning method which has stronger ability of noise exclusion and anti-sense expression, but when the category is too much, the error may be increased easier. M-SVM is a more popular method about text information classification, which can be used for a large scale in text data

X. Tao (✉)

Tourism College of Zhejiang, Hangzhou 311231, China

e-mail: powertxb@sina.com

mining. In this chapter, in order to have more exactly classification result about our digital archive for the project of the building about wisdom campus, we do some improvement about the classification method with M-SVM.

162.2 Several Conventional Text Categorization Methods

There are several conventional methods to solve the text categorization problems, our chapter will illustrate three algorithms. Different algorithms become different text classification result.

K-NearestNeighbors (KNN) [1, 3]. K Nearest Neighbor (K-NN) classification method is a traditional text classification method, the algorithm is simple, the result is stable and effective, the basic idea is that when the document which is classified called x , you must find K articles which is the most similar to x in the training set, according to the type of this article documents, the K is used to determine the document x 's type. The algorithm as Eq. 162.1 as follows:

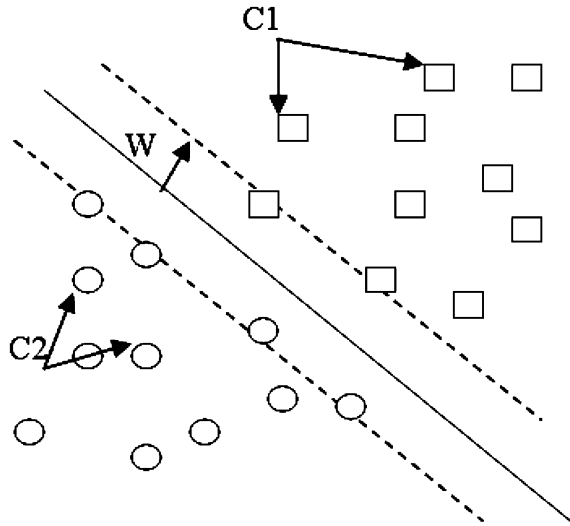
$$H_{knn}(x') = \text{sign} \left(\frac{\sum_{i \in knn(x')} y_i \cos(x', x_i)}{\sum_{i \in knn(x')} \cos(x', x_i)} \right) \quad (162.1)$$

$H_{knn}(x')$ shows that the indexes of the K documents which have the highest cosine with the document to classify x' .

Decision Tree [2]. The Decision Tree provides the method under the similar rule of what conditions obtained with the value. The upmost node of the decision tree is called the root node which is the beginning of the decision tree. The numbers of the child node with each father node determined by the relevant algorithm used by the decision tree. Each branch either a new decision node or the end of the tree which to be called leaves. When traversing along the Decision tree from top to bottom, each node will encounter a problem, on the issue of each node, different answers lead to different branches, and finally will reach one leaf node.

SVM classifier. In the early stage Vapnik put forward the support vector machines, mainly solved the problem of the II Type about linear classification. The thesis is also based on this support vector machine using secondary classification with the problem about digital archives classification. M-SVM is called Multi-SVM which is another classifier can give the category of more than one classification of the text, it have the same theory supposing that the training data $(x_1, y_1), \dots, (x_i, y_i), x \in R^n, y \in \{+1, -1\}$ can be separated by a hyperplane $(w \bullet x_i) + b$. If the vector set is separated by the hyperplane rightly and the nearest vector from the hyperplane and the distance between the hyperplane is the biggest, we said that this vector set to be separated by the optimal hyperplane (or the maximum interval hyperplane). The vector located in the $(w \bullet x_i) + b = +1$ and $(w \bullet x_i) + b = -1$ called support vector. Combine the above two equations: $g(x) = (w \bullet x_i) + b$, where $w \rightarrow$ is a vector perpendicular to the hyperplane which

Fig. 162.1 A separable support vector machine about Type II decision-making rule



defines the orientation of the hyperplane, and b defines the position of the hyperplane. when for the Type II decision-making rule is that: if $g(x) \geq 0$, then determine x is c_1 ; if $g(x) \leq 0$, then determine x belongs to c_2 , shown in Fig. 162.1 at follows:

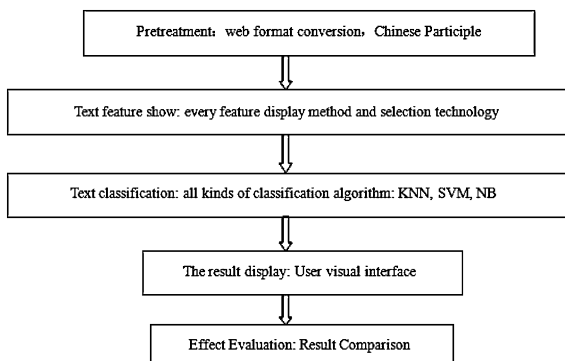
162.3 The Application of Digital Archives Classification with Progressive M-SVM

In this chapter, before using the M-SVM classification algorithm, according to particularity of the digital archives about Tourism College of Zhejiang, we should do some related process about all the digital text, so we should do some improvement about the building of corpus, the Chinese participle and feature calculate and selection, and the progressive M-SVM algorithm mainly improved the operational efficiency.

The General Process. Text classification process consists of the following steps: Corpus building, Chinese participle, feature computer and selection, text feature vector representation, text classification and classification performance evaluation. The detailed process in the below: Fig. 162.2

Corpus building and improvement [5]. Before the text classification work we should do some collection and maintenance about the digital archives corpus. In order to improve the precision and recall of the classification results. During the process of corpus production, we eliminate those redundant which may affect the classification effect. So we set up a special corpus for the wisdom campus building which is needed by digital archives, this is also the innovation of this article.

Fig. 162.2 The process of text classification



The corpus used in this chapter is derived from the Tourism College of Zhejiang digital archives, which has been taken nearly 100 samples from each category document.

Chinese Participle and improvement [6]. In this chapter, we used the publication “Mass Smart Segmentation Research Edition” which is researched by researchers come from Mass Intelligent Computing Research Center, in this chapter, we improve the participle system to the specific information of our school’s digital archives as follows:

Step 1 Append professional vocabulary to various categories. For example, we choose “Personnel”, “post”, “status”, “training” and other professional and decisive words to human resources classification, which have semantic features to our corpus. Therefore, it is one improvement of the chapter, it will through an interface function call, to loaded all the user-defined words into custom segmentation component so that to improve the participle results effectively.

Step 2 To correct the units and post name after Chinese participle. It was found that, although the participle system supports identification for short name, but for school activities, academic archives and other digital files in teaching subject, all the description is over three words length, so the participle system can’t incise exactly name.

Therefore, in the process of participle, we have done some adjustment slightly about our participle system, we examine the two consecutive name word further which be cut from the participle system. If the anterior noun representative name and the back noun also representative name have no consecutive on the sense and meanings, then we combined the two words in the middle as one word unit.

Feature calculation and selection [7]. When doing text mining, feature vector dimension is usually very large [7], this tends to reduce the efficiency and quality of text mining, hence there is the need for feature selection to reduce the vector dimension. Feature selection is from the original feature vector space to select the relevant characteristics which can reflect the characteristics of the relevant characteristics. After the words’ segmentation, there are still exist a large number of some trashy adjectives, nouns and verbs. In this chapter, we use the IG method can

remove a large number of meaningless words in semantic [8] analysis. Because of the feature column items in the table WORD including more larger IG value of all the text, so one text using a little features represent more classification information. Computer the word IG value method basic idea is that by setting a threshold value, and then take the greater threshold value as the characteristics of the effective words. Computer each words' information gain (Information gain, IG) according to stat the number of words appears or not appear in a document to predict the types of documents [3, 6], IG value represents the amount of information in its class, we made the scale between 0 and 1, the larger value IG is, the greater the words' contribution provided. And the IG calculation example is shown below.

$$\begin{aligned}
 IG(T) = H(C) - H(C/T) = & - \sum_{n=2}^n p(c_i) \log_2 p(c_i) + p(t) \sum_{n=1}^n p(c_i/t) \log_2 p(c_i/t) \\
 & + p(t) \sum_{n=1}^{-n} p(c_i/t) \log_2 p(c_i/t)
 \end{aligned}
 \tag{162.2}$$

$p(c_i)$ shows that the c_i appears' probability, using 1 divided by the total number of categories, $p(t)$ is the probability of the feature T appears, as long as there have been T , with the number of documents divided by the total number of documents, such as when $p(c_i/T)$ shows T appears that when the probability of class c_i appears, and with the emergence of T documents which belong to the class c_i divided by the number of documents t appears.

162.4 Specific Steps for Text Classification

The specific steps as follows:

Step 1. Extract the useful information of each document about Chinese digital archive, this function was accomplished by the information extraction module, the document which was extracted saved in the selected folder at the same level of the TXT folder, the file name was the same as the original document name.

Step 2. Do participle operation to each txt document with the mass intelligent participle system. Firstly, obtain the directory of the selected folder and search all the documents below the folder, do cycling operation to each files. Secondly, read the entire contents of the document in stream from and stored in string. Define the word list which used to return all of the participle word, initialized the sub-word dictionary, create the sub-word handle, pass the string to the interface function provided by intelligent participle system, if the sub-term success, search every word and add them to word list, and then computer times each word appeared in the current document and recorded as TCount. Finally, read all the word with their property and insert them into table WORD in database, including FILENAME,

Table 162.1 The dimension is 3000 before the improved of corpus and feature disposal

Types	Accuracy(%)	Recall(%)
Human Resources(1)	85.15	86.02
School Activities(2)	86.01	87.01
Academic Archives(3)	83.80	84.10
Research Files(4)	86.75	87.05
Financial Records(5)	85.08	86.03
Assets' Management Archives(6)	84.08	87.34

TYPE, WORDNAME, TCOUNT, FLAG, which the category is the one of the six categories correspond with figures. The tag is used to indicate the current word whether belong to the corpus, if they are marked as 1, the opposition marked as 2, that's because the follow participle word also stored in this table, so we used FLAG to have distinguish.

Step 3. Calculate the TF-IDF value for each word, which mainly used to reduce the dimension when making the template on the following.

Step 4. We calculate the IG value for each word after dimension reduction, The IG understands the information value of the word in it's type, we have scaled the value between 0 and 1.

Step 5. Generate the corpus samples, which need to use value be computered form steps 3 and 4, after obtained the samples using LIBSVM tool to generate SVM model, and finally writes the results to the classification table which shown through the visual interface display, and the category completed.

162.5 Experiment Design and Result

Generally, we use the accuracy and recall to evaluate the the reflection of text classification work, the formula as following:

Accuracy: the ration of the exact classification of total samples in all categories.

Recall: the percentage of the correct classification in the text which should be classified rightly.

The corpus used for testing is totally six types, each class have 100 files, that is: human resource files (1), school activities (2), academic archives (3), research files (4), financial records (5) and assets' files management archives (6), using 100 pages files to do the forecast work, each type including 20 pages. Put the 100 pages resumes as the training data and put into the SVM-TRAIN and use the LIBSVM to train modules and gain the result of SVM model txt, which used to be as the preparation when the last classification work. Withing the obtained training data by the LIBSVM with the SVM-PREDICT module, we can predict the classification result. In the experiment, we use the before and after situation about the corpus and feature disposal improved with M-SVM to observe the improvement

Table 162.2 The dimension is 2000 after the improved of corpus and feature disposal

Types	Accuracy(%)	Recall(%)
Human resources(1)	91.56	92.08
School activities(2)	92.06	93.85
Academic archives(3)	90.17	91.99
Research files(4)	91.01	91.23
Financial records(5)	91.06	92.77
Assets' management archives(6)	92.01	93.06

about our chapter's method. The accuracy of classification and recall test results are as follows (Tables 162.1, 162.2.)

Thus, in the case of the same experiment data, through the classification result described in corpus and feature selection improvement with M-SVM, we can find that 2000 dimension with the improvement method can have more effective classification result, and really help to save a lot of time in computing

162.6 Summary

This work is supported by Tourism College of Zhejiang under the project of the building about The Wisdom College.

References

1. Yuan P, Chen Y, Jin H (2008) MSVM-KNN: Combining SVM and K-NN for multi-class text classification [C]. In: Huang shan: IEEE international workshop on semantic computing and systems, pp 133–140
2. Appavu AB, Rajaram R (2007) Suspicious E-mail detection via decision tree: A data mining approach [J]. CIT. J Comput Inf Technol 15(2):161–166
3. Sun R-Z An improved KNN algorithm for text classification. Comput Knowl Technol. doi: CNKI:SUN:DNZS.0.2010-01-073
4. He B, Chuan KC, Han J (2004) Discovering complex matchings across web query interfaces: a correlation mining approach [C] KDD Washington
5. Beguel C (2005) Text mining and natural language processing technologies to support competitive intelligence efforts. Temis, London
6. Feng Y, Li H Zhong J, Ye C-X (2010) Text classification algorithm based on adaptive Chinese word segmentation and proximal SVM 01–064. doi: CNKI:SUN:JSJA.0
7. Jiang H, Chen L-Y (2010) A new feature selection method in SVM text categorization, Comput Technol Dev. 03-006. doi:CNKI:SUN:WJFZ.0
8. Li YJ et al (2008) Text document clustering based on frequent word meaning sequences. Data Knowl Eng 64:381–404

Chapter 163

Economic Growth Differential Model and Short-Term Economic Growth Momentum

Fan Decheng and Liu Kan

Abstract Based on the income determination theory, the Economic Growth Differential Model is established to help a country make reasonable economic strategy to achieve the short-term economic goals. Furthermore, there are empirical analysis of economic growth momentum in China in 2009, to identify the main problems and the main strategy of economic growth in present stage.

Keywords Economic growth differential model · Income determination theory · Propensity to consume · Propensity to investment · Spontaneous consumption · Spontaneous investment

Supported by: National Natural Science Foundation project (70773027), National Soft Science Research Program (2010GXQ5D327).

163.1 Introduction

Economic growth usually means the growth of GDP or GDP per capita in a country or region. After the financial crisis in 2007, the development of economic and economic environment is more complicated. Lots of countries have taken appropriate stimulus to pull the growth of its economy in short period of time. The main driving force of economic growth are different in states. China achieved great economic results, by stimulating investment, opening credit and making other

F. Decheng (✉) · L. Kan

Liu Kan.School of Economics and Business Administration, Harbin Engineering University, Harbin 150001, China

economic policies. Digging the main driving force of economic growth, and making in-depth study of the current model of economic growth, can help to make reasonable policies to keep sustainable growth of China's economy after the financial crisis. There are several perspectives about economic growth model, which put forward the corresponding strategies and recommendations. Li Fuqiang, Dong Zhiqing and Wang Linhui analyzed the important role of human capital and physical capital on China's economic growth through OLS and GMM estimates, based on characteristics of our market and property rights reform [1]. Pei Pingcao and Yuan Fang analyzed the role of export and investment on economic growth, using the cointegration of 1999–2007 data [2]; Liu Xihe made analysis of the US growth model about the global allocation of resources and the expansion of consumer lending [3]. These made in-depth analysis of the main mode of economic growth in China and the United States, and provides a theoretical basis for China's economic growth strategy. But there is not a reasonable model of economic growth for the country to achieve short-term economic growth. On the other hand, these researches are based on the period before the financial crisis, while the financial crisis changed the international economic situation and affected the economic growth mode in China. So it is necessary to make in-depth analysis of the main driver and patterns of economic growth in the short term in China under financial crisis, to make properiate strategy to support short term economic growth.

According to the role of input–output and demand, a short-term economic growth differential model is established to help a country make economic strategy to achieve the short-term economic goals reasonably. On this basis, there are empirical analysis of economic growth momentum in China in 2009, to identify the main problems and solutions of economic growth in present stage.

163.2 Economic Growth Differential Model

According to the national income identity:

$$\text{GDP}(t) = y^C(t) + y^I(t) + y^J(t) + y^S(t) \quad (163.1)$$

$y^C(t)$ represents the consumer spending in year t ; $y^I(t)$ represents the investment expenditure in year t ; $y^J(t)$ represents the net exports in year t ; $y^S(t)$ represents the amount of inventory in year t , reflects the current difference between output and final demand.

Based on the income determined theory, the total consumption $y^C(t)$ in year t depends on the previous year's national income $NI(t-1)$. Introduce b as the marginal propensity to consume, which means the rate of consumption increase in unit revenues increase, and introduce a as spontaneous consumption. Then, $y^C(t) = a + bNI(t-1)$.

Similarly, the total investment $y^I(t)$, is expressed as a function of national income $NI(t-1)$, with h as the tendency of whole social investment, q representing the spontaneous investments that has nothing to do with the national income. Then, $y^I(t) = q+h NI(t-1)$.

If depreciation isn't taken into account, GNP $(t-1)$ can be instead of the previous year's national income $NI(t-1)$ [4]. Thus there are: $y^C(t) = a+b GDP(t-1)$; $y^I(t) = q+h GDP(t-1)$. Then $GDP(t) = (b+h) GDP(t-1) + a + q + y^O + y^S - y^I$.

According to the input-output model of Leontif and income determination theory of macroeconomic, the relationship between income of last year and domestic demand, as well as the relationship between the demand and GDP of this year, together determine the growth of GDP. This model connects the GDP in year t with the GDP in year $(t-1)$, the spontaneous investment, the spontaneous consumption, as well as the net out exports and stock, which can clearly express the denemic process of the GDP growing.

Based on the model, the differential economic growth curve can be drawn, with $GDP(t)$ as the dependent variable, $b+h$ as the explanatory variables, a, q, y^O, y^S as the intercept, $GDP(t-1)$ as the slope, to reflect the true relationship between the growth trends of variable $GDP(t)$ based on $GDP(t-1)$ and other variables, such as spontaneous consumption a , spontaneous investment q , propensity to consume b , propensity to investment h , net exports y^O , stock y^S .

- (1) $(h+h)$ is the main indicator affecting the dependence of GDP on the previous year's GDP. While, the sum of b and h generally are oscillating around 1. And in a stable economic system, $(b+h)$ should be less than 1[5]. While expansionary economic policies will lead to $(b+h) > 1$. According to the model, $\Delta GDP(t) = \Delta(b+h) \cdot GDP(t-1)$. And the growth rate of GDP, can be expressed as $[GDP(t)-GDP(t-1)] \cdot 100\% / GDP(t-1)$. So there is :The growth rate of $GDP = (b+h-1) \cdot 100\% + (a+q+y^O+y^S) \cdot 100\% / GDP(t-1)$. Thus, the distance between $(b+h)$ and 1 is the main indicator affecting the growth of GDP, because of the expansion effect of percentage, tiny divation will cause significant chages of GDP growth rate.
- (2) $(a+q+y^O+y^S)$ expresses the intercept of the curve, which can affected the curve to move parallelly and affected the statistics of GDP in this year. Spontaneous consumption a is proposed In the income-expenditure model, refers to the consumption which has nothing to do with the income. Similarly, spontaneous investment q means the investment which has nothing to do with the income. Net exports y^O is the main part of GDP, which can be controlled by policy. The stock express the difference between the supply and the demands, which is a small part in GDP.
- (3) The investment structure, consumption structure, export structure, and imports structure, although do not performance in the model, but the changes of these factors will directly affect the propensity to consume, propensity to investment, and total net exports and inventories.

Table 163.1 Model test data Unit: billion yuan

Years	GDP	Consumption	Investment	Stock	Net expots
2002	120332.6893	71816.5	43632.1	1932.9	3094.1
2003	135822.7561	77685.5	53490.7	2472.3	2986.3
2004	159878.3379	87552.6	65117.7	4050.7	4079.1
2005	184937.369	99051.3	74232.9	3,624	10223.1
2006	216314.4259	112631.9	87954.1	5,000	16,654
2007	265810.3058	131510.1	103948.6	6994.6	23380.6
2008	314045.4271	152346.6	128084.4	10240.9	24229.4

Data Source: Calculated from National Statistical Yearbook 2010

163.3 The Reliability Empirical Test of Economic Growth Difference Model

Most of the models exist a gap between theory and reality, and often have large deviation in the practical application. Growth Differential Model reflects the relationship between GDP (t) and other economic indicators. Put other economic indicators as the known variables into the model to calculate GDP (t), and compare with the actual GDP. According to the error, the reliability of the model can be tested.

Data acquisition

The data between 2002 and 2009 will be used in the study, and the previous year's GDP, net exports and inventory can be obtained from the National Statistical Yearbook. While a , q , b and h , can not be directly from the Yearbook. They can be got respectively from the regression model: $y^C = a + b \text{GDP}(t-1) + \varepsilon$; $y^I = q + h \text{GDP}(t-1) + \varepsilon$.

According to the regression model, use AR (P) method in EViews. There is:

$$Y^I = -10466.22 + 0.5508 * \text{GDP}(t-1) + u(t); u(t) = 0.755683 u(t-1) + \varepsilon_t$$

Then, $q = -10764.28$, $h = 0.5304$

$$y^C = 18677.32 + 0.495 \text{GDP}(t-1) + u(t); u(t) = 0.588384 u(t-1) + \varepsilon_t$$

Then, $a = 22904.43$, $b = 0.474$

(B) Model checking

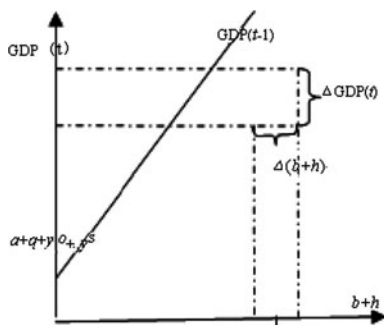
Put the other economic indicators which have been calculated into the model, calculate the GDP of next year, and compare it with the actual values to calculate the relative error. Relative error equals (actual value—predicted value)/actual value, reflecting the degree of deviation between actual value (Tables 163.1 and 163.2).

Compared the forecast with the actual value of 2002–2008, with the exception of 2002, relative errors are about 1%. GDP growth trend is almost consistent with the actual growth trend (see Fig. 163.1). It proves that the results and the actual is quite consistent, and the model is reliable. In the model tests, there are several way to eliminate the error. First, as the lack of direct data of a , q , b and h , they are

Table 163.2 :The predicted value of GDP and relative error

Year	2002	2003	2004	2005	2006	2007	2008
Predicted value	127352.284	138513.01	156749.1	186638.3	219625.3	259875.2	313705.4
Relative error	-0.0583349	-0.019807	0.01957	-0.0092	-0.01531	0.022328	0.001083

Fig. 163.1 Economic growth curve



calculated by fitting the regression in this paper, reflecting an approximation of the past few years, are not obtained for each year. If the correct a , q , b and h can be put into the model, the calculation of the data will be more accurate. Second, the depreciation in the model is negligible, which can bring a certain degree of error. If this problem can be considered in the later study, this part of the error can be eliminated (Fig. 163.2).

163.4 The Theoretical Significance of Economic Growth Difference Model

This model reflects the short-term economic growth patterns. According to the model, under current environment, the short-term trends and process of economic development in a country or a region can be analyzed. And the level of economic growth can be macro-controlled, by the optimal design of investment and consumer orientation, import and export.

- (1). Economic Growth Differential Model exceeded the existing limitations of current theory. The existing theory of economic growth mainly reflects the general quantitative relationship between the rate of economic growth and other basic variables under the conditions of balanced growth, with more assumption and more bias, and the content is relatively limited. Thus they will not be able to target economic strategy and provide implementation policy plan. This theory of Economic Growth Differential Model is simple, with less

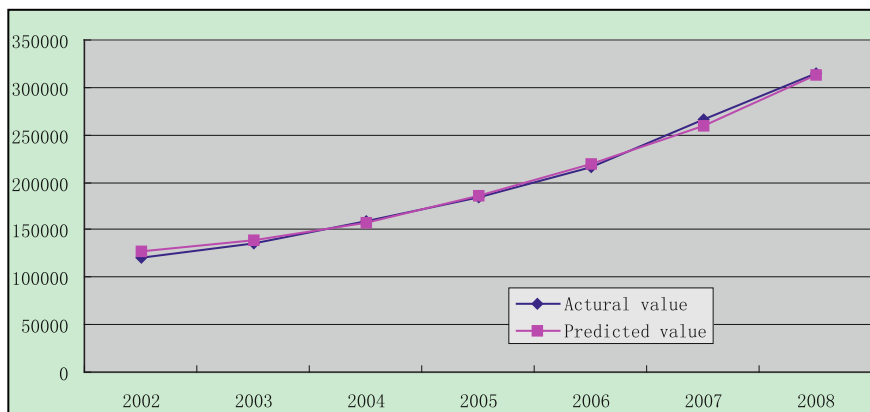


Fig. 163.2 The actual value and predicted value of GDP

- assumption and tightly with the real life. So it can help to judge the economic situation and provide strategic according to economic goals.
- (2). According to the theory that the ideology motives of economic growth is society's needs, economic growth differential model clearly describes the path of economic growth. The model adjusts the impaction of previous GDP on this year's GDP by adjusting b and h , adjusts the amount of GDP through spontaneous investment and consumption, as well as net exports. The model describes the conversion of the previous income to national income this year, and the conversion of domestic and external demand to the gross domestic product, thus achieving to forecast and plan the goal of GDP in this year. In the analysis of the economic situation, this model can be used to forecast trends and changes of economic growths, and be of great help to achieve economic growth targets directly through the regulation of various economic parameters, according to the environment of economic and national economic policy.
 - (3). Economic Growth Differential Model blends the economic indicators, such as tendency of investment and consumption, spontaneous investment and consumption, exports and so on. It reflects the role of accumulation, consumption, investment and exports, and reflects the ideas of the expanded reproduction. Not only predicts the rate of economic growth, but also deepened the role of the production process such as reproduction, distribution, investment, consumption, and exports in economic growth. At the same time, it also helps to provide a theoretical basis for the macro-control of economic indicators, to enable Countries achieve the objective of economic growth under the premise of reasonable balance between investment and consumption, as well as export.

163.5 Dynamic Analysis of Our Economic Recovery Strategy

In 2009, our government implemented a series of economic stimulus policies to promote the our economy, and GDP growth rate finally reached 8.7%. However, what is the driving force of the repaid growth, will it continue to maintain, and will this economic growth mechanism in 2009 can last forever, are the questions that we must analyze and study.

According to the model: $GDP(t) = (b + h)GDP(t-1) + a + q + y^0 + y^s - y^I$, There is:

$$\Delta GDP(t) = \Delta(b + h) \cdot GDP(t-1) + \Delta a + \Delta q + \Delta y^0 + \Delta y^s - \Delta y^I \quad (163.2)$$

So, the growth rate of GDP depends on the changes of propensity to consume and invest, spontaneous consumption and investment, as well as import and export. First of all, China's net exports in 2009 is amounted to 15,033.3, with the decrease of 919.61 billion yuan compared to 2008, resulting in the fact GDP falls about 919.61 billion yuan. Secondly, according to $b = \Delta y^c(t)/\Delta GDP$, the propensity to consume can be calculated as 0.498091 in 2009, with an increase of 0.066 points compared to 2008, which contributes 20,727 trillion yuan ($0.066 * GDP(t-1)$) for the GDP Growth. Third, according to $h = \Delta y^i(t)/\Delta GDP$, the propensity to invest can be calculated as 1.080644 in 2009, with an increase of 0.58 points compared to 2008, which contributes 182146.3 million ($0.58 * GDP(t-1)$) to GDP growth, almost more than half of the GDP level. Therefore, the economic policy and investment of four trillion tendency drives rapid increase in investment, and the propensity to invest is the main reason for economic recovery.

However, there are certain problems. First of all, there is an excessive investment growth and propensity to invest reached 1.080644 in 2009, casued by the formulation of economic policies. Such a high propensity to invest in the economic development process is rare. Thus the investment is too high and there isn't appropriate level of consumption to match, which will bring excess capacity and other issues and not conducive to the stability and development of long-term economic [6]; Second, the propensity to consume remains low. The propensity to consume is 0.474 in China in recent years. According to the historical practice of developed countries, the propensity to consume is too low in the period when per capita GDP belows 3,000 dollars[7]. The unreasonable relationship between the propensity to consume and investment, will turn to the result that high investment growth loses the support of final demand. And finally a large number of social product value can not be fulfilled, which will eventually lead to structural imbalances and economic turmoil [8].There fore, The growth of economic will face larger challenge. From the long term angel, the driving prower of economic should be shifted from investment to consumption. The export-oriented economy should be shifted to inward-oriented economic, which can help to start domestic

consumption driven model, promote rapid and sustainable steady growth of economic.

(1). Increase the propensity to consumption

First of all, increasing our propensity to consume, should be based on the trends and structure of demand. The industrial structure and product structure should be adjusted to meet the diversity consumption requirements of urban residents. Second, income distribution situation should be improved. Increase the income level of farmers and urban low-income persons, make sure the real office support work of minimum wage and minimum cost, and reduce the gap between rich and poor in China. Third, increase government's support of social security funds to enhance the sense of security of the residents, in order to promote consumption; Fourth, develop consumer credit, encourage the market of family car, encourage consumer installment credit for the big package goods in the nationwide, encourage excessive consumption [9].

(2). Gradually reducing the high propensity to invest. Gradually regulate Government's investment behavior and strengthen supervision of business investment. Government can use tax and credit measures such as leverage, make rational planning and guidance of industrial investment to avoid over-investment and waste of resources. Strictly supervise economic benefits and social benefits as well as resources consumption of investment projects, control and reduce the disorder of corporate investment behavior in the system environment. From the angel of the bank credit, improve the facilities of registration, approval and tax system. Strengthen the investment credit management, support the development of a conducive environment for the consumer credit business [10]. So that the propensity to invest will gradually come back to a reasonable level, and the proportion of investment and consumption will become reasonable gradually.

(3). Pull the spontaneous consumption and spontaneous investment

Spontaneous consumption and spontaneous investment are the part of domestic demand which are unrelated to the income changes. To improve our spontaneous consumption, there are several ways. On one hand, the product structure should be adjusted based on consumer preference. So that, the supply structure and the demand structure are consistent. On the other hand, the price levels of goods should be under reasonable regulation and strict supervision, floating in an acceptable price range. Furthermore, consumer culture should be promoted, in order to promote the enthusiasm of consumption. The level of spontaneous investment in China mainly depends on the policy, investors' preference, market environment, as well as the freedom and standardization of the market', etc. The policy of government, specially refers to the attitude and support of the government to the investors. The appropriate regulation and guidance of these factors is conducive to stimulate domestic demand, boost the level of investment, and thus promote GDP growth.

References

1. Fuqiang L, Zhiqing D, Linhui W (2008) The classification test of system led, elements contribution, and China's economic growth momentum [J]. *Econ Res* 4:53–59
2. Cao pp, Yuan F (2008) Dynamic analysis of China's economic growth [J]. *Soc Sci* 11:1–5
3. Xi L (2008) The power structure of American economic growth and dollar equilibrium exchange rate [J]. *Nankai Econ Stud* (4):142–145
4. Dornbusch R, Fischer S, Momtaz MR (2007) *Macroeconomics* [M], vol 11. China Renmin University Press, Haidian district, pp 24–26
5. Decheng F, Xisong L (2002) The dynamic input-output model [J]. *China Manag Sci* 5:42–45
6. Fangmei W (2006) Discussion on the consumption rate, investment rate and China's economic growth. *Econ issues* 9:9–13
7. Jianwei L (2007) The international comparison of evolution characteristic of investment rate and consumption rate. *China's Financial* 8:49–52
8. Decheng F, Liu K (2009) The study of the role Marginal propensity to consume plays on optimizing China's three industries structure [J]. *Econ Bus Manag* 11:48–51
9. Wentao Y (2006) The recommendations and causes of high investment rate and low consumption rate. *Macroecon Manag* 6:18–19
10. Dongli L (2006) China's high investment rate and low consumption rate of the phenomenon of Nanjing normal university (social sciences). 1:43–46

Chapter 164

A Study of Leading Industries Selection in Comprehensive Cities Based on Factor Analysis

Faming Zhang and Zhaoqing Ye

Abstract Combining theories on leading industries selection and taking into consideration of the characteristics of comprehensive cities, the paper establishes a leading industries selection index system of comprehensive cities. Factor analysis model is used to analyze the leading industries selection in comprehensive cities, and an empirical study of Hangzhou is carried out. The study shows Hangzhou should take financial industry, chemical raw material, chemicals manufacturing, mechanical manufacturing and intelligence transmission as its leading industries.

Keywords Factor analysis · Comprehensive cities · Leading industries · Selection

164.1 Introduction

With the development of economy, some cities in China, due to their superior geographic positions and obvious industrial advantages, have gained more rapid development, and become regional, national, even international economic and trade centers. They become comprehensive cities. But with the rapid development and requirement of industrial upgrading, they enter a period of sluggish development. Then the selection of leading industries is vital for them. The paper, combining the theories on leading industries selection [1–8] and taking into consideration characteristics of comprehensive cities, establishes a leading

F. Zhang (✉)

School of Economics and Management Nanchang University, Nanchang, China

Z. Ye

School of Foreign Languages Nanchang University, Nanchang, China

industries selection index system of comprehensive cities. Factor analysis [9, 10] model is used to analyze the leading industries selection in comprehensive cities, and an empirical study of Hangzhou is carried out. The study shows Hangzhou should take financial industry, chemical raw material, chemicals manufacturing, mechanical manufacturing and intelligence transmission as its leading industries.

164.2 Establishing the Index System

Based on the broad connotation of leading industries is very, and characteristics of diversity, complexity and concentration of comprehensive cities, a leading industries selection index system of comprehensive cities is constructed, following the principles of systematic, brief, feasible, universal, objective, and authentic. The indexes with high generality, easy access, and large quantity of information are chosen and screened. Finally the following indexes in Table 164.1 are selected.

164.3 Empirical Study

164.3.1 Factor Analysis Calculation

Based on the previous analysis, we conduct factor analysis in leading industries selection of Hangzhou. First, the basic conditions of leading industries selection of Hangzhou are analyzed. Second, taking into account the current state of industrial development, infrastructure and human Capital of Hangzhou, and including strategic new industries in the selection scope, we seek out the industries with potential advantages, on the list. Based on the ordering of total annual sales, we decide on a list of ten categories of industries as leading industries selection scope. They are textile industry, machinery manufacturing, wholesale and retail, paper-making and paper products industry, chemical raw materials and chemical products manufacturing, food manufacturing and processing industry, finance, real estate, clothing and leather products industry, information transmission, computer services and software industry. Third, by factor analysis, we establish a mathematical model and solve it to obtain the comprehensive assessment values of the ten industries. At last, we order the values and determine the leading industries, considering the regional resources endowment and the economic development conditions.

The economic and technologic indicators of ten industries in Hangzhou are selected. According to the evaluation index system of leading industries selection and the formula established above, and based on the Statistical Yearbook 2008 and 2009 of Zhejiang Province and Hangzhou city, the following basic data are obtained, as shown in Table 164.2.

Table 164.1 Leading industries selection index system of comprehensive cities

Datum layer	Index layer	Formula	Formula description
Industrial development elasticity	Income elasticity of demand	$x_1 = \frac{vQ_i}{vN_i} = \frac{Q_i}{N_i}$ $(i = 1, 2, \dots, n)$	$\frac{vQ_i}{Q_i}$ is the demand growth rate of the product of industry i; $\frac{vN_i}{N_i}$ is the national income per capita growth rate
	Industry growth rate	$x_2 = \sqrt[n]{\frac{Y_{it}}{Y_{i0}}} - 1$ $(i = 1, 2, \dots, n)$	Y_{it} is the demand of industry i in reporting period; Y_{i0} is the demand of industry i in base period; n is the gap year between the reporting period and base period
Comparative superiority of regional industries	Location quotient	$x_3 = \frac{Y_{ij}}{Y_{jt}} = \frac{Y_{ji}}{Y_{ni}} = \frac{Y_{ij}}{Y_{nt}}$ $(i = 1, 2, \dots, n)$	Y_{ij} is employment population or the added value of i industry in region j; Y_{jt} is the total employment population or the added value in region j; Y_{ni} is employment population or the added value in the country; Y_{nt} is the total employment population or the added value in the country
	Comparative labor productivity	$x_4 = \frac{g_i}{I_i} = \frac{g}{I}$ $(i = 1, 2, \dots, n)$	g_i and I_i represent the national income and labor population of leading industries in region i respectively; g and I represent the total national income and labor population in region i respectively
Technological progress	Labor productivity	$x_5 = \frac{vY_i}{L_i}$ $(i = 1, 2, \dots, n)$	vY_i is the added value of industry i; L_i is the employment population of industry i
	Increased rage of labor productivity	$x_6 = \sqrt[n]{\frac{L_{it}}{L_{i0}}} - 1$ $(i = 1, 2, \dots, n)$	L_{it} is the labor productivity of industry i in reporting period; L_{i0} is the labor productivity of industry i in base period N is the gap year between the reporting period and base period
Existing conditions	Employment absorption rate	$x_7 = \frac{L_i}{Y_i}$	L_i is the total employment population of industry i; Y_i is the total value of industry i

Then the basic data are nondimensionalized and Table 164.3 is obtained.

SPSS 17.0 is used to analyze the data in Table 164.3, and correlation matrix and communalities of variables are obtained. The results show that the correlation among original variables is greater and the communalities of various variables are large (all more than 90%), suggesting that when the variable space is transformed

Table 164.2 The index value of various industries in Hangzhou

Industry	X1	X2	X3	X4	X5	X6	X7
Y1	0.95	0.09	1.36	1.56	1.03	-0.94	13.07
Y2	0.97	0.09	0.46	3.34	2.25	-0.97	28.03
Y3	2.54	0.23	0.57	1.84	2.88	-0.88	15.49
Y4	5.93	0.53	0.38	4.77	13.94	-0.89	40.06
Y5	3.71	0.33	0.50	2.74	5.77	-0.88	23.04
Y6	0.55	0.05	0.86	0.93	0.37	-0.95	7.79
Y7	1.97	0.18	1.50	0.88	1.11	-0.80	7.38
Y8	1.60	0.14	3.11	0.73	0.78	-0.80	6.17
Y9	0.95	0.09	0.75	5.71	3.80	-0.98	47.99
Y10	2.77	0.25	0.33	7.16	12.01	-0.97	60.16

Note: Y1: textile industry; Y2: machinery manufacturing; Y3: papermaking and paper products industry; Y4: chemical raw materials and chemical products manufacturing; Y5: food manufacturing and processing industry; Y6: clothing and leather products industry; Y7: information transmission, computer services and software industry; Y8: wholesale and retail; Y9: real estate; Y10: finance

Table 164.3 Nondimensionalized data

Industry	X1	X2	X3	X4	X5	X6	X7
Y1	0.07	0.08	0.37	0.13	0.05	0.22	0.13
Y2	0.08	0.08	0.05	0.41	0.14	0.06	0.40
Y3	0.37	0.38	0.09	0.17	0.18	0.56	0.17
Y4	1.00	1.00	0.02	0.63	1.00	0.50	0.63
Y5	0.59	0.58	0.06	0.31	0.40	0.56	0.31
Y6	0.00	0.00	0.19	0.03	0.00	0.17	0.03
Y7	0.26	0.27	0.42	0.02	0.05	1.00	0.02
Y8	0.20	0.19	1.00	0.00	0.03	1.00	0.00
Y9	0.07	0.08	0.15	0.77	0.25	0.00	0.77
Y10	0.41	0.42	0.00	1.00	0.86	0.06	1.00

into factor space, much information is reserved. Thus the effect of factor analysis is remarkable. KMO test value is 0.619 (larger than 1.5), the chi-square value of Bartlett's sphericity test is 134.889, both significance levels are 0.000. It indicates that there is correlation among variables and they may share factors. Thus it is suitable to carry out factor analysis. As the factor loading matrix can not well reflect variables' load on share factors, we use maximum variance method to conduct an orthogonal rotation and get rotated factor loading matrix. And finally the R eigenvalue and contribution rate of the two main factors are obtained from factor model calculation (Table 164.3). From the variable correlation matrix we find two large characteristic roots, and the accumulative contribution rate reaches 89.76%. This suggests the two main factors contain the information of the original data. Following the principle of the value of characteristic root larger than 1, two shared factor are selected, and the information of the seven indicators can generally be represented (Tables 164.4 and 164.5)

Table 164.4 The total variance explained

Ingredient	Initial eigenvalues		
	Total	Variance %	Accumulation %
1	4.170	59.578	59.578
2	2.113	30.185	89.763
3	0.579	8.272	98.035
4	0.113	1.612	99.647
5	0.024	0.349	99.996
6	0.000	0.003	100.000
7	0.000	0.000	100.000

Table 164.5 Ingredient score coefficient matrix

	Ingredient	
	1	2
x1	-0.086	0.363
x2	-0.082	0.362
x3	-0.233	0.006
x4	0.267	0.012
x5	0.102	0.232
x6	-0.339	0.245
x7	0.266	0.013

According to scores of various industries' factors and the variance contribution rates of the main factors obtained by operating SPSS 17.0, we calculate the comprehensive scores of factors of different industries, as shown in Table 164.6.

164.3.2 Discussion

Priority is given to the five top industries in Table 164.6 as developing focus of the leading industries in Hangzhou. According to Table 164.6, these five industries are finance, chemical raw materials and chemical products manufacturing, real estate, machinery manufacturing, and food manufacturing and processing industry.

As the political center of Zhejiang Province, Hangzhou's city size, transportation, education and scale of investment are far better than those of other cities in the province; beside, it is close to China's financial center—Shanghai. All these elements contribute a superior environment for the development of finance industry in Hangzhou. Developing finance and high-tech industry is also in line with the regional industrial upgrading.

Chemical raw materials and chemical products manufacturing and machinery manufacturing are also competitive, due to the advantages of brands and talents. It is suggested that the development of these two industries combine with the strategic new industries recently promulgated by the government to seek a better development way.

Table 164.6 Score of industry factor and ranking

Industry	Factor score	Ranking
Finance	1.21	1
Chemical raw materials and chemical products manufacturing	0.92	2
Real estate	0.54	3
Machinery manufacturing	0.14	4
Food manufacturing and processing industry	0.14	5
Papermaking and paper products industry	-0.19	6
Textile industry	-0.41	7
Clothing and leather products industry	-0.45	8
Information transmission	-0.78	9
Wholesale and retail	-1.13	10

Real estate ranks third in the table, indicating its rapid development in the past two years. However, it is not recommended to treat this industry as a leading one. Although real estate develops quickly in recent years, it is mainly due to the influence of macroeconomic situation. Besides, real estate industry is frequently affected by the national policy regulation, and is closely related to livelihood and other social problems. Thus it is not appropriate to treat real estate as a regional leading industry and depend on it to drive economic growth.

The total annual sales of textile industry and clothing and leather products industry still rank high among all industries in Hangzhou. But from Table 164.6, we find its factor score is negative. As labor costs continue to increase, this labor-intensive industry has become unsuitable for development in the region. But it can take advantage of the existing size and brands of this industry, and make full use of the talents in Hangzhou, to keep design and other high added-value links in Hangzhou, while remove others links to some other places.

Information transmission industry ranks ninth in the table, but it is a high-tech industry, and can promote the development of fiancé industry. Thus it can also be treated as a leading industry.

164.4 Conclusions

Considered the leading industries selection index system established above and the local situations, the following industries should be taken as the leading industries in Hangzhou. They are fiancé, Chemical raw materials and chemical products manufacturing, machinery manufacturing, and information transmission.

The selection of these four industries as leading industries of Hangzhou, based on the leading industries selection index system of comprehensive cities, conforms to the economic developing rules and meets the requirement of industrial upgrading. It is also in line with the characteristics of comprehensive cities. Thus this index system is practical.

Acknowledgements The national natural science foundation of China (fund Numbers: 71001048); jiangxi province education science “The twelfth five-year plan” planning project (Numbers: 10YB222).

References

1. Nelson RR, Winter SG (1982) An evolutionary theory of economic change[M]. Belknap Press, Cambridge
2. Valente M, Andersen ES (2002) A hands_on approach to evolutionary simulation: Nelson & Winter models in the laboratory for simulation development[J]. *Electron J Evol Modeling Econ Dyn* 1:045–052
3. Tiaojun X, Zhaohan S, Shuping C (2002) Analysis of portrait enterprise group’s R&D and economy increase[J]. *J Manag Sci China* 5(4):1–6 (in Chinese)
4. Jonard N, Yildizoglu M (1999) Sources of technological diversity[C]. Paris: Cahiers de l’innovation, No. 99030, CNRS
5. Yildizoglu M (2002) Competing R&D strategies in an evolutionary industry model[J]. *Comput Econ* 19:51–65
6. Winter SG, Kaniovski YM, Dosi G (2003) A baseline model of industry evolution[J]. *J Evol Econ* 13:355–383
7. Winter SG, Kaniovski YM, Dosi G (2000) Modeling industrial dynamics with innovative entrants[J]. *Struct Chang Econ Dyn* 11:255–293
8. Kwasnicki W (2001) Comparative analysis of selected neo_schumpeterian models of industrial dynamics[R]. Druid, Aalborg
9. Reynolds DA (2003) Channel robust speaker verification via feature mapping. In: *Proceedings of the IEEE international conference on acoustics, speech and signal processing*, vol II. Hongkong, pp 53–56
10. Teunen R, Shahshahani B, Heck L (2000) A model-based transformational approach to robust speaker recognition. In: *Proceedings of the 6th international conference on spoken language processing*, vol II. Beijing, pp 495–498

Chapter 165

A Programme-Oriented Algorithm to Compute $\text{Ord}(f(x))$

Lihua Liu and Zhengjun Cao

Abstract Computing the order of a polynomial over finite fields is of importance for a variety of applications. The general algorithm for the problem is not iterated. The drawback incurs more cost and memory. Moreover, the programming work becomes more tedious. In this paper, we present an iterated algorithm to compute the order of an irreducible polynomial over finite fields. The algorithm captures good efficiency.

Keywords Finite fields · Repeated-squaring algorithm

165.1 Introduction

Computing the order of a polynomial over finite fields is of importance for a variety of applications, such as, investigating the ordericity properties of linear recurring sequences, generating cyclic codes [1, 2]. We refer to Ref. [3, 4] for the general algorithm to compute the order of an irreducible polynomial $f(x)$ of degree n over a finite field \mathbb{F}_q . The algorithm is not iterated. Concretely speaking, given the value of

$$x^{(q^n-1)/p_i^t} \pmod{f(x)}$$

L. Liu (✉)

Department of Mathematics, Shanghai Maritime University, Shanghai, China
e-mail: lhliu@yahoo.cn

Z. Cao

Department of Mathematics, Shanghai University, Shanghai, China
e-mail: caoamss@gmail.com

one can not recursively use the value to compute

$$x^{(q^n-1)/p_i^{t+1}} \pmod{f(x)}$$

where $q^n - 1 = p_1^{k_1} \dots p_r^{k_r}$, p_1, \dots, p_r are r distinct primes and k_1, \dots, k_r are r positive integers. The drawback incurs more cost.

In this paper, we present an iterated algorithm to compute the order of an irreducible polynomial over finite fields. The algorithm is based on the observation that

$$p(f) = p_1^{s_1} \dots p_r^{s_r}$$

where s_i is the least nonnegative integer such that

$$\left(x^{(q^n-1)/p_i^{k_i}}\right)^{p_i^{s_i}} \equiv 1 \pmod{f(x)}$$

for all $i = 1, \dots, r$. Since there exists the following iteration,

$$\left(x^{(q^n-1)/p_i^{k_i}}\right)^{p_i^{t+1}} \equiv \left(\left(x^{(q^n-1)/p_i^{k_i}}\right)^{p_i^t} \pmod{f(x)}\right)^{p_i} \pmod{f(x)}$$

it becomes easy to determine each s_i , for all $i = 1, \dots, r$. Thus, the new algorithm is efficient.

165.2 Preliminary

Let p be a prime, $q = p^j$, where j is a positive integer. We know the characteristic of the finite field \mathbb{F}_q is p . The following definition and results can be found in Ref. [4].

Definition 1 Let $f(x)$ be a polynomial over \mathbb{F}_q of degree ≥ 1 and with a nonzero term. The order of $f(x)$ is defined to be the least positive integer l such that

$$f(x)|(x^l - 1)$$

We use $p(f)$ to denote the order of $f(x)$. Hence, if $f(x)|(x^l - 1)$, then $p(f)|l$.

Theorem 1 Let $f(x)$ be a polynomial over \mathbb{F}_q of degree n and with a nonzero term. Then

$$p(f)|(q^n - 1)$$

and

$$\gcd(p(f), p) = 1$$

Theorem 2 Let $f(x)$ be a polynomial over \mathbb{F}_q of degree ≥ 1 and with a nonzero term and assume that $f(x) = f_1(x)f_2(x)$, where $\text{deg}f_1(x), \text{deg}f_2(x) \geq 1$, and $\text{gcd}(f_1(x), f_2(x)) = 1$, where $\text{gcd}(f_1(x), f_2(x))$ denotes the greatest common divisor of $f_1(x)$ and $f_2(x)$. Then

$$p(f) = \text{lcm}[p(f_1), p(f_2)]$$

where $\text{lcm}[p(f_1), p(f_2)]$ denotes the least common multiple of $p(f_1)$ and $p(f_2)$.

Theorem 3 Let $f(x)$ be an irreducible polynomial over $\mathbb{F}_q, f(0) \neq 0$, and let e be an integer ≥ 1 . Then

$$p(f^e) = p(f) \min\{p^i : p^i \geq e\}$$

where $\min\{p^i : p^i \geq e\}$ denotes the smallest power of p which is $\geq e$.

Theorem 4 Let $f(x)$ be a polynomial over \mathbb{F}_q with a nonzero term and assume that

$$f(x) = f_1(x)^{e_1} f_2(x)^{e_2} \dots f_r(x)^{e_r}$$

where $f_1(x), \dots, f_r(x)$ are r distinct irreducible polynomial over \mathbb{F}_q and e_1, \dots, e_r are r positive integers, then

$$p(f) = \text{lcm}[p(f_1), p(f_2), \dots, p(f_r)] \min\{p^t : p^t \geq e_1, \dots, e_r\}$$

where $\min\{p^t : p^t \geq e_1, \dots, e_r\}$ denotes the smallest power of p which is $\geq e_1, \dots, e_r$.

By Theorem 4, to determine the order of a polynomial $f(x)$ over \mathbb{F}_q of degree ≥ 1 and with a nonzero constant term, we may proceed in two steps. The first step is to factorize $f(x)$ into a product of irreducible polynomials in \mathbb{F}_q . Concerning this topic, Berlekamp's algorithm [5] can be applied. The second step is to determine the orders of the irreducible factors of $f(x)$. As for the topic, we have

Theorem 5 Let $f(x)$ be an irreducible polynomial of degree n over \mathbb{F}_q with a nonzero constant term and let

$$q^n - 1 = p_1^{k_1} p_2^{k_2} \dots p_r^{k_r}$$

where p_1, \dots, p_r are r distinct primes and k_1, \dots, k_r are r positive integers. Then for each $i = 1, \dots, r$, there is a nonnegative integer $t_i \leq k_i$ such that

$$x^{(q^n - 1)/p_i^{t_i}} \equiv 1 \pmod{f(x)}$$

$$x^{(q^n - 1)/p_i^{t_i + 1}} \not\equiv 1 \pmod{f(x)}$$

Consequently,

$$p(f) = p_1^{k_1 - t_1} p_2^{k_2 - t_2} \dots p_r^{k_r - t_r}$$

The general algorithm for OIP

Input: an irreducible polynomial $f(x)$ of degree n over \mathbb{F}_q with a nonzero constant term

- 1) Factorize $q^n - 1$. Assume that $q^n - 1 = p_1^{k_1} p_2^{k_2} \dots p_r^{k_r}$, where p_1, \dots, p_r are r distinct primes, k_1, \dots, k_r are r positive integers.
- 2) Calculate $x^{(q^n-1)/p_i} \pmod{f(x)}$, $x^{(q^n-1)/p_i^2} \pmod{f(x)}$, $x^{(q^n-1)/p_i^3} \pmod{f(x)}$, ..., $x^{(q^n-1)/p_i^3} \pmod{f(x)}$, ..., for $i = 1, \dots, r$, until a nonnegative integer $t_i \leq k_i$ such that $x^{(q^n-1)/p_i^{t_i}} \equiv 1 \pmod{f(x)}$, $x^{(q^n-1)/p_i^{t_i+1}} \not\equiv 1 \pmod{f(x)}$.

Output: $p(f) = p_1^{k_1-t_1} p_2^{k_2-t_2} \dots p_r^{k_r-t_r}$

For convenience, we use the abbreviation OIP for the order of an irreducible polynomial over finite fields, henceforth.

165.3 The General Algorithm for OIP

The following description of the general algorithm to compute the order of an irreducible polynomial over finite fields, comes from [4]:

In order to facilitate the computation of

$$x^{(q^n-1)/p_i^t} \pmod{f(x)}, t = 1, 2, \dots; i = 1, 2, \dots, r$$

one can proceed as follows:

2.1). Calculate

$$x^0 \pmod{f(x)}, x^q \pmod{f(x)}, x^{2q} \pmod{f(x)}, \dots, x^{(n-1)q} \pmod{f(x)}$$

Assume that

$$x^{jq} \equiv \sum_{i=0}^{n-1} a_{ij} x^i \pmod{f(x)}, j = 0, 1, \dots, n-1$$

Let

$$A = \begin{pmatrix} a_{00} & a_{01} & \cdots & a_{0,n-1} \\ a_{10} & a_{11} & \cdots & a_{1,n-1} \\ \vdots & \vdots & & \vdots \\ a_{n-1,0} & a_{n-1,1} & \cdots & a_{n-1,n-1} \end{pmatrix}$$

which is an $n \times n$ matrix over \mathbb{F}_q .

2.2). Calculate

$$x \pmod{f(x)}, x^q \pmod{f(x)}, x^{q^2} \pmod{f(x)}, \\ \dots, x^{q^{n-1}} \pmod{f(x)},$$

Note that when $x^{q^i} \pmod{f(x)}$ has been calculated, assume that

$$x^{q^i} \equiv \sum_{j=0}^{n-1} b_j x^j \pmod{f(x)},$$

then

$$x^{q^{i+1}} \equiv \sum_{i=0}^{n-1} \left(\sum_{j=0}^{n-1} a_{ij} b_j \right) x^i \pmod{f(x)},$$

That is, if we let

$$x^{q^{i+1}} \equiv \sum_{j=0}^{n-1} b'_j x^j \pmod{f(x)},$$

We have

$$\begin{pmatrix} b'_0 \\ b'_1 \\ \vdots \\ b'_{n-1} \end{pmatrix} = A \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{n-1} \end{pmatrix}$$

2.3). Express $(q^n - 1)/p_i^t$ as a q -ary number. Assume that

$$(q^n - 1)/p_i^t = a_0 + a_1 q + \dots + a_{n-1} q^{n-1}, 0 \leq a_i \leq q - 1$$

2.4). Calculate

$$x^{(q^n - 1)/p_i^t} \equiv x^{a_0} (x^q)^{a_1} (x^{q^2})^{a_2} \\ \dots (x^{q^{n-1}})^{a_{n-1}} \pmod{f(x)}$$

Remark 1 Given the value of $x^{(q^n-1)/p_i'} \pmod{f(x)}$, one can not recursively use the value to compute $x^{(q^n-1)/p_i^{r+1}} \pmod{f(x)}$. The drawback incurs more cost and memory. Moreover, the programming work becomes more tedious.

165.4 A Programme-Oriented Algorithm for OIP

In this section, we propose a new algorithm for PIP.

A new algorithm for OIP

Input: an irreducible polynomial $f(x)$ of degree n over \mathbb{F}_q with a nonzero constant term

- 1) Factorize $q^n - 1$. Assume that $q^n - 1 = p_1^{k_1} p_2^{k_2} \dots p_r^{k_r}$, where p_1, \dots, p_r are r distinct primes, k_1, \dots, k_r are r positive integers.
- 2) Compute $\left(x^{(q^n-1)/p_i^{k_i}} \pmod{f(x)}\right), \left(x^{(q^n-1)/p_i^{k_i}}\right)^{p_i} \pmod{f(x)}, \left(x^{(q^n-1)/p_i^{k_i}}\right)^{p_i^2} \pmod{f(x)}, \dots, \left(x^{(q^n-1)/p_i^{k_i}}\right)^{p_i^{s_i}} \pmod{f(x)}, \dots$, for $i = 1, \dots, r$, until a nonnegative integer $s_i \leq k_i$ such that $\left(x^{(q^n-1)/p_i^{k_i}}\right)^{p_i^{s_i}} \equiv 1 \pmod{f(x)}$.

Output: $p(f) = p_1^{s_1} p_2^{s_2} \dots p_r^{s_r}$

The following theorem is the base of the new algorithm.

Theorem 6 *Let $f(x)$ be an irreducible polynomial of degree n over \mathbb{F}_q with a nonzero constant term and let*

$$q^n - 1 = p_1^{k_1} p_2^{k_2} \dots p_r^{k_r}$$

where p_1, \dots, p_r are r distinct primes and k_1, \dots, k_r are r positive integers. Then for each $i = 1, \dots, r$, there is the least nonnegative integer $s_i \leq k_i$ such that

$$\left(x^{(q^n-1)/p_i^{k_i}}\right)^{p_i^{s_i}} \equiv 1 \pmod{f(x)},$$

Consequently,

$$p(f) = p_1^{s_1} p_2^{s_2} \dots p_r^{s_r}$$

Proof Since $q^n - 1 = p_1^{k_1} p_2^{k_2} \dots p_r^{k_r}$, $p(f) | q^n - 1$, there exist r nonnegative integers s_1, \dots, s_r , such that $p(f) = p_1^{s_1} p_2^{s_2} \dots p_r^{s_r}$. Clearly, $0 \leq s_i \leq k_i$, for all $i = 1, \dots, r$. Since

$$\begin{aligned} ((q^n - 1)/p_i^{k_i})p_i^{s_i} &= p_1^{k_1} \dots p_i^{s_i} \dots p_r^{k_r}, \\ p_1^{s_1} \dots p_i^{s_i} \dots p_r^{s_r} &| p_1^{k_1} \dots p_i^{s_i} \dots p_r^{k_r} \end{aligned}$$

we have

$$\left(x^{(q^n-1)/p_i^{k_i}}\right)^{p_i^{s_i}} \equiv x^{p_1^{k_1} \dots p_i^{s_i} \dots p_r^{k_r}} \equiv 1 \pmod{f(x)}$$

Suppose that there is a nonnegative integer λ_i such that

$$\lambda_i < s_i, \text{ and } \left(x^{(q^n-1)/p_i^{k_i}}\right)^{p_i^{\lambda_i}} \equiv 1 \pmod{f(x)},$$

By the definition of $p(f)$, we have

$$p_1^{s_1} \dots p_i^{s_i} \dots p_r^{s_r} | p_1^{k_1} \dots p_i^{\lambda_i} \dots p_r^{k_r}$$

This leads to a contradiction, since p_1, \dots, p_r are r distinct primes.

165.5 Comparison

There is a dramatic difference between the general algorithm and the new algorithm. See the following table.

The general algorithm	The new algorithm
Compute	Compute
$x^{(q^n-1)/p_i} \pmod{f(x)},$	$\left(x^{(q^n-1)/p_i^{k_i}}\right) \pmod{f(x)}$
$x^{(q^n-1)/p_i^2} \pmod{f(x)},$	$\left(x^{(q^n-1)/p_i^{k_i}}\right)^{p_i} \pmod{f(x)}$
$x^{(q^n-1)/p_i^3} \pmod{f(x)}, \dots$	$\left(x^{(q^n-1)/p_i^{k_i}}\right)^{p_i^2} \pmod{f(x)}, \dots$
until $t_i \leq k_i$ such that	until $s_i \leq k_i$ such that
$x^{(q^n-1)/p_i^{t_i}} \equiv 1 \pmod{f(x)},$	$\left(x^{(q^n-1)/p_i^{k_i}}\right)^{p_i^{s_i}} \equiv 1 \pmod{f(x)}$
$x^{(q^n-1)/p_i^{t_i+1}} \not\equiv 1 \pmod{f(x)}$	
$p(f) = p_1^{k_1-t_1} p_2^{k_2-t_2} \dots p_r^{k_r-t_r}$	$p(f) = p_1^{s_1} p_2^{s_2} \dots p_r^{s_r}$

Notice that

$$\begin{aligned} & \left(x^{(q^n-1)/p_i^{k_i}}\right)^{p_i^{t_i+1}} \\ \equiv & \left(\left(x^{(q^n-1)/p_i^{k_i}}\right)^{p_i^{t_i}} \pmod{f(x)}\right)^{p_i} \pmod{f(x)} \end{aligned}$$

The iteration is more applicable to programme. The repeated-squaring algorithm works well in this setting [6].

To make a further cost comparison, we need [7]:

Theorem 7 *Let r be a ring, $f(x) \in R[x]$ be a polynomial of degree n , e be a nonnegative integer. $g(x)$ is in the residue class ring $R[x]/(f(x))$. Using the repeated-squaring algorithm, $g(x)^e \pmod{f(x)}$ can be done using $\mathcal{O}(n^2 \log_2 e)$ operations in R .*

In the general algorithm, the cost of step 2 is dominated by the computation

$$x^{(q^n-1)/p_i^{t_i}} \pmod{f(x)}, j = 1, \dots, t_i$$

for each $i = 1, \dots, r$. Roughly speaking, it requires

$$n^2 \left(\sum_{j=1}^{t_i} \log_2 (q^n - 1)/p_i^j \right)$$

operations in \mathbb{F}_q . That is, it can be done using

$$\mathcal{O}(t_i n^2 (n \log_2 q - t_i \log_2 p_i))$$

operations in \mathbb{F}_q to obtain t_i .

In the new algorithm, it is dominated by the following computation

$$\left(x^{(q^n-1)/p_i^{k_i}}\right)^{p_i^{s_i}} \pmod{f(x)}, j = 1, \dots, s_i,$$

for each $i = 1, \dots, r$. Using the above iteration, it requires

$$n^2 (\log_2 (q^n - 1)/p_i^{k_i} + s_i \log_2 p_i)$$

operations in \mathbb{F}_q . Notice that $s_i = k_i - t_i$. Thus, it can be done using

$$\mathcal{O}(n^2 (n \log_2 q - t_i \log_2 p_i))$$

operations in \mathbb{F}_q to obtain s_i . Apparently, the new algorithm captures good efficiency.

165.6 Conclusion

In this paper, we present a programme-oriented algorithm for computing the order of an irreducible polynomial over finite fields. The algorithm differs the general algorithm because it is iterated and captures good efficiency.

Acknowledgements We acknowledge the Sciences & Technology Program of Shanghai Maritime University.

References

1. Garrett P (2003) The mathematics of coding theory. Prentice-Hall, New Jersey
2. Roman S (1992) Coding and information theory. Springer, New York
3. Lidl R, Niederreiter H (1986) Introduction to finite fields and their applications. Cambridge University Press, Cambridge
4. Wan ZX (2003) Lectures on finite fields and galois rings. World Scientific, Lynwood
5. Berlekamp E (1968) Algebraic coding theory. McGraw-Hill, New York
6. Gathen J, Gerhard J (2003) Modern computer algebra. 2nd edn. Cambridge University Press, Cambridge
7. Shoup V (2005) A computational introduction to number theory and algebra. Cambridge University Press, Cambridge

Chapter 166

Finding the Academic Collaboration Chance in Open Research Community

Bin Xu, Huanyu Zhou, Weigang Chen and Yujia Ge

Abstract Due to the advance of information technology and the widely used social network on internet, research is being done in an open environment. While the researchers may share the published research achievements in paper format, it's essential for the research groups to find the potential academic collaboration chance between them and the large scale academic social network. This article is to present our solution to find the chance of academic collaboration between different research groups in such an open research community. A prototype has been made and the core algorithms and an experiment with main interfaces are introduced.

Keywords Social network · Research community · Academic collaboration · Collaborative innovation

166.1 Introduction

Facilitated by the advance of information technology and the widely used internet, research is being done in an open environment, which we name it “research community” in this article. While blog, Wikipedia, and professional search engines provide plenty information for research, researchers could share

B. Xu (✉) · H. Zhou · W. Chen · Y. Ge
College of Computer Science and Information Engineering,
Zhejiang Gongshang University, Hangzhou, China
e-mail: xubin@mail.zjgsu.edu.cn

H. Zhou
e-mail: zhouhytx@hotmail.com

achievements together, and they look forward to efficient collaborative innovation between them.

Collaborative innovation begins when universities and industries come together to solve problems and/or develop customer-centric solutions that are beyond the scope, scale or capabilities of the individual universities or companies. Collaborative innovation has been paid great attention by different domains during these year, people establish the networks and centers for collaborative innovation [1, 3–5].

Peter A. Gloor in his book [1] introduced “Swarm Creativity”, a collaborative innovation network. He recognized collaborative innovation network as a cyber-team of self-motivated people with a collective vision, enabled by technology to collaborate in achieving a common goal innovation by sharing ideas, information, and work. He explained the traits that characterize the network members and their behavior. His solutions include creation of self-motivated teams, collective vision, enabled by technology, common goal, and sharing of ideas, information and work.

Haibin Zhu introduced role-based collaboration model E-CARGO for collaboration [2]. He suggested to establishing the development/business environment as a role net. Each role provides a certain services and applies a certain services in the proposed role net. His solutions to build a more efficient collaborative system includes regarding roles as agent dynamics in multi-agent systems, reducing the impact of role transfer in emergency management systems.

The collaborative innovation with business partners, customers, consultants, associations and even competitors have many benefits such as access to markets and customers, higher quality, sharing of risk, financial and intellectual benefits, reduction in technology gap quickly, leveraging of shared infrastructure, significant scale, faster time-to-market and time-to-revenue and increased customer loyalty [3]. However, the pitfalls are also significant, which include

“Not Invented Here” Syndrome,
Co-ordination failure,
Risk perception, and
Opportunism and monitoring costs.

In this Chapter, the authors focus the research on establishing stronger relationship between other research groups with similar interest. The authors view the research community as a social network which made of research groups [6, 7]. Within each research group, the researchers are connected by similar research interest and domain knowledge. The authors will propose a model to value the similarity of research interest between two artifacts, and then determine the similarity of research interest between every two groups by analyzing the similarity of their artifacts. After that, collaboration graph will be generated and used to illustrate the research similarity between different research groups, high value relationship denotes large similarity between the two groups and low value denotes a small similarity between the two groups. Finally, the possible chance of academic collaboration can be found out so as to make the research activities more efficient.

There are some similar work has been done.

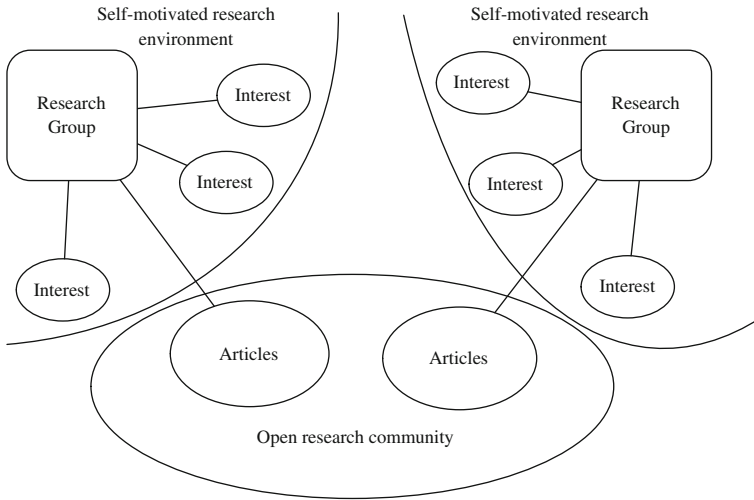


Fig. 166.1 A simple research environment model

The rest of this paper is organized as follows: [Section 166.2](#) models and defines the research environment, and [Sect. 166.3](#) will describe a core algorithm. Prototype will be introduced in [Sect. 166.4](#) and conclusion will be stated in [Sect. 166.5](#).

166.2 Definition of Research Environment

166.2.1 Research Environment Modeling

As demonstrated in [Fig. 166.1](#), research environment can be simply modeled as self-motivated research environment contain private research groups and open research community containing the journals, conference, blogs, Wikipedia and other published source. Each research group may have several research interests, and may publish a set of articles to the open research community. The research group know the articles other groups published but don't know the detail interests the other groups had.

166.2.2 Related Definitions

In order to find the collaborative chance between different research groups, the similar research interest should be identified and the research interests with largest similarity will be suggested to be a collaboration chance between two groups.

Because the research interests may not be opened to others, they must be determined by the analyzing on the published articles. In such way, research interest, academic article, research group should be defined in our research [8].

Definition 1 Research interest A research interest is a research direction belongs to a research group, it is defined as $ri:: = \langle id, dms, goals, keywords \rangle$, where

id is the identification of the research interest,
 dms is a set of domains the research group is working on,
 $goals$ is a set of goals the research group wants to achieve in the research work,
 $keywords$ is a set of keywords contained in related papers.

Definition 2 Academic article An academic article is the achievement published by a research group. It is defined as $aa:: = \langle id, dm, goal, achievement, title, abstract, keywords, reference \rangle$, where

id is the identification of the academic article,
 dm is the domain which the article is focused,
 $goal$ is the research goal of the article,
 $achievement$ is the achievements made by the work introduced in the article, which includes methodology, practice, framework, prototype, system and algorithms,
 $title$ is the title of the concrete paper,
 $abstract$ is the abstract of the concrete paper,
 $keywords$ is the a set of keywords listed in the concrete paper,
 $references$ is a serial of references listed in the concrete paper.

Definition 3 Research group A research group contains a set of research members, and has several research interests and published articles. It is defined as $rg:: = \langle id, name, ris, aas, members \rangle$, where

id is the identification of the research group,
 $name$ is the name of the group,
 ris is a set of research interests of the group,
 aas is a set of academic articles published by the group,
 $members$ refers to the staff of the group.

166.3 Core Algorithms for Collaboration Chance Finding

To find the collaboration chance, we need to calculate the similarity of the published articles and determine the most suitable research collaboration between different research groups. Here, the related algorithms will be designed to calculate the similarity of the text, the article and determine the most valuable research collaboration for group pairs.

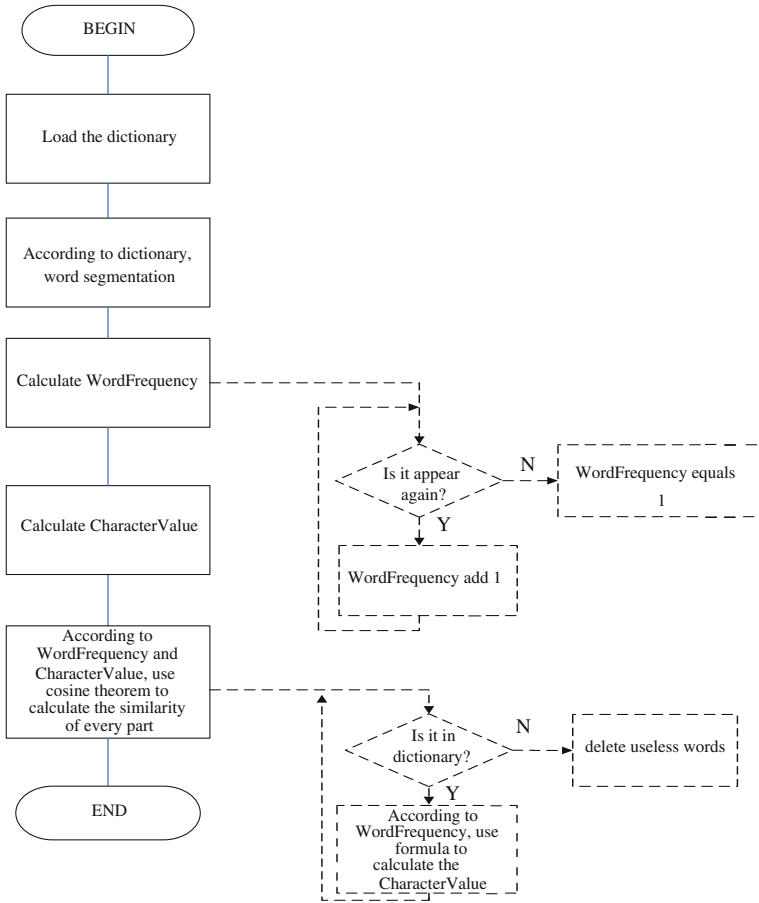


Fig. 166.2 The procedure of text similarity calculation

166.3.1 Procedure of Text Similarity Calculation

(Figure 166.2).

166.3.2 Algorithm for Article Similarity Calculation

Algorithm: article similarity calculation

Input: Article1, Article2; Factors // Two published articles to be compared. Factors are the adjust factors to the different parts of the article.

```

Output: SV_article // the similarity of the two articles.
Fetch title1, abstract1, keywords1, and reference1 from
article1.
Fetch title2, abstract2, keywords2, and reference2 from
article2.
SV_title = textSimilarity(title1, title2);
SV_abstract = textSimilarity(abstract1, abstract2);
SV_keywords = textSimilarity(keywords1, keywords 2);
num_of_ref1 = num_of_ref2 = 0;
FOR (each iref1 in reference1)
    num_of_ref1 = num_of_ref1 + 1;
FOR (each iref2 in reference2)
    num_of_ref2 = num_of_ref2 + 1;
min_num_of_ref = min(num_of_ref1, num_of_ref2);
same_ref = similar_ref = 0;
FOR (each iref1 in reference1)
    FOR (each iref2 in reference2)
        IF (textSimilarity(iref1, iref2) ==1)
            same_ref = same_ref + 1;
        ELSEIF
            (textSimilarity(iref1, iref2) >=lowestBoundary)
                similar_ref = similar_ref + 1;
SV_reference = (same_ref + similar_ref * min_factor_ref) /
min_num_of_ref;
SV_article = SV_title * factor_title
            + SV_abstract * factor_abstract
            + SV_keywords* factor_keywords
            + SV_reference * factor_reference;

```

This algorithm calculates the similarity of reference different from the similarity calculation of title, abstract and keywords. Fetch title, abstract, keywords and reference from article can be automatically operated but is manually input in our prototype. The function text similarity is the algorithm to calculate the similarity of two texts. Regarding the reference similarity calculation, different articles published by the same authors can be considered but is not realized in the prototype.

To be mentioned, the keywords similarity can also be calculated as that of reference. The entire procedure of the article similarity calculation is stated as Fig. 166.3.

After the article similarity has been calculated, the largest similarity article-pair can be calculated for each research group pair. The related research topic and research interests can be considered as the research chance between two research groups. Of course, the lowest boundary should be defined so as to remove the unnecessary research collaboration. Besides, when there are several similar research interests between the research group pair, we should reduce the granularity of research group down to research group and research interests.

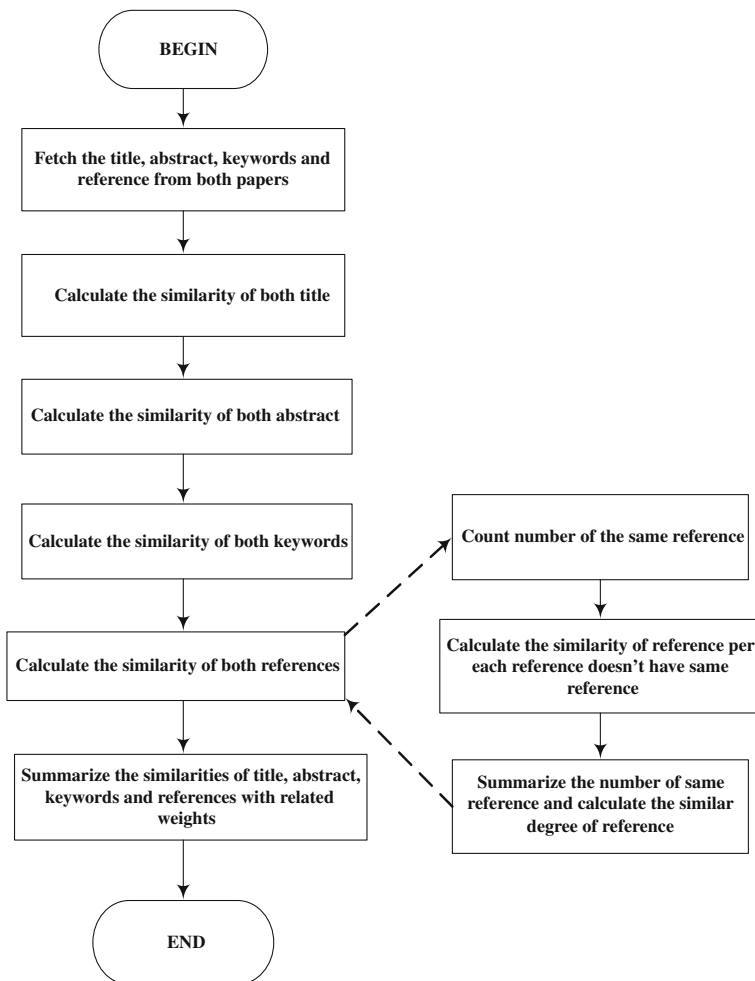


Fig. 166.3 The entire procedure of article similarity calculation

166.4 Experiment and Prototype

The authors have realized the algorithms of text similarity and article similarity calculation and developed a prototype for them.

The following screen snapshots illustrate the article similarity calculations (Figs. 166.4 and 166.5).

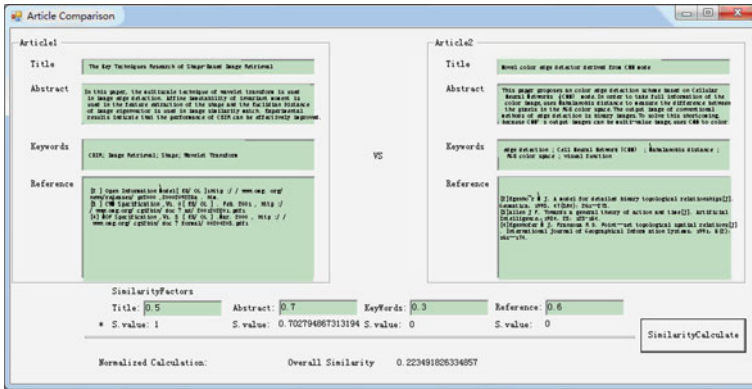


Fig. 166.4 Article similarity calculation (example for two low similarity articles)

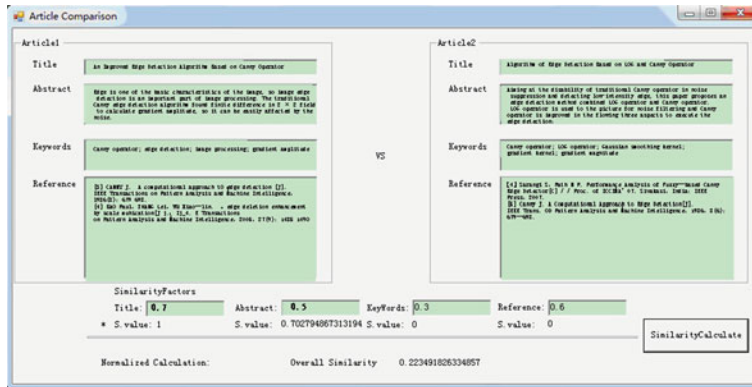


Fig. 166.5 Article similarity calculation (example for two high similarity articles)

166.5 Conclusion and Status of Research

The advantage and issues of collaborative innovation has been introduced. Academic research chance has been researched in this paper, and a prototype has been developed to demonstrate the features of the work. Research environment has been modeled and the core concepts have been defined, several core algorithms have been introduced or detail stated.

Currently, we are going to develop a framework with the involved algorithms and would like to try it in finding the possible research collaboration between several groups in computer science.

Acknowledgment This research is supported by Zhejiang Gongshang University, China with No. XGZ1102 to Dr. Weigang Chen. This research was also financial supported by National Natural Science Foundation of China with No. 60873022 to Dr. Hua Hu, The Science and Technology Department of Zhejiang Province, China with No. 2008C11009 to Dr. Bin Xu, and No. 2008C13082 to Dr. Xiaojun Li, and Education Department of Zhejiang Province with No. yb07035 to Prof. Yun Ling and No. 20061085 to Dr. Bin Xu.

References

1. Gloor PA (2006) *Swarm creativity : competitive advantage through collaborative innovation networks*. Oxford University Press, Oxford
2. Zhu H, Zhou MC (2006) Role-based collaboration and its Kernel mechanisms. *IEEE Trans Syst Man Cybern C* 36(4):578–589
3. Venkatesh G (2006) Collaborative innovation. In: *Proceedings of international conference on managing technological innovation in IT*, The Leela Palace, Bangalore, Keynote speech December 2006.
4. Mcdaniel EA (2003) Facilitating cross-boundary leadership in emerging E-government leaders, *Informing Science*
5. Klein M, Sayama H, Faratin P, Bar-Yam Y (2006) The dynamics of collaborative design: insights from complex systems and negotiation research. In: Braha D, Minai A, Bar-Yam Y (eds) *Complex engineered systems*. Springer, Berlin
6. Sebastiani F, Ricerche CND (2002) Machine learning in automated text categorization. *ACM Comput Surv* 34(1):1–47
7. Tang J, Zhang D, Yao L (2007) Social network extraction of academic researchers. In: *Proceedings of ICDM'2007*, pp 292–301
8. Extraction and Mining of an Academic Social Network

Chapter 167

Analysis of Competition and Cooperation of Ningbo-Zhoushan Port and Shanghai Port

Peihua Fu and Yangfei Chen

Abstract Aimed at thermal phenomena for Port Competition of China's Yangtze River Delta, this chapter selecting Ningbo-Zhoushan Port and Shanghai Port to analyze how to make competition and cooperation between ports in the Yangtze River Delta. This chapter analyzed the overall competitiveness of Ningbo-Zhoushan Port and Shanghai Port respectively from four aspects, represented the relationships of competition and cooperation between the two ports, learn the three modes of cooperation from foreign experience in the ports and proposed the two ports should be join hands in the competition and should be guided by the Government to grasp the general direction and develop dislocatedly.

Keywords Ningbo-Zhoushan Port · Shanghai Port · Competition · Cooperation

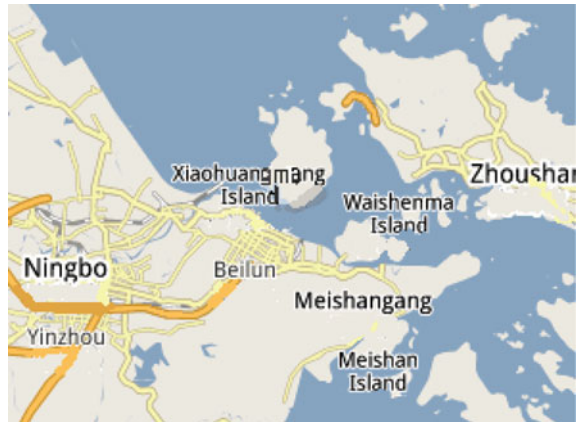
167.1 Introduction

With global economic integration, marine transport is still the main form of global cargo transport. The port is the meeting point of various modes of transport, so the port's role in promoting the regional economy can not be overstated. Like many industries, the port industry is also faced with many opportunities and challenges.

Fund project: Social Science Foundation of Zhejiang Province (10JDLG03YB).

P. Fu · Y. Chen (✉)
College of Computer and Information Engineering, Zhejiang Gongshang University,
Hangzhou, 310018, China
e-mail: faye.chen1988@163.com

Fig. 167.1 Ningbo-Zhoushan Port schematic



Such as continuous improvement of infrastructure has reduced the time and cost of regional transportation; overlapping scope of the port hinterland to further increase, the ports competed in the common hinterland; ships of large-scale required infrastructure facilities of ports to further increase and so on. For the problems faced by the port, foreign adopted a series of measures to ensure the coordinated development of the ports. Such as the New York–New Jersey Port Group in U.S. [1] set up Port Authority jointly, through the development of port economy and promote comprehensive development of the port group; European port group established a European Sea Ports Organization (ESPO) [2] to coordinate and manage the harbor of the entire European region; Tokyo Bay Port Group has a district port linkage [3], a clear division of labor, not only does not form a vicious competition, but to achieve a win–win situation and so on.

For the Yangtze River Delta, like Ningbo-Zhoushan Port and Shanghai Port [4], the study of the competition and cooperation between the two ports are all the more significant, and this is a hot and focus research at home and abroad now.

167.2 Comparison of Competitiveness of Ningbo-Zhoushan Port and Shanghai Port

167.2.1 Geographical Location Condition

Ningbo-Zhoushan Port which in Ningbo and Zhoushan provinces belongs to the city of Yangtze River Delta, shown in Fig. 167.1. Ningbo-Zhoushan Port's coastline is rich in resources. As China's economic hot spots, highlight the advantages of its location, strong industrial base, high dependence on outside world. Ningbo-Zhoushan Port is one of China's coastal hub port and an important hub of an integrated transport system. Ningbo-Zhoushan Port through rivers and

Fig. 167.2 Shanghai Port schematic



seas, along the Yangtze River can be traced back to Jiangsu, Anhui, Jiangxi, Hunan, Hubei, Chongqing, Sichuan and other provinces, can reach Busan, South Korea, Japan, Kobe, Singapore and other ports through the coastal, with the geographical advantage that domestic and international shipping Hub should have.

Total area of Shanghai Port which is the largest port in China is 3619.6 square kilometers [5]. It is located in the intersection of east–west transport routes of the Yangtze River and north–south transport corridor of sea, shown in Fig. 167.2. It through the north and south coasts of China and oceans of the world, and consistent the Yangtze River and Jiangsu, Zhejiang, Anhui River, Taihu Lake basin. It is connected by the Shanghai-Hangzhou and Shanghai-Nanjing trunk lines to national railway, and leading to Yantai, Urumqi, Kunming and Lhasa. It has favorable natural conditions. Its hinterlands have a developed economy and has open channels for set sparse.

The contrast of Ningbo-Zhoushan Port and Shanghai Port's basic situation is in Table 167.1.

167.2.2 Supply Condition of the Hinterland

Ningbo-Zhoushan port's economic hinterland includes seven provinces and two cities (Shanghai, Jiangsu, Zhejiang, Anhui, Jiangxi, Hunan, Hubei, Sichuan, Chongqing) which compose of the Yangtze River economic belt, and Ningbo-Zhoushan Port's the most direct economic hinterland is Ningbo and its surrounding area in Zhejiang Province. The completion of the Hangzhou Bay Bridge have promoted the Ningbo-Zhoushan Port's regional industrial structure adjustment and industrial training community, have made Ningbo and the surrounding areas accept better radiation of Shanghai, and is conducive to using foreign investment and introducing technology to promote the development of Ningbo's

Table 167.1 The contrast of Ningbo-Zhoushan Port and Shanghai Port

Port	Geographic coordinates	Port type	Carrier range	Harbor distribution
Ningbo-Zhoushan Port	29°56'N	River port	Crude oil, ore, coal, container, liquid chemical products, etc	Port Yongjiang, Zhenhai, Beilun, Chuanshan, Daxie of 19
	121°50'E	Seaport		
Shanghai Port	121°29'04.5"E	River port	Container, coal, metal ores, petroleum and its products, steel, mineral building materials, machinery and equipment, etc	Baoshan, Zhanghuabang, military Road, Waigaoqiao, Gongqing, Gao Yang, Zhujiamen, livelihood, Xinhua, revival, Kaiping, Dongchang and other port
	31°14'18.8"N	Seaport		

processing industry, manufacturing and other foreign industries. Ningbo-Zhoushan Port's hinterland but also extended from eastern Zhejiang to southern and northern Jiangsu, Anhui, and then extended to the entire Yangtze River basin region.

Shanghai, as the Chinese financial, economic and maritime center and the leading of economic development in the Yangtze River Delta region, has a good location, personnel and liberal investment environment. Shanghai Port uses Shanghai as the basis, backed by the Yangtze River, have a vast economic hinterland, 31 provinces (including Taiwan) have the goods handling or dress re-exports through Shanghai port. The major economic hinterland to Shanghai Port are as the same as Ningbo-Zhoushan Port. And Shanghai's land and water transport facilities is convenient, collection and distribution channels is open. Through the highway and national highway, railway lines and coastal transport network can be radiated to the Yangtze River or even the whole nation. Close to the world shipping routes, and locate at the edge of the world's sea routes. Shanghai also has developed air transport. Excellent location and convenient collection and distribution transport, which provides Shanghai a good supply conditions.

167.2.3 Hardware Condition

The area of Ningbo-Zhoushan Port has the best and richest resources. It is the largest port of large deep-water berths in inland. In the construction of Beilun Phase IV and Phase V international container terminal water area and Jintang Island container terminal development project, more large container ships can berth. The Shanghai Port was in lack of hardware facilities when comparing with

Table 167.2 The hardware comparison of Ningbo-Zhoushan Port and Shanghai Port

Port	Sea shoreline (km)	Production berths (seat)	Throughput capacity (10,000 ton)
Ningbo-Zhoushan Port	4,750	723	20,000
Shanghai Port	589	1,140	17,051

Table 167.3 Ningbo-Zhoushan Port and Shanghai Port throughput comparison

Port	Cargo throughput (million tons)	Increase the amount of cargo throughput (%)	Container throughput (myria-TEU)	Increase the amount of container throughput (%)
Ningbo- Zhoushan Port	6.33	9.7	1314.7	25.2
Shanghai Port	6.53	10.4	2906.9	16.3

the Ningbo-Zhoushan Port, but it has more berths. Detailed comparison is in Table 167.2.

167.2.4 Overall Development Level

The level of the overall development of the port here mainly shows by cargo throughput and container throughput. So these two figures are chosen to show the overall development level of Ningbo-Zhoushan Port and Shanghai Port. The compare in 2010 and the increase amount is in Table 167.3.

The growth and the growth rate of cargo throughput and container throughput of the two port in 2010 were a record high. Shanghai Port's container throughput for the first time surpass the Singapore Port and became NO.1 in the world. The Ningbo-Zhoushan Port has surpassed the Dubai Port and Guangzhou Port and became the world's sixth.

After several years of development, Ningbo-Zhoushan Port has formed a trunk and four bases, namely, the container ocean main ports, China's largest ore transit base, the largest crude oil transit base, China's largest liquid chemical storage and transportation along the coast and an important base for coal transportation in the East China region. It is an important part and the deep water outer harbour of Shanghai International Shipping Center, and is the fastest growing integrated port.

Shanghai Port is the hub of north-south sea and Sea Combined Transport. Although the world number one, but dispersed Shanghai Port, port and railway line separated, the development of sea and railway transport is restricted. Also, because the deviation from the main international shipping lines, are rarely the first arrival

and late departure of the Spanish-American and European routes, international transit obvious disadvantage. Meanwhile, a number of the neighboring international port's construction and expansion and business alliances of shipping companies have brought Shanghai's development tremendous pressure.

167.3 Analysis of the Two Port's Competition and Cooperation Modes by Studying the Cooperation Mode of Foreign Ports

Ningbo-Zhoushan Port and Shanghai Port are in the most economically developed Yangtze River Delta region. Recent years, two ports' trade increased rapidly, they have become the bridgehead of economic development and external communication in Yangtze River Delta regional. However, in the two ports' development process, function of the port is similar and the economic hinterlands are the same, some competition is essential, and there are many similar cases in foreign countries. Analyze Ningbo-Zhoushan Port and Shanghai Port's competition and cooperation relationship by studying the cooperation model of famous foreign ports, recognizing the urgent need to establish two-port's healthy competition and cooperation relations.

167.3.1 Comparison Mode of Cooperation of Foreign Ports

Mode of cooperation of foreign ports to guide the reform of our port operation mode makes sense. This chapter elects three well-known cooperation modes of foreign ports to analysis. See Table 167.4.

From the comparison we can see the three cooperation modes apply to the Yangtze River Delta region directly are defective. The first mode has not yet been accepted by the local government; the second does not take the port play a great role in the local economy in China into account, the competition between the port city is long-term existence; the third is not suitable for the development degree of market economy is not high enough and the legal system sophistication is not high enough.

167.3.2 Port Competition

The great development of the port also breeds the seeds of competition. In recent years, the rapid development of the ports in Yangtze River Delta, port pattern developed from the thriving port of Shanghai to the Shanghai Port and Ningbo-Zhoushan Port "dual hub" mode. Shanghai Port directes at a world-class port and Ningbo-Zhoushan Port strives to enter the top three ports in the world.

Table 167.4 Ningbo-Zhoushan Port and Shanghai Port throughput comparison

Mode	Port	Characteristic	Scope of application
Place led	New York– New Jersey Port	Joint establishment of regulatory bodies to achieve unified planning, development and management	Smaller regions, especially within the port location is very close to the case of a limited number of ports is more suitable
State led	Tokyo Bay	Led by the State, the Ministry of Transport is responsible for the coordination of the various ports	The idea of government about the port city are the same, areas of high authority about ports planning
Association led	European Seaports	To maintain the independence of the ports, form and protect the reasonable playing field between the ports	In the highly developed market economy, the legal system is more comprehensive

They are separated by a narrow Hangzhou Bay, shared waters, sharing of international routes, so many resources in the hinterland have staggered even overlapping, the hinterland of the port's power is directly related to supply volume, so competition is inevitable.

Shanghai Port layout has actions as early as 8 years ago, the first is cooperating with Wuhan Port, shares of 40% of the establishment of Wuhan Port Container Co. Ltd. and then Shanghai alliance Wuhu, Chongqing, Nanjing, Nantong Port. In 2008, the Shanghai International Port Group Co. Ltd.'s invested in Yibin and Chongqing Port's container terminal. With the cooperation of different levels along the Yangtze River port, Shanghai has basically completed "Yangtze strategy" layout. In essence, it is mainly to expand the port hinterland and to attract more goods.

Ningbo Port is also frequently go north to layout the "golden waterway" along Yangtze River and with Shanghai's "sphere of influence" close combat. Last March, Taicangwu Port and Taicang Articles International Terminal invested by Ningbo Port Group have been officially put into operation. Meanwhile Ningbo Port have also set up a business berths in Nanjing and Suzhou Port which in the Yangtze River. In addition, Ningbo to Nantong, Zhangjiagang or Taicang container routes are starting operations. In response to the construction of Shanghai Yangshan Deepwater Port [6–7], Zhejiang established Hangzhou Bay Bridge and implement the Hangzhou-Ningbo Canal Improvement Project, which aims to increase Ningbo-Zhoushan Port's radiation.

167.3.3 Port Cooperation

Strengthening the cooperation between ports, especially a cluster cooperation, is to prevent ports' excessive competition on development and operation terms, thus reasonable use harbor resource is an effective means and an inevitable choice.

167.3.3.1 Dislocation Development

Port dislocation is to give port group within different port to a different port functional, positioning and division, make port group developed into a rational layout, complete functions, a clear division of work, complementary advantages and overall coordination of the organic whole. Its purpose is to foster the port's own advantage. Ningbo-Zhoushan port and Shanghai port must unite ideas, set up the whole ideas and concepts. Especially in making their development planning and policy, to foster strengths and circumvent weaknesses, outstanding advantages and grasp its own accurate positioning.

Shanghai international shipping center, localization should focus on the development of shipping high-end service and Shanghai is big cities, do not have let large shipments in the carriage through downtown and the conditions to transport to the suburbs, or the city will be more congestion. Such as Shanghai and suburban enterprise's demand for coal, ore, food, petrochemical and wood, etc., has to be transported by Ningbo-Zhoushan Port. Ningbo-Zhoushan Port is located for the Chinese coastal and the Yangtze river delta region's important comprehensive transportation hub, international container shipping lines and bulk reserving base to the Yangtze River basin, should focus on the development of deep-water wharf. Anyhow, two ports should be based on their own conditions for positive science division of labour, formed division of distinct feeder port, main port, or inland fed port, and rational layout of regional container port network.

167.3.3.2 Policy Support

Due to the administrative divisions, the government requires to be guided, to grasp the general direction and form a long-acting connective mechanism between Ningbo-Zhoushan Port and Shanghai Port to make sure its firm of cooperation. The government should strengthen the overall planning and the implementary efforts of the development of the Ningbo-Zhoushan Port and Shanghai Port. Formulate relevant policies and regulations, set up effective administrative coordination mechanism to promote two ports' division of labor and cooperation, ensure the development of two port dislocating smoothly. In recent years, the government has taken some measures to ensure the Yangtze river delta region ports, major events in Table 167.5.

167.4 Summary

Competition and cooperation are the two most important market economy of economic characteristics. Under the environment of the global economic, trade, logistics development, Ningbo-Zhoushan Port and Shanghai Port depending on the advantages, such as interior economy, geographic location and port conditions, and

Table 167.5 Events to promote port cooperation in Yangtze River Delta region

Time	Sponsors	Events
1997	The Ministry of Communications and two provinces and one city in the Yangtze River Delta	Establishment of the Shanghai Combination Port Management Committee and the Office
2006	16 cities' Port and Shipping Management in the Yangtze River Delta	Joint meeting of the Yangtze River Delta port management sector
March, 2009	The State Council	The views of Shanghai should speed up the development of building an international financial center and an international shipping center

learn from practical experience in foreign ports to avoid “homogeneous competition” and take the road of cooperation and competition. Work together to build Shanghai International Shipping Center, will surely exceed and improve the port function, increase port integrated development.

References

1. Jiang L (2010) Analyzed and countermeasures discussion about logistics development in Ningbo-Zhoushan Port [J]. *Economist* 3:255–257
2. Zhang X (2009) Port competitiveness and cooperative game analysis—take Shanghai Port and Ningbo Port as examples [J]. *Logist Eng Manag* 31(9):13–15
3. Zhang S, Lu Z (2010) The game analysis of Shanghai and Ningbo-Zhoushan Port's competition and cooperation [J]. *Port Waterw Eng* 5:79–82
4. Lu HS (2010) Competition and cooperation path analysis about Shanghai and Ningbo Ports [J]. *China Collect Econ* 4:55–56
5. Wang L, Mao B (2010) Different modes of cooperation in foreign port system and inspiration for the Yangtze River Delta [J]. *Soc Sci* 6:37–44
6. Wang X (2008) The competition and cooperation between Ningbo-Zhoushan Port and Shanghai Port [J]. *Port Waterw Manag* 30(4):9–12
7. Wang B (2005) Some discussions of competition and cooperation about Chinese coastal port [J]. *Port Econ* 4:20–21

Chapter 168

The Study About Long Memory and Volatility Persistence in China Stock Market Based on Fractal Theory and GARCH Model

Liu Cheng, Liu Jiankang and He Guozhu

Abstract The research to volatility characteristics in financial market is foundation to the problems of capital assets pricing and avoiding strategy of financial risk. This chapter gives the empirical study about the long memory and volatility persistence in China stock market, using fractal spectral density estimation for time series and GARCH model, and gets some conclusions.

Keywords Fractal theory · Volatility persistence · GARCH model · Long memory characteristic

168.1 Introduction

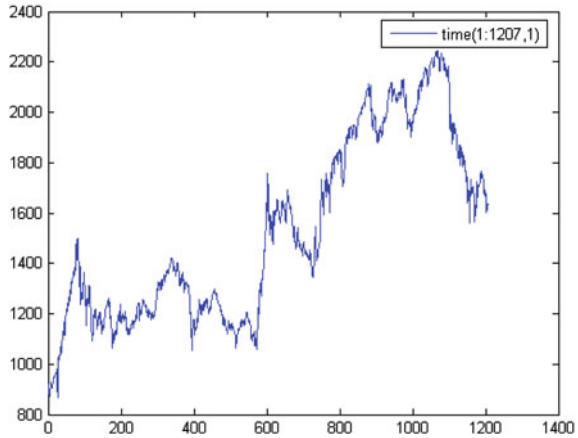
China's stock market has been over 20 years of development and improvement, from the beginning of China's reform and opening up, but, stock market, as a barometer of the market economy, still has many problems underdeveloped and immature, In not yet fully functioning market economy. Long memory is very obvious characteristics of the market, for the mature stock markets. The existence of long Memory in China's Stock Market has become the main contents of many scholars today. This article will use method of the time series fractal dimension

L. Cheng (✉) · L. Jiankang · H. Guozhu
College of Business, Sichun Agricultural University, Dujiangyan 611830, China
e-mail: matlabcheng@163.com

L. Jiankang
e-mail: ljiankang@scfc.edu

H. Guozhu
e-mail: hguozhu@scfc.edu

Fig. 168.1 Data for long memory characteristic



estimation, and research long memory of the Chinese stock market by verifying the fractal dimension characteristics of stock data. And on this basis, research the existence of volatility persistence in China's Stock Market further, using ARCH modeling method.

168.2 The Study About Long Memory of the Chinese Stock Market

In the Chinese stock market, government macro-control on the stock market play a decisive role, which does not appear under the conditions of mature market economy. To research the true nature of the Chinese stock market this article select Shanghai composite Index closing data from January 2, 1997 to December 31, 2001 adding up to 1207 (Fig. 168.1), because of Chinese stock market has not been significantly affected in the regulation of government macro-control, which can reflect the real situation of Chinese market.

Analysis is as follows:

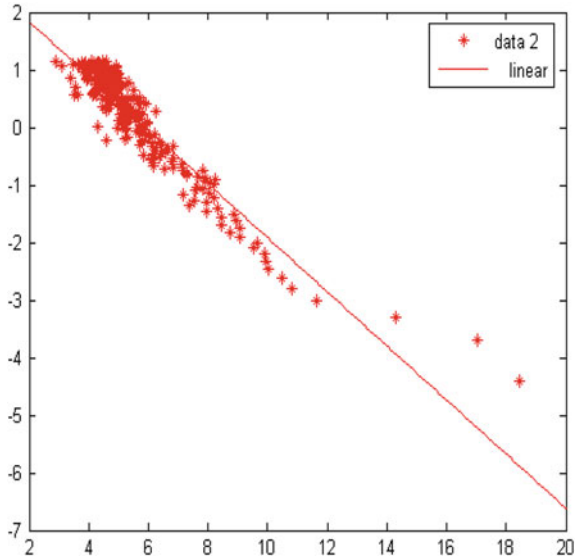
- (1) Calculate the Spectral density function [1] of time series x_t ($t = 0, 1, \dots, 1207$), let

$$I(\omega) = \frac{1}{2\pi T} \left| \sum_{t=1}^T x_t e^{it\omega} \right|, \quad (0 < \omega < \pi) \quad (168.1)$$

when $I(\omega)$ is the asymptotically unbiased estimates for spectral density $f(\omega)$ of x_t .

- (2) Combine Eq. 168.1, according to the related fractal nature of time series [2], we get

Fig. 168.2 Regression line



$$f(\omega) = \left(2 \sin \frac{\omega}{2}\right)^{-2H}, \quad 0 < \omega \leq \pi \tag{168.2}$$

$$f(\omega) \sim \omega^{-2H}, \quad \omega \rightarrow 0 \tag{168.3}$$

which, H is Hurst index.

- (3) Have logarithmic of Eq. 168.3, get $\ln I(\omega) \sim \ln(\omega^{-2H})$, then regress $\ln I(\omega)$ to $\ln(\omega)$ (regression results shown in Fig. 168.2), get the the slope of the regression line is $1-2H$, to obtain estimates of H.
- (4) Hurst index is a important parameters to descript fractal time series [2], and

$$H = 0.5 + d \tag{168.4}$$

which, d is a important parameters to descript the long memory of time series in ARFIMA.

- (5) According to the above analysis, obtain the regression equation

$$y = -0.46969x + 2.7805$$

and use (168.4) we can get $d = 0.2348$. According to ARFIMA (p, d, q), when $0 < d < 0.5$, this time series has obvious characteristics of long memory.

- (6) After the above empirical analysis, Shanghai stock market has a very distinct characteristics of long memory in empirical data collection time period. This chapter argues that, different conclusions may be due to different empirical data.

Fig. 168.3 Data for volatility persistence

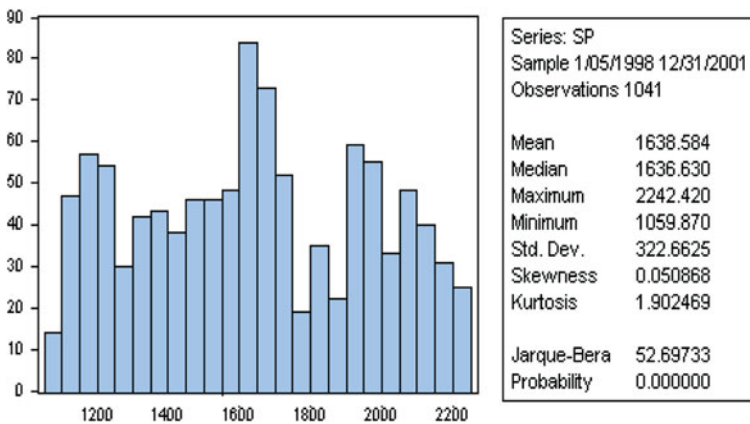
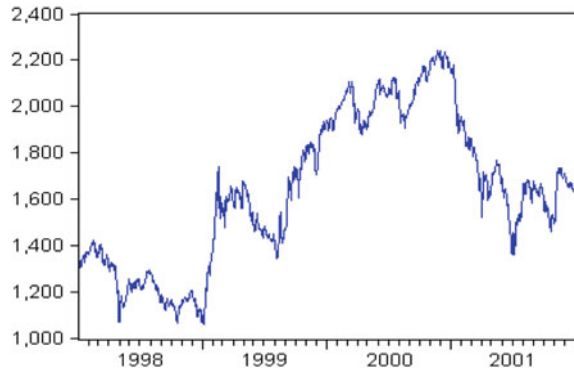


Fig. 168.4 Statistical distribution

168.3 Empirical Study about Volatility persistence in China's Stock Market

On the basis of the study of long memory of volatility, we further analysis the existence of the volatility persistence. This chapter select Shanghai Composite Index closing data, from January 5, 1998 to December 31, 2001, adding up to 940, as the sample sequence $\{sp_n\}$ (Fig. 168.3).

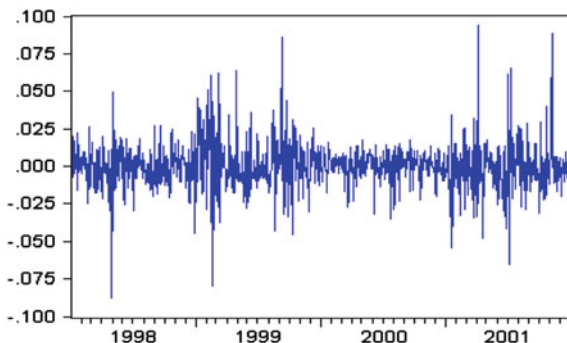
Analysis is as follows:

- (1) First, calculate the basic statistical properties of the sample sequence (Fig. 168.4). Easy to see from Fig. 168.4, that sample sequence does not meet the standard normal distribution, and show obvious peaks heavy tail, which is the typical characteristics of the financial volatility.

Table 168.1 Regression equation statistical test

Variable	Coefficient	Std. Error	t-Statistic	Prob.
LOG (SP(-1))	1.000028	6.44E-05	15516.95	0.0000
R-squared	0.994215	Mean dependent var		7.381604
Adjusted R-squared	0.994215	S.D. dependent var		0.201867
S.E. of regression	0.015354	Akaike info criterion		-5.513878
Sum squared resid	0.245186	Schwarz criterion		-5.509125
Log likelihood	2870.974	Durbin-Watson slat		1.935631

Fig. 168.5 Regression residuals



- (2) Do natural logarithm transformation on the sequence. Taking into account stocks index series use commonly Random Walk model [3], we selection basic model as following

$$\ln(sp_t) = \gamma \times \ln(sp_{t-1}) + u_t \tag{168.5}$$

which, u_t is regression residual, γ is regression coefficient.

- (3) Using the least square method, getting the estimate of the regression equation as following

$$\ln(\hat{sp}_t) = \underset{(15517)}{1.000028} \times \ln(sp_{t-1}) \tag{168.6}$$

and $R^2 = 0.994$, log-likelihood value is 2871, $AIC = -5.51$, $SC = -5.51$.

- (4) Statistical test: do statistical test on the above results (Table 168.1).The results show that regression equation is statistically significant characteristics, and the degree of fitting is good.
- (5) Autoregressive conditional heteroskedasticity (ARCH Effect) test [4, 5]. Make the residual plot of the regression equation (Fig. 168.5). It can be seen from Fig. 168.5 that the residual sequence has the characteristic of volatility clustering, which can preliminary determine that order ARCH effects exist in residual series. Do ARCH-LM test of conditional heteroskedasticity for regression equation further [6-8], and get the test results when lag order $p = 3$

Table 168.2 Test results of ARCH-LM

F-statistic	19.86353	Probability	0.000000
Qbs*R-squared	56.57756	Probability	0.000000

(Table 168.2). From the test results, we can see that Residual series exists ARCH effect.

- (6) Re-estimate Eq. 168.5, using GARCH model, results are as follows mean equation:

$$\ln(sp_t) = 1.00003 \times \ln(sp_{t-1})$$

(23249)

Variance equation:

$$\hat{\sigma}_t^2 = 1.19 \times 10^{-5} + 0.251 \times \hat{u}_{t-1}^2 + 0.731 \times \hat{\sigma}_{t-1}^2$$

(5.27) (11.49) (33.38)

R2 = 0.994, D.W. = 1.94, log-likelihood value is 3003, AIC = -5.76, SC = -5.74.

- (7) The results show that: the coefficients of ARCH items and GARCH items are significant, and the log-likelihood value increased, AIC and SC became smaller, which illustrate this model can fit the data better. Then do ARCH-LM test for this equation, getting dependent probability is P = 0.91, which illustrate GARCH model eliminated ARCH effects in the original regression equation.

Acknowledgment This work was supported by Double support plan (school-level special) of Sichuan Agricultural University, 2011.

References

1. Edgar E (1991) Chaos and order in the capital markets. Wiley, New York
2. Fan Z, Zhang S (2002) The efficiency of financial markets and fractal market theory. Syst Eng Theo Pract 22(3):13–19
3. Hosking M (1981) Fractional differencing. Biometrika 68:165–176
4. Engle F (1982) Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. Econometrica 50:987–1008
5. Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. J Ecometrics 31:307–327
6. Li H, Zhang S, Fan Z (2002) Continuity and avoidance strategy of financial risks. Syst Eng Theo Pract 22(5):31–36
7. Geweke J, Porter-Hudak S (1983) The estimation and application of long memory time series models. J Time Series Anal 4:221–237
8. Sarkar S (2003) The effect of mean reversion on investment under uncertainty. J Econ Dyn Control 28:377–396

Chapter 169

Research on Home-Textile Enterprise-Oriented Comprehensive Management Integration Platform and Its Application

Penghui Zhan, Zailin Lu, Renwang Li, Zhigang Bao, Qian Yin and Yongxian Chen

Abstract Aiming at the practical situation of house-textile industry in China and combining the current situation of informatization and the trend of development of domestic enterprises, this article mainly elucidates the production and management platform for comprehensive integration of theory of the small and medium-sized household textile enterprises, analyses the characteristics of the production management, the article focuses on probing the design issues of small-scale household textile's production and operations management of integrated design platform, including production and management of an enterprise architecture system, and finally describes the system platform development and its application.

Keywords Home-textile enterprises · ERP · Comprehensive integration platform

P. Zhan (✉) · R. Li · Z. Bao · Q. Yin · Y. Chen
Institute of Advanced Manufacturer Technology, Zhejiang Sci-Tech University,
Hangzhou 310018, China
e-mail: zhanpenghui@126.com

R. Li
e-mail: renwangli@foxmail.com

Z. Bao
e-mail: BaoZhigang@126.com

Q. Yin
e-mail: YinQian@126.com

Y. Chen
e-mail: ChenYongxian@126.com

Z. Lu
Department of Applied Engineering, Hangzhou Wanxiang Polytechnic, Hangzhou
310023, China
e-mail: zailinzju@sohu.com

169.1 Introduction

Home-textile industry is an important part of China's textile industry. Over the past ten years, the house-textiles industry in China has developed rapidly by the annual growth rate of more than 20%, in 2009 the national household textile industry output value has reached 960 billion Yuan and the total market has rose to 500 billion Yuan. Since 2010, China's house-textile market will maintain rapid growth in the future [1]. Generally speaking, Home-textile production management mainly includes the production process, labor organization, work quota, material, equipment, energy, quality, safety and environmental protection, etc. [2].

Home-textile production and management of household objects are rational organization of the productive forces, how to produce effective integration of various elements together to form an organic whole, under the constraints of internal and external conditions in the enterprise, with minimal input for maximum output. Therefore, when textile production management, it can be divided into planning, doing, checking and action in four phases, to affirm the experience of success, so that standardization, and formulate plans according to standards of production directly [2]. Failure to draw lessons and prevent a similar situation, while the content once again fed into the next plan, the formation of PDCA cycle management [3].

Home-textile companies and other enterprises both in the production of common management, but also personality, in the home-textile products failed quality testing when finished or semi-finished products, subject to change film, apart redo can become qualified; In the home-textile production process, the sewing machine sewing bedding films, the working hours of sewing time, only 40% of the entire work is about to take the rest of the time, lace, piece goods on the bed, change lines, cut lines, recording, processing, and other matters contact the time spent. Currently, management of textile production is particularly important, China's home-textile production management foundation is weak, most factories are their experience and subjective will arrange the production line. In order to meet the needs of the market economy, in the production process to conduct scientific management, efficient use of manpower, material, equipment and funds, the use of reasonable methods in order to produce affordable, quality home-textile products.

According to the characteristics of the house-textile industry, house-textile enterprises should follow the "unified planning, unified standards, unified platform, unified standard" principle and the "overall planning, step by step, focused, application-first" approach, according to their different characteristics and the actual situation, Implementation of the application of ERP, SCM, CRM and CAD/CAM and other information technology systems. When the ERP system can meet the "standardized, industry-based, platform" [4] of the three major characteristics, the only truly meet the long-term development of house-textile enterprises.

169.2 System Architecture Design for Comprehensive Integration Platform

Overview of comprehensive integrated platform function. Through establishing comprehensive management integrated platform for home-textile enterprises, all business processes such as production process, materials transportation, transaction processing, cash flowing, customer communication, etc., were digitized and translated into new information resources by all sorts of information systems, and were presented to employers and employees, in order to optimize factors of production and allocation of resources, which could help enterprise to survive in the rapidly diversifying and increasingly competitive market. And based-on business reengineering, process optimizing, and department flat restructuring, effective Check-and-Balance Mechanism is established and significant economical and social benefits would be achieved.

Establishing comprehensive management integrated platform home-textile enterprises could help enterprise to accomplish three types of function as follows: (1) Implementing lean management of dynamic cost calculation based-on product varieties and types. (2) Developing supply chain management system, including product planning management, sales management, manufacturing management etc., which could complete evaluating and choosing partners, establish suppliers evaluating base and SCM partners relationship base, and development safety and quality approving management for suppliers. (3) Establishing collaborative commerce system among enterprises by using collaborative commerce technology and enterprise application integration technology, including the following functions: on-lined ordering, real-timely tracking of orders, product safety and quality approving, and feedback of quality information. Based-on lean management of cost, supply chain management and collaborative commerce platform, constructing comprehensive integrated platform of production and operation management for home-textile enterprises could achieve the integration of information flow, materials flow, cash flow and production and operation management.

Composition for the comprehensive integration platform. According to the characters of home-textile industry and the requirements of a home-textile company in Xiaoshan, Zhejiang, the architecture of above comprehensive integration platform is designed. Based-on the original ERP system, related function modules are added, mainly including lean management system of product cost, SCM system and collaborative commercial platform among enterprises.

The aim of lean management system is to realize lean management of product cost. The main design method is adding appropriate functions to achieve lean management of cost to the original ERP system by secondary development. According to the theory of supply chain management, the proposed product supply chain management system is designed, which has close relationships with suppliers of the chain based on one certain product. In the system data interchange interface for partners and suppliers, and information communication function among enterprise, partners and suppliers are provided. In order to facilitate the

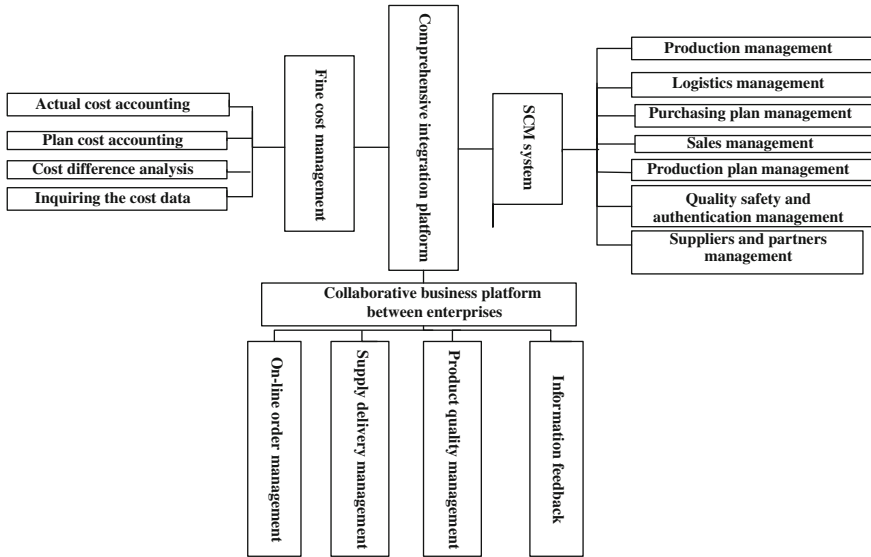


Fig. 169.1 The function tree for production and management Integration platform

communication with suppliers and partners, B/S structure is needed. This enterprise enjoys not only C/S system. Thus, considering effective integration of enterprise original system and new developing system, the architecture combining C/S and B/S is chosen finally. The aim of collaborative commerce platform is enhance the communications between this home-textile enterprise and customers. According to the theory of collaborative commerce and customer relationship management related function modules are divided and designed.

According to business demands of the home-textile company, functions of related projects in production and operation management integration are decomposed. Based-on the requirement of independence among functions, the system is divided into several separated sub-systems, and each sub-system is divided into several modules. During the decomposition the function of each module. The comprehensive management integrated platform in this home-textile company includes three sub-systems: lean management system of product cost, product supply chain management system and collaborative commerce platform, which function tree is as Fig. 169.1.

Technical architecture of the comprehensive integration platform. The home textile company developed and run different heterogeneous application systems in different period, such as ERP system, OA system, and the new developed supply chain management system and collaborative commerce system, which are distributed in related department of this company [5]. How to integrate these systems to make them to meet new function to make them adapt to the new requirements of the workflow, becomes one of key technological questions we

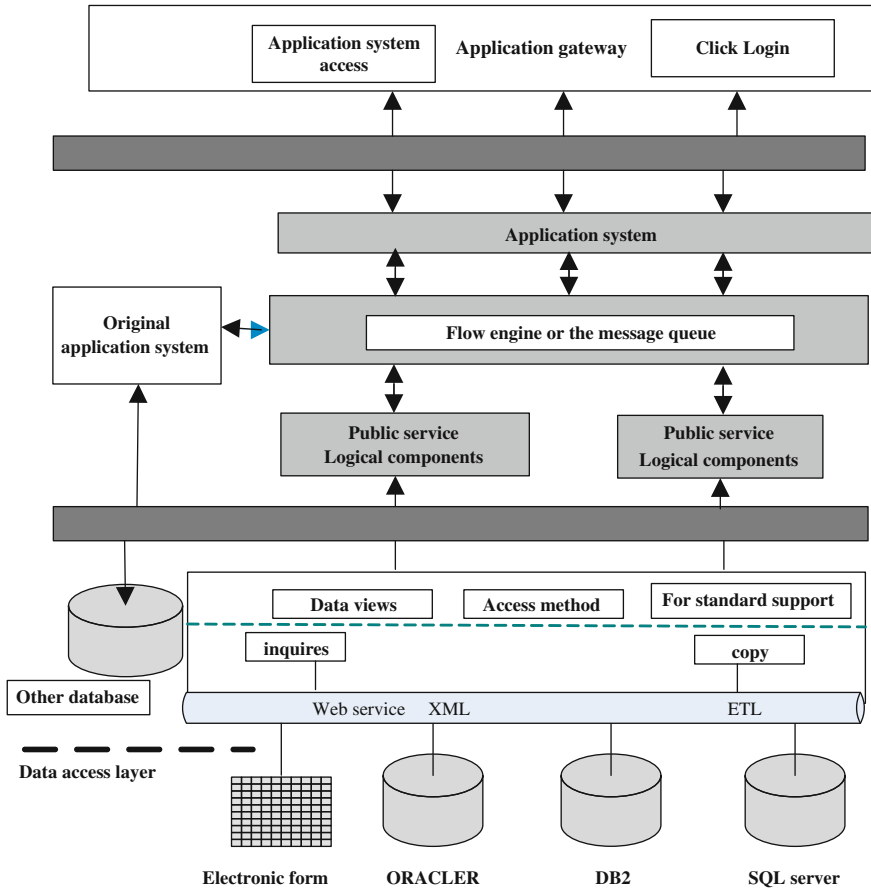


Fig. 169.2 The overall technical architecture of the integration platform

need to solve. After investigation and data analysis, we believe that enterprise application integration technologies could help home-textile enterprises to establish comprehensive management integration platform.

In order to enter a certain systematic interface of another one from a system, need to go through a series of keyboard input, including passwords, menu selection. Therefore, simulates the technology which and the tool the keyboard input and the screen capture in the contact surface integration pattern uses frequently. As for enterprise’s management integration platform for the home-textile, we use the interface layer and data layer integrates technology. Its technical architecture is as shown in Fig. 169.2

What is located at the first layer is the data integration, which is associated with the interaction between heterogeneous database access, which is to use data replication for heterogeneous data access among, whereas the top is the portal

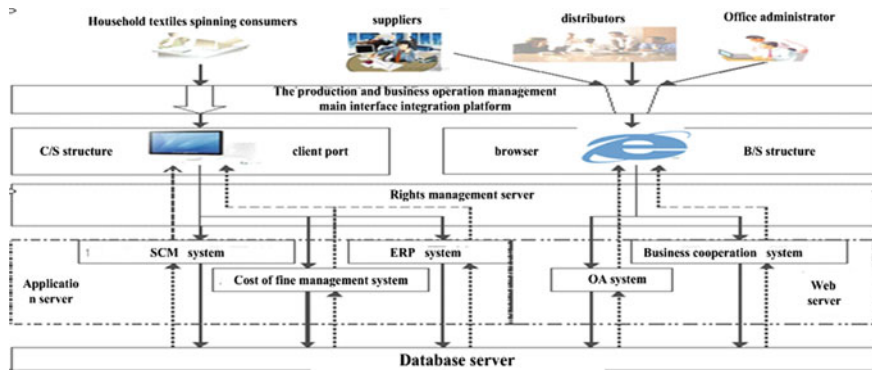


Fig. 169.3 The overall application system structure of the integration platform

integration, it is related to the construction and application portal, the main to achieve authentication and single sign-on, and all access applications.

169.3 The Development and Application of Integrated Platform

The Development of Integrated Platform. The production and business operation management Platform for small and medium home-textile enterprises integrated platform C/S (Client/Server), B/S mixed structure, as shown in Fig. 169.3. Some enterprise users use client-side applications, through the deployment of the application server’s connect to the database server and client program to the server sends request in accordance with the request of the user, the client program sends a request to the server, server in accordance with user requests, the associated database operations, and results of the database returned to the client application. Suppliers, distributors, direct store around using the IE browser, access to collaborative commerce system and OA system, the WEB server browser is responsible for the interpretation of the user requests from execution, related database operation, and the result of database operation to web returned to the user.

The Application of Comprehensive Management Integration Platform. For home-textile enterprise, established integrated platform on the enterprise information construction have play an important role. After an enterprise had adopted the comprehensive integration platform, it has broken information island phenomenon between the production department, for enterprise high-level strategic decision to bring convenience, for enterprise’s production management, home-textiles sales to bring the direct effect, making the development of enterprise specialization, standardization, platform digestion. Figure 169.4 for some home textile enterprise integrated with the Comprehensive management platform.



Fig. 169.4 Management integration platform interface of some home-textile enterprise

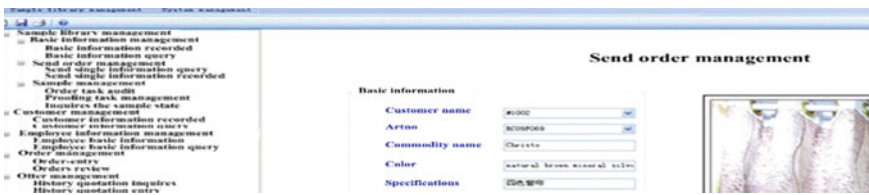


Fig. 169.5 Sending single information recorded



Fig. 169.6 Sending single information query

In the production and business operation management, including the workbench integrated supply chain management is the main component of the platform, and the comprehensive integration platform each system function module most parts, in the process of the system implementation has carried on many times the improvement. Figure 169.5 for sending single information recorded interface diagram.

In an enterprise production management process, home-textile enterprise general managers generally reflect this enterprise products production situation is difficult to understand, thereby causing directly to enterprise information lag of enterprise, to the detriment of more effective and more direct management and decision-making. So, while in improving SCM system in the process, we should

pay special attention to production, product from the order of dynamic inquires again to sales. Figure 169.6 for sending single information query interface diagram. To the process of solving this problem, is us who see more informatization bringing to the enterprise bigger opportunity for enterprise received good economic and social benefits.

169.4 Summary

The production and business operation management platform for the home-textile enterprise, making a fine management system, supply chain management systems, and collaborative business platform between enterprises information integration and sharing of resources, from the initial facing client system development, to establish the business cooperation based on Web server platform.

The integrated platform specially is applied to the current home-textile industry cluster development of the small medium-sized home-textile enterprise strategic positioning new request. And competitive advantage of industrial cluster scale returns, scope economy, regional economic development and competition play a more and more important role, so the study of home-textile enterprise for management comprehensive integration platform has important economic and social benefits.

Acknowledgments The work is supported by the Natural Science Foundation (Grant No. R6080403, No. Z6090572) and Key Science and Technology Creating Team Project (Grant No. 2011R09015) of Zhejiang Province, and the Scientific Research Fund of Zhejiang Provincial Education Department (Y200909779).

References

1. Xu XY(2010) Depth research about Fuanna company. QiLu Securities Institute, Shanghai (in press)
2. Zhu YZ (2009) Household textile production management and cost accounting. China Textile Press, BJ (in press)
3. Wang Y (2009) House-textile marketing. China Textile Press, BJ (in press)
4. Information on <http://info.textile.hc360.com/2010/01/221320148586-2.shtml>
5. Gao P, Xie YC (2009) China management informationization, vol 12, p 57

Chapter 170

Location-Aware Elderly Personal Safety Checking in Smart Home

Bin Xu, Zepeng Hu, Peiyu Han and Yujia Ge

Abstract Due to the increasing ratio of elderly, elderly personal safety has become a social problem in many countries. As a result, smart home has been paid much more attention globally. This paper intends to provide an economic solution to monitor the personal safety of the elderly in smart home according to the location movement status, and then check the safety with a set of light living pattern rules. It is assumed to be unsafe if the elderly remains at a space beyond the period defined in the rules. The family members, neighbor or doctor will be informed consequently if the elderly is recognized to be unsafe by the system. A prototype has been developed and the approaches were approved to be sufficient to check the scenarios. The pattern rules with trial duration were ignored and all the rules were customizable and could be trained all the time.

Keywords Location-aware · Elderly safety checking · Intelligent environment

B. Xu (✉) · Z. Hu · P. Han · Y. Ge
College of Computer and Information Engineering, Zhejiang Gongshang University,
Hangzhou, Zhejiang, China
e-mail: xubin@zjgsu.edu.cn

Z. Hu
e-mail: zepeng_china@163.com

P. Han
e-mail: happyyou@zjgsu.edu.cn

Y. Ge
e-mail: yjge@zjgsu.edu.cn

170.1 Introduction

There are several countries have entered the era of aging population, including America, Korea, Japan and China [1]. By the end of 2009, the national aging population of China has increased to 1.6714 billion, which is 12.5% in total population. Compared with the year 2009, elderly population increased 725 million, with the proportion increased by 0.5%, most elderly suffering from chronic illnesses such as hypertension and diabetes in 2010. Therefore, health care for the elderly has been recognized as a social problem [2]. Smart home turns out to be an efficient way to enable the elderly life happy with less support [1, 3–7] while a safety enhancing mechanisms is proposed [3], web service is used to manage the medication [4], human intention change is detected and analyzed [1, 5], and integrated mobile solution for personal health care is provided [6, 7].

Considering that the sudden disease, such as cardiovascular disease, cerebral thrombosis, and stroke, has an upward trend in the elderly in recent years, some experts argued that fast rescue and timely treatment should be provided to the sudden illness to enable the elderly survive from the emergency easily [3, 4, 6]. Though there are some systems enable the elderly to call for help when they are in trouble or emergency, these systems are not sufficient when the elderly lose consciousness [6, 7]. To detect the emergency as early as possible is essential in such situation, which is the main purpose of our research.

In this research, a smart home is divided into several rooms and each room can be further divided into several spaces. It is assumed to be unsafe if the elderly remains at a space beyond threshold period. The family members, neighbor or doctor will be informed consequently if the elderly is recognized to be unsafe by the system. Initial rules can be established for the safety checking, while the living pattern of the elderly can be used to refine the rules so as to better detect the unsafe situation. Typically the living pattern of the elderly can be reflected as a set of movement queue [1], but here in this paper, the spaces where the elderly stay with trial duration are ignored in the queue. Therefore the living pattern in this research contains only those spaces where the elderly stay with more than threshold duration. In such way, the volume of pattern and related rules are simplified and the rules can be defined and trained all the time.

There are many available location detection techniques [8–14]. Though ZigBee technique [8] and RSSI-based estimation algorithms [15, 16] are used in our research, other location techniques could be used to enhance the accuracy of the location detection when there is no critical budget limitation.

This chapter is organized as follows. Location detection of the elderly in smart home is suggested in Sect. 2. Section 3 demonstrates the implementation of algorithm. The study case is showed in Sect. 4. Conclusion is made in Sect. 5.

170.2 Location Detection of the Elderly

RSSI-based location approach is used in this research to detect the location of the elderly. Several anchor nodes have been deployed around the smart home and a blind node is suggested to be attached with the elderly. Assuming the four anchor nodes have been known as $P_i = (x_i, y_i, z_i)$ where $i = 1, 2, 3, 4$, and blind node $p(x, y, z)$ with its distance d_i , where $i = 1, 2, 3, 4$, then there is the following Eq. 170.1.

$$\begin{cases} (x - x_1)^2 + (y - y_1)^2 + (z - z_1)^2 = d_1^2 \\ (x - x_2)^2 + (y - y_2)^2 + (z - z_2)^2 = d_2^2 \\ (x - x_3)^2 + (y - y_3)^2 + (z - z_3)^2 = d_3^2 \\ (x - x_4)^2 + (y - y_4)^2 + (z - z_4)^2 = d_4^2 \end{cases} \tag{170.1}$$

Formula (170.1) can be translated into Eq. 170.2.

$$\begin{cases} 2(x_2 - x_1)x + 2(y_2 - y_1)y + 2(z_2 - z_1)z = x_2^2 - x_1^2 + y_2^2 - y_1^2 + z_2^2 - z_1^2 - d_2^2 + d_1^2 \\ 2(x_3 - x_1)x + 2(y_3 - y_1)y + 2(z_3 - z_1)z = x_3^2 - x_1^2 + y_3^2 - y_1^2 + z_3^2 - z_1^2 - d_3^2 + d_1^2 \\ 2(x_4 - x_1)x + 2(y_4 - y_1)y + 2(z_4 - z_1)z = x_4^2 - x_1^2 + y_4^2 - y_1^2 + z_4^2 - z_1^2 - d_4^2 + d_1^2 \end{cases} \tag{170.2}$$

The equation set can be expressed as Eq. 170.3.

$$AX = b \tag{170.3}$$

where,

$$A = \begin{bmatrix} x_2 - x_1 & y_2 - y_1 & z_2 - z_1 \\ x_3 - x_1 & y_3 - y_1 & z_3 - z_1 \\ x_4 - x_1 & y_4 - y_1 & z_4 - z_1 \end{bmatrix}$$

$$b = \begin{bmatrix} x_2^2 - x_1^2 + y_2^2 - y_1^2 + z_2^2 - z_1^2 - d_2^2 + d_1^2 \\ x_3^2 - x_1^2 + y_3^2 - y_1^2 + z_3^2 - z_1^2 - d_3^2 + d_1^2 \\ x_4^2 - x_1^2 + y_4^2 - y_1^2 + z_4^2 - z_1^2 - d_4^2 + d_1^2 \end{bmatrix}$$

The final result of the Least Squares Estimation can be expressed as Eq. 170.4.

$$X = (A^T A)^{(-1)} A^T b \tag{170.4}$$

We can get the unknown point p according to the formula above. However, the distance d_i , where $i = 1, 2, 3, 4$, which we get with RSSI technology has some error. Therefore, the final results also have a certain degree of errors. To address this issue, we get different results from different anchor nodes and take the average of these results as the final data, so as to reduce the error to some degree.

170.3 Algorithms for the Elderly Personal Safety Checking

The space is defined for the initial position indicates in the default rules. A space can be a room or part of a room when there are different rules for different parts.

Definition 1 Space is defined as $S = (\text{sid}, x_1, y_1, z_1, x_2, y_2, z_2)$ where sid is the identification no of the space; x_1, y_1, z_1 and x_2, y_2, z_2 indicate a cube which is wrapped by the planes $x = x_1, y = y_1, z = z_1, x = x_2, y = y_2$ and $z = z_2$.

Definition 2 Space cluster is a group of spaces which have the similar rules in the smart home. It is defined as a set of space, $SC = \{\text{scid}, s_1, s_2, \dots, s_n\}$.

Definition 3 Movement record is the record of movement of the elderly in the smart home. It is defined as $MR = (T, x, y, z)$, where T is the time for the record, x, y and z is the location of the elderly. All the location is captured by the location estimation algorithm.

Definition 4 Living pattern $LP = (\text{sc}, \text{minD}, \text{mediumD}, \text{maxD}, \text{alertD})$, where sc is the space cluster, minD, mediumD, and maxD refers to the minimal, medium and maximal duration the elderly has ever stayed at the space cluster each a time. All these durations are measured and refined in the training. The alertD is the threshold duration for the elderly to stay at the space cluster and the system would alert the related person for this unsafe situation. Living pattern in this paper has been simplified as some mapping rules, which are defined on the basis of space cluster. In order to calculate the medium duration, the summary SumD and count CD of the duration for this space cluster will be calculated at first and the value of mediumD is the factor of the SumD and CD ($\text{medium} = \text{SumD}/\text{CD}$).

Definition 5 Safety checking system is defined as $\text{SACHS} = (S, SC, MR, LP, F, \text{Default})$, where S is a set of spaces, SC is a set of space clusters, MR is all the movement record, LP is the living pattern which indicates the rules, F is a serial of algorithms including the algorithms for location estimation, hot space detection, space cluster manual merging, and safety checking. Default contains the default parameters including the pre-defined inferior limit (PIL: unit s), pre-defined upper limit (PUL: unit s), and the error of location estimation (ELE: unit m).

Hot space detection algorithm. Hot space is a kind of space where the elderly stays for a long duration more than a pre-defined inferior limit. Hot space detection is a training algorithm.

Input: MR, SACHS

Output: Hot space

Sort MR in the order of T;

2 ele = SACHS. Default.ELE, pil = SACHS. Default.PIL, pul = SACHS.

Default.PUL;

3 FOR (each mr in MR) {

3 BeginT = MR.T;

4 beginL.x=MR.x, beginL.y=MR.y, beginL.z=MR.z;

```

5 DO
6 fetch next mr in MR;
7 WHILE (square(beginL.x-MR.x)+square(beginL.y-MR.y)+square(beginL.z-
MR.z)<ele)
8 IF (MR.T - BeginT>pil) {
9 tempSid = SACHS.Default.newSpaceID(beginL.x, beginL.y, beginL.z);
10 IF(tempSid>0)//the space has been created already
11 tempscid = SACHS.SC.findBySpace(tempSid);
12 ELSE {
13 SACHS.S.add(temps = new s(-tempSid, beginL.x-ele, beginL.y-ele, be-
ginL.z-ele, beginL.x+ele, beginL.y+ele, beginL.z+ele)
14 tempscid = SACHS.Default.newSpaceClusterID();
15 SACHS.SC.add(tempssc = new sc(tempscid,temps));
16 }
17 lp = SACHS.LP.findBySC(tempscid);
18 IF(lp < 0)
19 SACHS.LP.add(tempscid, MR.T - BeginT);
20 ELSE
21 SACHS.LP.update(tempscid, MR.T - BeginT);
22 }
23 }

```

In the training model, when a hot space *hs* is detected by the algorithm, a related space cluster will be generated, which only contain one item *hs*. Such space cluster can be refined or merged to some similar space cluster manually. *SACHS.Default.newSpaceID* is used to generate a new space id if there not exists any space contain the input location, considering the error *ele*. *SACHS.LP.add* is used to add the *tempscid* into the living pattern, init the *minD*, *mediumD*, *maxD*, *SumD* with the input duration and init the *CD* with 1. *SACHS.LP.update* will update the related living pattern with input duration accordingly.

170.4 Case Study

Movement record was simulated by a program which provided random duration, random 3-D movement (*x*, *y* and *z*). The program generated the record every 5 s. Table 170.1 lists the hot spaces detected by the algorithm. *AlertD* is assigned with the *maxD* plus 300 s.

For each space, it is recognized as safe if the elderly stay within the *maxD*, unsafe is the elderly stay beyond the *maxD* but no exceed *alertD*. If the elderly stay more time than the *alertD*, the system assumes that there is a danger. Any observation on non-hot space but with duration more than the *PIL* will be recognized as abnormal as no hot space could be detected. Because the data volume is small, manual calibration is easy and enabled all the time. The space cluster can be manually merged when there are of similar rules. For example, in Table 170.1,

Table 170.1 Hot spaces detected by the algorithm

Space cluster	Coordinate (m)		minD (s)	MediumD (s)	MaxD (s)	AlertD (s)
1	(1, 6, 1)	(1.5, 6.5, 1.5)	60	300	500	800
2	(4, 5, 1)	(4.5, 5.5, 1.5)	60	600	1,200	1,500
3	(0, 4, 1)	(0.5, 4.5, 1.5)	60	600	1,740	2,040
4	(1, 4, 1)	(1.5, 4.5, 1.5)	60	600	7,200	7,500
5	(3, 4, 0.5)	(3.5, 4.5, 1)	60	1,200	2,400	2,700
6	(4, 4, 0.25)	(4.5, 4.5, 0.75)	60	14,400	28,800	29,200
7	(4, 6, 1)	(4.5, 6.5, 1.5)	60	600	1,200	1,500
8	(5, 6, 1)	(5.5, 6.5, 1.5)	60	600	1,200	1,500

Table 170.2 Elderly personal safety checking

Observation	Location	Time	Safety checking
1	(1.25, 6, 1.25)	12:10	Safe in space cluster 1
2	(0.25, 4.25, 1.25)	12:15	Unsafe in space cluster 3
3	(1.25, 4.5, 1.25)	12:45	Unsafe in space cluster 4
4	(4.5, 4.5, 0.25)	14:45	Safe in space cluster 6
5	(3, 4, 0)	15:50	Abnormal space cluster NA
6	(3, 4, 0.5)	16:00	Safe in space cluster 5
7	(4, 5, 1)	16:40	Danger in space cluster 2
8	–	17:20	–

space cluster 2, 7, and 8 could be merged to one space cluster. In such way, the rules could be reduced though the spaces remain the same.

Table 170.2 shows the checking results on some test cases. All these test cases were designed after the hot spaces had been detected. For the observation 1, 4 and 6, the elderly stayed in space cluster 1, 6, and 5. Since the durations (300 s, 3,900 s, and 2,400 s) were no more than the related maxD (500 s, 28,800 s, and 2,400 s), these three observations are safe. For the observation 2, the elderly was in space cluster 3, however the duration is 1,800 s which is between the maxD (1,740 s) and alertD (2,040 s), the situation is unsafe. For the observation 7, the elderly stayed in space cluster 2 for 2,400 s which is larger than alertD (1,500 s), it is assumed to be danger. Observation 5 indicated that the elderly stay in some other places other than available hot spaces and is recognized as abnormal.

170.5 Conclusion and Discussion

Elderly personal safety has become a social problem globally due to the increasing ratio of elderly. Smart home has been suggested to solve such problem. Different with other research, here in this paper, an economic solution for elderly personal safety checking has been proposed in this paper. ZigBee and RSSI based location

estimation technique have been suggested; however, other research may choose some expensive solutions to get accurate location. Space, space cluster, movement record, living pattern and SACHS system have been defined while living pattern has been simplified so as to reduce the data volume of the rules and enable better calibration of the rules. Hot space has been used to indicate the locations where the elderly prefer to stay and checking rules have been defined as 4 levels of duration including minimal, medium, maximal and alert duration. Different status could then be recognized, such as safe, unsafe, abnormal, and danger.

Though the error of the location detection has been considered in this research, and a parameter name *ele* has been introduced in the hot space algorithm, the impact of error is still there. The more accurate location detection results in more accurate safety checking. Due to the cost of infrastructure, smart home may not deploy with some expensive location detection solution. Our group will try to design better accurate location detection solution with reasonable cost in the future research.

Acknowledgments This research is supported by Zhejiang Gongshang University, China with No#Xgz1102 to Dr. Weigang Chen (High Education Research Project). This research was financial supported by National Natural Science Foundation of China with No. 60873022 to Dr. Hua Hu, The Science and Technology Department of Zhejiang Province, China with No. 2008C11009 to Dr. Bin Xu, and Education Department of Zhejiang Province with No. 20061085 to Dr. Bin Xu.

References

1. Chang CK, Jiang H, Ming H, Oyama K (2009) Situ: a situation-theoretic approach to context-aware service evolution. *IEEE Trans Serv Comput* 2(3):261–275
2. The Xinhua news agency: Xinhua news (2010-01-03). <http://www.chinamail.com.cn/jfjbmap/content/19524.html>
3. Yang H, Helal A (2008) Safety enhancing mechanisms for pervasive computing systems in intelligent environments. In: *Proceedings of PerCom*, pp 525–530
4. Álamo JMR, Wong J, Babbitt R, Yang H, Chang CK (2009) Using web services for medication management in a smart home environment. In: *ICOST*, pp 265–268
5. Ming H, Chang CK, Oyama K, Yang H (2010) Reasoning about human intention change for individualized runtime software service evolution. In: *COMPSAC*, pp 289–296
6. Zhou F, Yang H, Álamo JMR, Wong JS, Chang CK (2010) Mobile personal health care system for elderly with diabetes. In: *ICOST*, pp 94–101
7. Helal S, Mitra S, Wong JS, Chang CK, Mokhtari M (2008) Smart homes and health telematics. In: *Proceedings of the 6th international conference, ICOST*, Ames. Springer, Berlin
8. Wheeler A (2007) Commercial applications of wireless sensor networks using ZigBee. *IEEE Commun Mag* 45(4):70–77
9. Hightower J, Borriello G (2001) Location systems for Ubiquitous computing. *IEEE Comput* 34(8):57–66
10. Pahlavan K, Li X, Makela JP (2002) Indoor geolocation science and technology. *IEEE Commun Mag* 40(2):112–118

11. Small J, Smailagic A, Siewiorek PD (2001) Determining user location for context aware computing through the use of a wireless LAN infrastructure. *ACM mobile networks and applications*, vol 6. pp 0256–0262
12. Xiang Z, Song S, Chen J, Wang H, Huang J, Gao X (2004) A wireless LAN-based indoor positioning technology. *IBM J Res Dev* 48(5/6):617–626
13. Kaemarungsi K (2005) Design of indoor positioning systems based on location fingerprinting technique. Ph.D.Thesis, University of Pittsburgh, vol v1. pp 181–186
14. Bahl P, Padmanabhan VN (2000) Radar: an in-building rf-based user location and tracking system. In: *Proceedings of IEEE INFOCOM*
15. Aamodt K (2011) Application note AN042, CC2431 location engine, Chipcon products from Texas instruments. Available at <http://focus.ti.com/lit/an/swra095/swra095.pdf>. Accessed 30 April 2011
16. Álvarez Y, de Cos ME, Lorenzo J, Las-Heras F (2010) Novel received signal strength-based indoor location system: development and testing. *EURASIP J Wirel Commun Netw* 2010:11–18

Chapter 171

Research on Maximum Benefit of Tourist Enterprises Based on the Influence of Scenic Spot Ticket Discount Amount

Li Jian, Han Na and Shu Bo

Abstract Price of scenic spot ticket has a great influence on tourism demand, so the ticket discount amount has an important effect on the amount of tourists. Based on this view, this paper researched the maximum benefit of the tourist alliance composed of scenic spot and travel agency from the view of ticket discount amount. Study found that scenic spot should cooperate with travel agency, and adopted a certain discount to tourists. Then the alliance could achieve the maximum benefit. Finally, we worked out the shapely value for each enterprise in the case of cooperation. By comparison, we found that the shapely value was much higher than the benefit in the case of non-cooperation.

Keywords Tourism alliance · Ticket discount · Shapely value

L. Jian (✉)

School of Social Science, Beijing Institute of Technology, Beijing 100081, China
e-mail: lijianbit@bit.edu.cn

H. Na · S. Bo

School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China
e-mail: hanna110@126.com

S. Bo

e-mail: shubo@ysu.edu.cn.

S. Bo

School of Economics and Management, Yanshan University, Qinhuangdao 066004, China

171.1 Proposing the Problems

Tourist products have the characteristic of comprehensiveness, any single tourist enterprise cannot provide all the products. In order to meet the needs of tourist, related enterprises must support each other. So these enterprises combine into a tourist alliance in order to share economic benefit. During the process of purchasing and distribution of tourist products, travel agency becomes a mediator connecting the upstream and downstream firms. On the one hand, travel agency becomes the important distribution institution of scenic routes; On the other hand, travel agency integrates tourist products and then sales them to tourists. So scenic spot and travel agency should form a tourist alliance, to achieve the overall maximum benefit. In recent years, tourist enterprises have made a great progress, but they still have many problems, such as low price competition, target market positioning inaccurate [1][2], small scale, weak market competitiveness[3], lack of vision of brand construction[1] and so on. These problems lead to the phenomenon of serious product convergence, profit space sharply reducing, and corporate reputation badly damaged, which go against the long-term development of enterprises. So, the problem of how to improve the benefit of tourist enterprises becomes an urgent need to resolve.

Because of the characteristic of service, the price of tourist product has obvious effect on purchasing option. So we believe that tourist product price determines the amount of tourist to some extent. From the psychology perspective price discount has big effect on consumer purchasing behavior. During the process of development, enterprises also adopt price promotion to attract customer in order to improve short-term sales. So, no matter for enterprise or consumers, product price discount can bring certain advantage. Price promotion is an important means of the marketing combination. It refers to a marketing tool that in a certain period the manufacturers or channel participants apply the lower price of product, or increase the quantity of a commodity with the original price [4]. Price promotion is probably the most common method in many promotion means, and is widely used for various demand purpose. Such as boosting demand, improving sales, maintaining or improving market share, etc. As we found, the appropriate price promotion can bring sales growth, but deep discount promotion has negative influence on the perceived quality, and then reduces corporate image and brand equity [5]. So the degree of discount taken by tourist enterprise becomes an unavoidable problem. So far, the research on the relationship between tourist product price discount amount and enterprise benefit is still tiny. So we think that it is necessary to study on the profit of alliance composing of scenic spot and travel agency from the perspective of scenic spot ticket discount amount, to explore the optimal discount amount, which can provide theoretical basis for the scenic spot tickets distribution.

171.2 Game-Theory Analysis of the Alliance Based on the Influence of Scenic Spot Tickets Discount Amount

Scenic spot, travel agency combine a benefit alliance. They can share resources, strengthen the learning ability of organization, improve the competitiveness of enterprises and the ability of anti-risk. However, a successful benefit alliance is based on reasonable profit distribution, namely the relationship between the resources contribution. The distribution of benefit decides this alliance's stability [6]. Scenic spot ticket not only ensures the rights of tourist, improves tourist utility, but also guarantees scenic spot obtain profit to maintain normal development of scenic spot protection and repair. As a mediator, travel agency connects with scenic spots and tourists, this organization integrates the tourist products and sales to tourists to achieve profit. Tourist wants to pursue a happy experience with reasonable price. If scenic spot overlooks the distribution channel—travel agency, it would lose a lot of tourists. Likewise, if travel agency makes the price blindly, regardless of scenic spot, also incurs tourists' discontent. So any decision will affect other participants' benefit. These two participants how to make decision rationally, realize the maximum profit of the alliance becomes inevitable problems. This paper uses game theory to analyze enterprises' collaboration and benefit allocation from the view of scenic spot ticket discount amount.

There are two paths to distribute scenic spot tickets: one is the direct route, tourists buy the tickets directly in the scenic spot, without discount; another is indirect path, scenic spot sales tickets to travel agency with some discount. Then visitors buy tickets from travel agency with a certain discount. In the indirect path, the price that scenic spot sales tickets to travel agency should lower than the price that visitors buy tickets from travel agency, or travel agency will loss. How to weigh the benefit of scenic spot and travel agency, whether there is a most point, which becomes the starting point of this paper.

Due to the tourists' consumption gradually maturity, transparent of the market information and tourist enterprises know the visitors' demand function, in the real operation the order is: firstly, scenic spot formulates preferential policy; secondly, travel agency determines whether to promote this tourist product according to the preferential policy; thirdly, visitor determines whether to travel according to the price. Therefore, we think that the model belongs to the complete information dynamic game. In this model, the game of the participants are scenic spot and travel agency, while the Nash equilibrium that meets the need of these two participants corresponding to the ticket discount amount is the optimal solution.

171.2.1 Profit Function of the Benefit Participants

Tourist demand refers to the quantity that in a certain time customers are willing and able to buy with certain price [7]. Due to the characteristic of service, tourist

enterprises must take full account of the degree of price sensitive when making price adjustment to avoid the mistakes of price decision [8]. When ticket price changes, the change of tourist routes demand is determined by the degree of price sensitive [9]. In the case of considering ticket price discount amount, the author supposes that demand function is: $Q = D - \lambda\beta_iP$, D indicates tourist market's largest demand, λ denotes the degree of price sensitive, β_i denotes different discount amount, P denotes ticket price. Q_1 denotes amount that visitors purchase tickets directly in scenic spot, β_1 denotes scenic spot ticket for tourist discount; β_2 denotes scenic spot ticket for travel agency discount, β_3 denotes travel agency for visitor discount; Q_2 denotes amount that visitors purchase ticket through travel agency. Then we can know: $Q_1 = D - \lambda\beta_1P, Q_2 = D - \lambda\beta_3P$, and calculate the profit function of the benefit subjects.

The profit function of scenic spot is:

$$\pi_1 = Q_1\beta_1P + Q_2\beta_2P = (D - \lambda\beta_1P)\beta_1P + (D - \lambda\beta_3P)\beta_2P \tag{171.1}$$

The profit function of travel agency is:

$$\pi_2 = Q_2(\beta_3P - \beta_2P) = (D - \lambda\beta_3P)(\beta_3P - \beta_2P) \tag{171.2}$$

Constraint condition is: $\beta_3 \geq \beta_2$.

When making preferential policies, scenic spot can choose to formulate preferential policies based on the actual situation and travel agency just passive accepts. Also, scenic spot can negotiate with travel agency and form a benefit alliance. So there are two situations would be discussed in this paper, one is that scenic spot and travel agency don't cooperate, another is these two enterprises cooperate.

171.2.2 Model of Non-cooperation

When these two enterprises don't cooperate, travel agency just passive accepts the discount amount made by scenic spot, which forms a dynamic game model, so both enterprises aim at respective maximum profit. According to the model of Stackelberg, we adopt reverse induction to solve. Travel agency considers its own maximum profit, we could get a formula as follow:

$$\beta_3 = \frac{D + \lambda\beta_2P}{2\lambda P} \tag{171.3}$$

According to the formula 171.1 and 171.3, and the principles of scenic spot profit maximization, we could get: $\beta_1 = \frac{D}{2\lambda P}$, $\beta_2 = \frac{D}{2\lambda P}$, meeting constraint condition $\beta_3 \geq \beta_2$. Under the equilibrium condition, maximum profit of scenic spot is: $\pi_1 = \frac{3D^2}{8\lambda}$, maximum profit of travel agency is: $\pi_2 = \frac{D^2}{16\lambda}$, the gain for these two enterprises is: $\pi_N = \frac{7D^2}{16\lambda}$

Table 171.1 Calculation of the basic benefits distribution of travel agency

S	{Travel Agency}	{Travel Agency,scenic spot}
$V(S)$	π_2	π
$V[S \setminus \{i\}]$	0	π_1
$V(S) - V[S \setminus \{i\}]$	π_2	$\pi - \pi_1$
$ S $	1	2
$\frac{(S -1)!(n- S)!}{n!}$	$\frac{1}{2}$	$\frac{1}{2}$
$\phi_i(V) = \sum_{S(i \in S)} \frac{(S -1)!(n- S)!}{n!} [V(S) - V[S \setminus \{i\}]]$	$\frac{\pi_2}{2}$	$\frac{\pi - \pi_1}{2}$
Shapley value of Travel Agency	$\pi'_2 = \frac{\pi_2}{2} + \frac{\pi - \pi_1}{2} = \frac{7D^2}{32\lambda}$	

171.2.3 Model of Cooperation

If scenic spot and travel agency cooperate to form a benefit alliance, they negotiate the discount amount β_i , in this case the game forms a static model. The overall benefit function of the alliance is: $\pi = \pi_1 + \pi_2 = D\beta_1P - \lambda\beta_1^2P^2 + D\beta_3P - \lambda\beta_3^2P^2$. As we can see, this profit function doesn't have β_2 , so in the case of cooperation, β_2 has no effect on the overall benefit function of the alliance. According to the principle of the maximum benefit, then:

$$\beta_1 = \beta_3 = \frac{D}{2\lambda P} \tag{171.4}$$

The maximum benefit of the alliance is: $\pi_C = \frac{D^2}{2\lambda}$

Comparing these two different overall benefit π_N and π_C , we could find that the overall benefit in the case of cooperation higher than non-cooperation, so scenic spot and travel agency should cooperate.

171.3 Benefit Distribution Based on Shapley Value

Shapley value [10] is a concept belongs to cooperative game theory, it is a solution used to distribute benefit in the case of cooperation, the benefit for every single participant is called shapely value, written for:

$$\phi_i(V) = \sum_{S(i \in S)} \frac{(|S|-1)!(n-|S|)!}{n!} [V(S) - V[S \setminus \{i\}]].$$

$|S|$ refers to the participant amount

in the cooperation, $V(S)$ refers to benefit of the alliance, $V(S \setminus \{i\})$ refers to the benefit achieved by all the participant other than the participant 'I' [11]. Using Shapley value, we could work out benefit for every cooperator. The benefit of travel agency is shown in Table 171.1.

Then we can work out the shapely value of scenic spot: $\pi'_1 = \pi - \pi'_2 = \frac{9D^2}{32\lambda}$. Compared with the case of non-cooperation, the shapely values of tourist

enterprises are higher than the benefit of non-cooperation. So it is advantage to cooperate. In the future, scenic spot and travel agency should cooperate and sale tickets to tourist at a discount: $\beta_1 = \beta_3 = \frac{D}{2\lambda P}$, to increase the amount of tourist and improve benefit of tourism enterprises.

171.4 Conclusion

Applying the game theory, this paper analyzed the interests of scenic spot and travel agency from the angle of tickets discount amount. Through the research, the author found that scenic spot could cooperate with travel agency, and sale tickets to visitor at a certain discount, then the alliance composed of these two participants could achieve maximum benefit. Finally, the author distributed the benefit for each participant adopting shapely value, finding that the benefit in the case of cooperation was much higher.

During the research, the paper was discussed under a complete information situation. In the future, we should pay more attention to cooperation under the situation of incomplete information. Also, in this paper the author just considered two enterprises, in the future, we could consider some other tourist enterprises to compose an alliance.

References

1. Li L (2008) Remarks on brand development strategy of tourism enterprises in China [J]. *Enterp Econ* 2:108 (in Chinese)
2. Lijian Z, Qiaolin G (2006) Tourism image positioning and the analysis of error root [J]. *J Tour* 6(21):50–51 (in Chinese)
3. Lingge S (2004) Cooperation marketing: the way to win for the tourism enterprise [J]. *Enterp vitality* 5:40 (in Chinese)
4. Raghur P, Coffman K (1999) When do price promotion affect pretrial brand evaluation [J]. *J Mark* 36:211–222
5. Minghua J, Weimin D (2003) Empirical study on the effect of price promotion discount on brand equity [J]. *J Peking Uni philos soc sci ed* 5(40):54 (in Chinese)
6. Renjun Z, Bo W (2008) The study of evolutionary game theory with the strategic alliance of tourism enterprise [J]. *Stat Decis* 21:183 (in Chinese)
7. Jigang B, YiFang C, Hua P (1993) *Tourism geography* [M]. Beijing: Senior Education Press, (in Chinese)
8. Jiang D (2001) *Travel service management* [M]. Tianjin: Nankai University Press, (in Chinese)
9. Yang L, Bangyi L, Weiguo L (2009) The study on profit distribution and coordination of tourism supply chain based on the pricing of tourism product. coordination [J]. *Ecol Econ* 2(24):107 (in Chinese)
10. Shapley LS (1953) A value for n-person games [J]. *Ann Math Stud* 28:307–318
11. Guangming H, Cunjin L (2005) *Management game theory* [M], Vol 2. Beijing Institute of Technology Press, Beijing, p 19

Part XVI
Web Service and Data Mining

Chapter 172

A Fuzzy Multi-Criteria Group Decision Making Approach for Hotel Location Evaluation and Selection

Santoso Wibowo and Hepu Deng

Abstract This paper presents a fuzzy multi-criteria group decision making approach for effectively solving the hotel location evaluation and selection problem. Pairwise comparison is used to help individual decision makers make their subjective assessments in evaluating the performance of alternative hotel locations and the relative importance of the selection criteria in a cognitively less demanding manner. A consensus building process is proposed for ensuring the achievement of consensus at an acceptable level in the evaluation process. An algorithm is developed for determining the overall performance of each alternative location across all the criteria on which the selection decision is made. An example is presented for demonstrating the applicability of the approach for solving the location selection problem in real world situations.

Keywords Uncertainty and imprecision · Multi-criteria analysis · Group decision making · Hotel location evaluation and selection

172.1 Introduction

The tourism industry is growing rapidly nowadays [1, 2]. A forecast from the Asian Pacific Travel Association indicates that this industry will be the fastest growing industry over the next decade [1]. Within such an important industry, the development of new hotels is a critical part of the tourism business [3].

S. Wibowo (✉) · H. Deng (✉)

School of Business IT and Logistics, RMIT University, Melbourne, Australia
e-mail: santoso.wibowo@rmit.edu.au

H. Deng

e-mail: hepu.deng@rmit.edu.au

To develop and maintain the competitive advantage in the tourism industry, selecting the most appropriate location for hotel development is critical [2]. This is because the selection of the most suitable hotel location has strategic implications to the development of the tourism business including the increase of market share and the improvement of profitability [3]. As a result, evaluating and selecting the most suitable location from available locations to develop becomes a critical decision to be made.

Evaluating and selecting the most suitable location for development is challenging. This is due to (a) the involvement of multiple decision makers (DMs) and the presence of multiple, often conflicting criteria, (b) the need for adequately modelling the uncertainty and imprecision present, and (c) the cognitive demanding on the DMs. Furthermore, it is critical to reach a certain level of agreement among the DMs for facilitating the acceptance of the decision made. As a result, structured approaches are desirable for effectively solving the location evaluation and selection problem.

This paper presents a fuzzy multi-criteria group decision making approach for effectively solving the hotel location evaluation and selection problem. A pairwise comparison process is used to help individual DMs make their subjective assessments in evaluating the performance of alternative hotel locations and the relative importance of the selection criteria in a cognitively less demanding manner. A consensus building process is proposed for ensuring the achievement of consensus at an acceptable level. A fuzzy multi-criteria algorithm is developed for evaluating the overall performance of alternative hotel locations across all the criteria on which the selection decision is made. An example is presented for demonstrating the applicability of the proposed approach for solving the hotel location evaluation and selection problem.

172.2 Some Preliminary Concepts

A fuzzy number is a convex fuzzy set [4], characterized by a given interval of real numbers, each with a grade of membership between 0 and 1. Triangular fuzzy numbers are a special class of fuzzy number, defined by three real numbers expressed as (a_1, a_2, a_3) whose membership function is described as

$$\mu_A(x) = \begin{cases} \frac{x-a_1}{a_2-a_1}, & a_1 \leq x \leq a_2, \\ \frac{a_3-x}{a_3-a_2}, & a_2 \leq x \leq a_3, \\ 0, & \text{otherwise.} \end{cases} \quad (172.1)$$

where a_2 is the most possible value of fuzzy number A , and a_1 and a_3 are the lower and upper bounds respectively used to illustrate the fuzziness of the data evaluated.

Table 172.1 Linguistic variables and their fuzzy number approximations

Linguistic variables	Fuzzy numbers	Membership function
Very poor (VP)	$\tilde{1}$	(1, 1, 3)
Poor (P)	$\tilde{3}$	(1, 3, 5)
Fair (F)	$\tilde{5}$	(3, 5, 7)
Good (G)	$\tilde{7}$	(5, 7, 9)
Very good (VG)	$\tilde{9}$	(7, 9, 9)

Fuzzy numbers are widely used to approximate the linguistic variables used for expressing the DM’s subjective assessments in decision making. To facilitate the making of pairwise comparison, linguistic variables [5] are used. These linguistic variables are approximated by triangular fuzzy numbers defined as in Table 172.1.

Fuzzy extent analysis is widely used for deriving the criteria weights and the alternative performance ratings from the reciprocal matrices resulting from the pairwise comparison process [6, 7] due to its simplicity in concept and computational efficiency. Assume that $X = \{x_1, x_2, \dots, x_n\}$ is an object set, and $U = \{u_1, u_2, \dots, u_m\}$ is a goal set. Fuzzy assessments are performed on each object for each goal respectively, resulting in m extent analysis values for each object, given as $\mu_i^1, \mu_i^2, \dots, \mu_i^m, i = 1, 2, \dots, n$, where all $\mu_i^j (j = 1, 2, \dots, m)$ are fuzzy numbers representing the performance of the object x_i on goal u_j . Using fuzzy synthetic extent analysis, the overall performance of the object x_i across all goals can be determined by

$$S_i = \frac{\sum_{j=1}^m \mu_i^j}{\sum_{i=1}^n \sum_{j=1}^m \mu_i^j}, \quad i = 1, 2, \dots, n. \tag{172.2}$$

172.3 A Fuzzy Multi-Criteria Approach

Evaluating and selecting alternative hotel locations usually involves in (a) discovering all alternative locations $A_i (i = 1, 2, \dots, n)$, (b) identifying the selection criteria $C_j (j = 1, 2, \dots, m)$, (c) assessing the performance of alternative locations and the weight of criteria by DMs $D_k (k = 1, 2, \dots, s)$, (d) aggregating the assessments for producing an overall performance index for each alternative location across all the criteria, and (e) selecting the best alternative location [8].

The selection process starts with determining the performance of alternative locations with respect to each criterion and the importance of the criteria. To reduce the cognitive demanding on the DMs, pairwise comparison [5] using the linguistic variables described in Table 172.1 is used, leading to a pairwise judgment matrix as

$$A = \begin{bmatrix} 1 & a_{12} & \dots & a_{1k} \\ a_{21} & 1 & \dots & a_{2k} \\ \dots & \dots & \dots & \dots \\ a_{k1} & a_{k2} & \dots & 1 \end{bmatrix} \tag{172.3}$$

Using the fuzzy synthetic extent analysis in (172.2), the criteria weightings and performance rating for DMs D_k with respect to criterion C_j can be obtained, resulting in the determination of the fuzzy decision matrix for the alternatives and the fuzzy weighting vector for the selection criteria respectively as

$$Y^k = \begin{bmatrix} y_{11}^k & y_{12}^k & \dots & y_{1m}^k \\ y_{21}^k & y_{22}^k & \dots & y_{2m}^k \\ \dots & \dots & \dots & \dots \\ y_{n1}^k & y_{n2}^k & \dots & y_{nm}^k \end{bmatrix} \tag{172.4}$$

$$w^k = (w_1^k, w_2^k, \dots, w_m^k) \tag{172.5}$$

By averaging the fuzzy assessments of individual DMs as given in (172.4) and (172.5), the overall fuzzy decision matrix and the fuzzy weight vector can be obtained as

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \tag{172.6}$$

$$W = (w_1, w_2, \dots, w_m) \tag{172.7}$$

where $x_{ij} = \frac{\sum_{k=1}^s y_{ij}^k}{s}$ and $w_j = \frac{\sum_{k=1}^s w_j^k}{s}$.

To explore the degree of consensus among DMs, the proximity measure is introduced. It measures the distance between the assessments of individual DMs and the group assessments. The proximity measure helps DMs determine the direction of changes in their assessments for improving their consensus level.

Several proximity measures are developed [9–11] for solving group decision making problems with sounding applications. They, however, suffer from various shortcomings including (a) the need for tedious mathematical computation and (b) cognitively very demanding on the DMs. To overcome these shortcomings, the graded mean integration representation distance based measure [12] is introduced due to the accuracy [12], simplicity in concept, and efficiency in computation. The degree of proximity between DMs on the performance ratings of alternatives is determined as

$$S_{ij}^k = \frac{1}{1 + \left[\left(\frac{y_{ij(L)}^k + 4y_{ij(M)}^k + y_{ij(R)}^k}{6} \right) - \left(\frac{x_{ij(L)} + 4x_{ij(M)} + x_{ij(R)}}{6} \right) \right]} \tag{172.8}$$

where $y_{ij(L)}^k$, $y_{ij(M)}^k$, and $y_{ij(R)}^k$ represent the lower bound, middle bound, and upper bound of individual DM's assessments, and $x_{ij(L)}$, $x_{ij(M)}$, and $x_{ij(R)}$ are the lower bound, middle bound, and upper bound of the group assessments about the performance rating of location alternative A_i with respect to criterion C_j respectively.

Similarly, the degree of similarity between DMs on criteria weighting is

$$T_j^k = \frac{1}{1 + \left[\left(\frac{w_{j(L)}^k + 4w_{j(M)}^k + w_{j(R)}^k}{6} \right) - \left(\frac{w_{j(L)} + 4w_{j(M)} + w_{j(R)}}{6} \right) \right]} \tag{172.9}$$

where $w_{j(L)}^k$, $w_{j(M)}^k$, and $w_{j(R)}^k$ represent the lower bound, middle bound, and upper bound of individual DM's assessments, and $w_{j(L)}$, $w_{j(M)}$, and $w_{j(R)}$ represent the lower bound, middle bound, and upper bound of the group assessments respectively.

A consistency measure (CM) is proposed for identifying whether individual DMs' opinions are within the acceptable level of the consensus threshold. It is calculated by comparing the CM of individual DMs with the consensus threshold. A CM larger than the consensus threshold indicates that the DM's opinion is consistent to the group opinion. Otherwise, the DM is requested to change the assessments. The CM for the group on the performance rating and the criteria weight can be defined as

$$CM = \max d(S_{ij}^k, T_j^k) \tag{172.10}$$

The consensus building process can be summarized as follows:

Step 1. Obtain the criteria weighting and performance rating with respect to criterion C_j for each DM using fuzzy synthetic extent analysis in (172.2).

Step 2. Determine the decision matrix by each DM as in (172.4).

Step 3. Determine the fuzzy weighting for the selection criteria by each DM as in (172.5).

Step 4. Calculate the overall fuzzy decision matrix for the DMs by averaging the fuzzy assessments made by individual DMs as given in (172.6).

Step 5. Calculate the fuzzy weight for all the DMs by averaging the fuzzy assessments made by individual DMs as given in (172.7).

Step 6. Calculate the proximity measure between individual DMs' assessments and the group assessments for the ratings on each criterion by (172.8).

Step 7. Calculate the proximity measure between individual DMs' assessments and the group assessments for the criteria weights on each criterion by (172.9).

Step 8. Determine whether individual DMs opinions are within the acceptable level of consensus by (172.10). If the CM of a DM is less than the consensus threshold, the DM goes back to Step 1 for adjusting the assessments. Otherwise, the consensus building process is finalized.

The weighted fuzzy performance matrix representing the overall performance of each alternative on each criterion can then be determined by (172.11) as follows

$$Z = \begin{bmatrix} w_1x_{11} & w_2x_{12} & \dots & w_mx_{1m} \\ w_1x_{21} & w_2x_{22} & \dots & w_mx_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ w_1x_{n1} & w_1x_{n2} & \dots & w_mx_{nm} \end{bmatrix} \tag{172.11}$$

Given the fuzzy vector $(w_jx_{1j}, w_jx_{2j}, \dots, w_jx_{mj})$ for criterion C_j , the fuzzy maximum (M_{\max}^j) and the fuzzy minimum (M_{\min}^j) [13] which represent respectively the best and the worst fuzzy performance ratings among all the alternatives with respect to criterion C_j can be determined as

$$\mu_{M_{\max}^j}(x) = \begin{cases} \frac{x-x_{\min}^j}{x_{\max}^j-x_{\min}^j}, & \mu_{M_{\min}^j}(x) = \begin{cases} \frac{x_{\max}^j-x}{x_{\max}^j-x_{\min}^j}, \\ 0, \end{cases} \end{cases} \tag{172.12}$$

where $x_{\max}^j = \sup(\text{supp } \bigcup_{i=1}^n (w_jx_{ij}))$, and $x_{\min}^j = \inf(\text{supp } \bigcup_{i=1}^n (w_jx_{ij}))$.

The degree to which alternative A_i is the best alternative with respect to criterion C_j can then be determined by calculating the Hamming distance between its weighted fuzzy performance (w_jx_{ij}) with the fuzzy maximum and the fuzzy minimum [13] respectively, given as in (13).

$$h_i^+ = \sum_{j=1}^m H(w_jx_{ij}, M_{\max}^j), \quad h_i^- = \sum_{j=1}^m H(w_jx_{ij}, M_{\min}^j), \tag{172.13}$$

With the use of triangular fuzzy numbers, the Hamming distance between two fuzzy numbers $A = (a_1, a_2, a_3)$ and $B = (b_1, b_2, b_3)$ can be calculated as [14].

$$H(A, B) = |a_1 - b_1| + |a_2 - b_2| + |a_3 - b_3| \tag{172.14}$$

An overall performance index for alternative A_i across all the criteria can be determined by (172.15). The larger the index, the more preferred the alternative.

$$P_i = \frac{(h_i^-)^2}{(h_i^+)^2 + (h_i^-)^2} \tag{172.15}$$

172.4 An Example

To demonstrate the applicability of the proposed approach, an example of evaluating and selecting a suitable hotel location involving three hotel managers is presented. Based on a thorough investigation, four hotel location alternatives are identified with respect to four criteria including Geographical Location (C_1), Traffic Condition (C_2), Hotel Facilities (C_3), and Operational Convenience (C_4) [2, 3].

Geographical location (C_1) refers to the subjective assessment of the DM on the location of the hotel for achieving the competitive advantage. It is measured by the proximity of the location to public facilities, the distance to existing competitors, the security around the location, the natural resources available, and the nearby rest facilities. Using the pairwise comparison, a fuzzy reciprocal judgment matrix for the performance of hotel locations on criterion C_1 by each DM can be determined as

$$\begin{aligned}
 C_{1,D_1} &= \begin{matrix} & A_1 & A_2 & A_3 & A_4 \\ A_1 & \tilde{1} & \tilde{5} & \tilde{5} & \tilde{3} \\ A_2 & \tilde{5}^{-1} & \tilde{1} & \tilde{9} & \tilde{3} \\ A_3 & \tilde{5}^{-1} & \tilde{9}^{-1} & \tilde{1} & \tilde{3}^{-1} \\ A_4 & \tilde{3}^{-1} & \tilde{3}^{-1} & \tilde{3} & \tilde{1} \end{matrix}, & C_{1,D_2} &= \begin{matrix} & A_1 & A_2 & A_3 & A_4 \\ A_1 & \tilde{1} & \tilde{3}^{-1} & \tilde{7}^{-1} & \tilde{5}^{-1} \\ A_2 & \tilde{3} & \tilde{1} & \tilde{9}^{-1} & \tilde{3} \\ A_3 & \tilde{7} & \tilde{9} & \tilde{1} & \tilde{3}^{-1} \\ A_4 & \tilde{5} & \tilde{3}^{-1} & \tilde{3} & \tilde{1} \end{matrix}, \\
 C_{1,D_3} &= \begin{matrix} & A_1 & A_2 & A_3 & A_4 \\ A_1 & \tilde{1} & \tilde{3} & \tilde{5} & \tilde{3}^{-1} \\ A_2 & \tilde{3}^{-1} & \tilde{1} & \tilde{9} & \tilde{5} \\ A_3 & \tilde{5}^{-1} & \tilde{9}^{-1} & \tilde{1} & \tilde{3}^{-1} \\ A_4 & \tilde{3} & \tilde{5}^{-1} & \tilde{3} & \tilde{1} \end{matrix}
 \end{aligned}$$

Traffic condition (C_2) focuses on the subjective assessment of the DM on the level of convenience of the situated hotel to various locations of interest. This is measured by the distance to airport or freeway, the distance to downtown area, the distance to scenic spots, the parking area, the convenience, and the convenience to scenic spots. Using pairwise comparison, a fuzzy reciprocal judgment matrix for the performance of alternative hotel locations on criterion C_2 for each DM can be determined as

$$\begin{aligned}
 C_{2,D_1} &= \begin{matrix} & A_1 & A_2 & A_3 & A_4 \\ A_1 & \tilde{1} & \tilde{7} & \tilde{5} & \tilde{7}^{-1} \\ A_2 & \tilde{7}^{-1} & \tilde{1} & \tilde{9} & \tilde{3} \\ A_3 & \tilde{5}^{-1} & \tilde{9}^{-1} & \tilde{1} & \tilde{3}^{-1} \\ A_4 & \tilde{7} & \tilde{3}^{-1} & \tilde{3} & \tilde{1} \end{matrix}, & C_{2,D_2} &= \begin{matrix} & A_1 & A_2 & A_3 & A_4 \\ A_1 & \tilde{1} & \tilde{3}^{-1} & \tilde{5} & \tilde{5}^{-1} \\ A_2 & \tilde{3} & \tilde{1} & \tilde{5}^{-1} & \tilde{3} \\ A_3 & \tilde{5}^{-1} & \tilde{5} & \tilde{1} & \tilde{3}^{-1} \\ A_4 & \tilde{5} & \tilde{3}^{-1} & \tilde{3} & \tilde{1} \end{matrix}, \\
 C_{2,D_3} &= \begin{matrix} & A_1 & A_2 & A_3 & A_4 \\ A_1 & \tilde{1} & \tilde{5} & \tilde{5} & \tilde{3} \\ A_2 & \tilde{5}^{-1} & \tilde{1} & \tilde{9} & \tilde{7}^{-1} \\ A_3 & \tilde{5}^{-1} & \tilde{9}^{-1} & \tilde{1} & \tilde{3}^{-1} \\ A_4 & \tilde{3}^{-1} & \tilde{7} & \tilde{3} & \tilde{1} \end{matrix}
 \end{aligned}$$

Hotel facilities (C_3) concern about the ability of the hotel to provide both facilities and services for fulfilling the requirements of the customer. This includes the indoor leisure facilities, the diversity of restaurants, the amalgamation with local culture, and the convenience of obtaining nearby land. A fuzzy judgment matrix for the performance of alternative hotel locations on C_3 for each DM can be determined as

$$\begin{aligned}
 C_3, D_1 &= \begin{matrix} & A_1 & A_2 & A_3 & A_4 \\ A_1 & \tilde{1} & \tilde{7}^{-1} & \tilde{7} & \tilde{5} \\ A_2 & \tilde{7} & \tilde{1} & \tilde{9} & \tilde{9}^{-1} \\ A_3 & \tilde{7}^{-1} & \tilde{9}^{-1} & \tilde{1} & \tilde{5}^{-1} \\ A_4 & \tilde{5}^{-1} & \tilde{9} & \tilde{5} & \tilde{1} \end{matrix}, & C_3, D_2 &= \begin{matrix} & A_1 & A_2 & A_3 & A_4 \\ A_1 & \tilde{1} & \tilde{3}^{-1} & \tilde{7}^{-1} & \tilde{7} \\ A_2 & \tilde{3} & \tilde{1} & \tilde{9}^{-1} & \tilde{3}^{-1} \\ A_3 & \tilde{7} & \tilde{9} & \tilde{1} & \tilde{7}^{-1} \\ A_4 & \tilde{7}^{-1} & \tilde{3} & \tilde{7} & \tilde{1} \end{matrix}, \\
 C_3, D_3 &= \begin{matrix} & A_1 & A_2 & A_3 & A_4 \\ A_1 & \tilde{1} & \tilde{9} & \tilde{7} & \tilde{3}^{-1} \\ A_2 & \tilde{9}^{-1} & \tilde{1} & \tilde{9} & \tilde{3} \\ A_3 & \tilde{7}^{-1} & \tilde{9}^{-1} & \tilde{1} & \tilde{3}^{-1} \\ A_4 & \tilde{3}^{-1} & \tilde{3}^{-1} & \tilde{3} & \tilde{1} \end{matrix}
 \end{aligned}$$

Operational convenience (C_4) involves with the subjective assessment of the DM on the key resources for supporting the business operations of the hotel. This is assessed from the sufficiency of human resources, the quality of manpower, and the regulation restrictions. A fuzzy reciprocal judgment matrix for the performance of alternative hotel locations in regard to criterion C_4 for each DM can be determined as

$$\begin{aligned}
 C_4, D_1 &= \begin{matrix} & A_1 & A_2 & A_3 & A_4 \\ A_1 & \tilde{1} & \tilde{5} & \tilde{9}^{-1} & \tilde{3}^{-1} \\ A_2 & \tilde{5}^{-1} & \tilde{1} & \tilde{9} & \tilde{3}^{-1} \\ A_3 & \tilde{9} & \tilde{9}^{-1} & \tilde{1} & \tilde{9}^{-1} \\ A_4 & \tilde{3}^{-1} & \tilde{3} & \tilde{9} & \tilde{1} \end{matrix}, & C_4, D_2 &= \begin{matrix} & A_1 & A_2 & A_3 & A_4 \\ A_1 & \tilde{1} & \tilde{3}^{-1} & \tilde{7}^{-1} & \tilde{7} \\ A_2 & \tilde{3} & \tilde{1} & \tilde{9}^{-1} & \tilde{7} \\ A_3 & \tilde{7} & \tilde{9} & \tilde{1} & \tilde{3}^{-1} \\ A_4 & \tilde{7} & \tilde{7}^{-1} & \tilde{3} & \tilde{1} \end{matrix}, \\
 C_4, D_3 &= \begin{matrix} & A_1 & A_2 & A_3 & A_4 \\ A_1 & \tilde{1} & \tilde{9} & \tilde{5} & \tilde{9} \\ A_2 & \tilde{9}^{-1} & \tilde{1} & \tilde{9} & \tilde{5} \\ A_3 & \tilde{5}^{-1} & \tilde{9}^{-1} & \tilde{1} & \tilde{9} \\ A_4 & \tilde{9}^{-1} & \tilde{5}^{-1} & \tilde{9}^{-1} & \tilde{1} \end{matrix}
 \end{aligned}$$

To determine the weights of the selection criteria, pairwise comparison is used based on the linguistic variables defined as in Table 172.1, resulting in the determination of a fuzzy judgment matrix for each DM as

$$\begin{aligned}
 W, D_1 &= \begin{matrix} & C_1 & C_2 & C_3 & C_4 \\ C_1 & \tilde{1} & \tilde{7} & \tilde{9} & \tilde{5} \\ C_2 & \tilde{7}^{-1} & \tilde{1} & \tilde{9} & \tilde{3} \\ C_3 & \tilde{9}^{-1} & \tilde{9}^{-1} & \tilde{1} & \tilde{3}^{-1} \\ C_4 & \tilde{5}^{-1} & \tilde{3}^{-1} & \tilde{3} & \tilde{1} \end{matrix}, & W, D_2 &= \begin{matrix} & C_1 & C_2 & C_3 & C_4 \\ C_1 & \tilde{1} & \tilde{7} & \tilde{7} & \tilde{5} \\ C_2 & \tilde{7}^{-1} & \tilde{1} & \tilde{3} & \tilde{3} \\ C_3 & \tilde{7}^{-1} & \tilde{3}^{-1} & \tilde{1} & \tilde{5}^{-1} \\ C_4 & \tilde{5}^{-1} & \tilde{3}^{-1} & \tilde{5} & \tilde{1} \end{matrix}, \\
 W, D_3 &= \begin{matrix} & C_1 & C_2 & C_3 & C_4 \\ C_1 & \tilde{1} & \tilde{3}^{-1} & \tilde{9}^{-1} & \tilde{3} \\ C_2 & \tilde{3} & \tilde{1} & \tilde{9}^{-1} & \tilde{3} \\ C_3 & \tilde{9} & \tilde{9} & \tilde{1} & \tilde{3}^{-1} \\ C_4 & \tilde{3}^{-1} & \tilde{3}^{-1} & \tilde{3} & \tilde{1} \end{matrix}
 \end{aligned}$$

Table 172.2 The consistency measure of decision makers

Decision maker	Consistency measure			
	A_1	A_2	A_3	A_4
D_1	0.73	0.65	0.72	0.77
D_2	0.81	0.69	0.64	0.71
D_3	0.74	0.73	0.67	0.70

Table 172.3 The overall performance index and ranking of hotel location alternatives

Hotel location	Index	Ranking
A_1	0.73	3
A_2	0.86	1
A_3	0.64	4
A_4	0.78	2

The proximity measure between individual DMs’ assessments and the group assessments for the performance rating and the criteria weight is calculated by (172.8) and (172.9). In this situation, the consensus threshold value is set at 0.60. By using (10), the CM for the group on the performance ratings and the criteria weights is obtained in Table 172.2. It is observed that the CM value of individual DMs on all alternatives is more than the consensus threshold. Therefore, the consensus building process is finalized.

An overall performance index for each location across all criteria is calculated by (172.11) to (172.15). Based on Table 172.3, A_2 is the most suitable location.

172.5 Conclusion

The hotel location evaluation and selection process is challenging as it involves several DMs, multiple selection criteria, numerous hotel location alternatives, the presence of subjective and imprecise assessments, and the pressure to reach a certain level of agreement among the DMs. To effectively solve this problem, this paper has presented a fuzzy multi-criteria group decision making approach for solving the hotel location evaluation and selection problem. An example is presented that shows the approach is capable of effectively addressing the hotel location selection problem.

References

1. Johnson C, Vanetti M (2005) Locational strategies of international hotel chains. *Ann Tour Res* 32:1077–1099
2. Hsieh LF, Lin LH (2010) A performance evaluation model for international tourist hotels in Taiwan-An application of the relational network DEA. *Int J Hosp Manage* 29:14
3. Chou TY, Hsu CL, Chen MC (2008) A fuzzy multi-criteria decision model for international tourist hotels location selection. *Int J Hosp Manage* 27:293–301
4. Zadeh LA (1965) Fuzzy sets. *Inf. Control* 8:338–353
5. Saaty TL (1990) *The analytical hierarchy process, planning, priority, resource allocation*. RWS Publications, Pittsburgh
6. Chang DY (1996) Applications of the extent analysis method on fuzzy AHP. *Eur J Oper Res* 95:649–655
7. Deng H (1999) Multi-criteria analysis with fuzzy pairwise comparison. *Int J Approx Reason* 21:215–231
8. Wibowo S, Deng H (2010) Risk-oriented group decision making in multi-criteria analysis. In: *Proceedings of the ninth IEEE/ACIS international conference on computer and information science*, Japan
9. Chen SH (1999) Ranking generalized fuzzy number with graded mean integration. In: *Proceedings of the eighth international fuzzy systems association world congress*, Taiwan
10. Hsu HM, Chen CT (1996) Aggregation of fuzzy opinions under group decision making. *Fuzzy Sets Syst* 79:279–285
11. Lee HS (1999) An optimal aggregation method for fuzzy opinions of group decision. In: *Proceedings of the IEEE conference on systems, man, and cybernetics*, Japan
12. Hsieh CH, Chen SH (1999) Similarity of generalized fuzzy numbers with graded mean integration representation. In: *Proceedings of the eighth international fuzzy systems association world congress*, Taiwan
13. Chen SH (1985) Ranking fuzzy numbers with maximising set and minimising set. *Fuzzy Sets Syst* 17:113–129
14. Klir GR, Yuan B (1995) *Fuzzy sets and fuzzy logic theory and applications*. Prentice-Hall, Englewood Cliffs

Chapter 173

Using Online Self-Adaptive Clustering to Group Web Documents

Noel Catterall

Abstract In this chapter an approach that is using online self-adaptive, incremental (on-line) clustering to automatically group relevant Web-based documents is proposed. The proposed online self-adaptive classifier has learning capability—it improves the result on-line with any new document that has been accessed. Therefore, the proposed approach is characterized by low complexity. This chapter reports the results of research on development of a novel clustering method that is suitable for real-time implementations. It is based on evolution principles and tries to address the limitations of existing clustering algorithms which cannot cope in an online mode with high dimensional datasets. This evolution-inspired and nature-inspired approach introduces the new concept of potential values which describes the fitness of a new sample (web document) to be the prototype of a new cluster without the need to store each previously encountered documents but taking into account the contextual similarity density between all previous documents in a recursive and thus computationally efficient way. This chapter also examines the clustering of documents by contextual similarity using extracted keywords represented in a vector space model.

Keywords Clustering · Information retrieval · Self-adaptive on-line clustering · Contextual similarity

N. Catterall (✉)
SYM Global Limited, Manchester, M1 2AT, UK
e-mail: noel@symblobal.com

173.1 Introduction

Dramatic developments in Internet and information systems lead to the situation when we are ‘drowning in information and starving for knowledge’ according to the expression by Roger. Despite the existence of huge amount of information on the Web its use is still hampered by the fact that it is disorganized and often irrelevant to our aims and objectives. One approach to cope with this problem is the Semantic Web, but it requires a bottom-up work on rebuilding the monstrous amount of web pages that already exists. An alternative is to ‘get the best out of the current situation’. The on-line clustering method developed in this chapter can be considered as a technique associated to the latter approach when we try to extract useful knowledge of the data stream represented by the Web documents [1, 2].

Cluster analysis involves a number of data analysis algorithms and techniques for grouping similar objects into categories so that those within a cluster have greater similarity to each other [3]. It has been applied in Information Retrieval applications such as SMART [4]. Contextual clustering is the partitioning of a dataset of contextually unordered documents into subsets so that data in each subset have similar traits [5]. Before a clustering algorithm can be applied, the contextually important keywords need to be extracted by means of regular expressions and noise or irrelevant words need to be filtered by means of a stop list [6]. The document is then transformed into a numerical representation using the vector space model approach [7]. Frequency information is used in the Euclidean, Levenstein [8], or ‘cosine’ distance measures [9] to apply weights to words so that a word that occurs frequently in all documents is given a lower weight than words that appear frequently in one document only. Normalisation of the sum of weighted words is applied so that very large documents do not skew the results of small documents.

Recently, approaches for on-line clustering textual documents [10, 11] have been developed. In the approach that is proposed in this chapter contextual similarity to all documents is used without storing them or the information about each one of them.

The software realization of the proposed approach (SmartSearch, available at www.onlineclustering.com) is intended to be an add-on to existing search engines and this is achieved by using web services which work using requests and response messages using the Remote Procedure Call (RPC) pattern. Web Services allow programmers to create add-ons to existing web based applications in a programming language independent manner. The “Google API” and “Alexa Web Information Service” were also used.

173.2 Paper Preparation

A novel approach to clustering the data streams in real-time was recently developed [12, 13]. In this chapter we develop further this approach and apply it to clustering Web-based documents in on-line mode with potential to be

implemented in real-time. The proposed online self-adaptive clustering method has the following specific features that make it suitable for this task:

High dimensionality: The algorithm is able to cope with thousands of samples (Web-based documents) without the need to store all the information in memory.

Unsupervised: The algorithm is able to divide a dataset into subsets without requiring initial training data.

Online: The algorithm is suitable for continuously clustering new data due to its recursive nature.

The online self-adaptive Clustering (e-Clustering) method is a prototype-based [12, 14]. Fuzzy rules are generated on-line around prototype Web Documents, D . A vector containing the frequencies, F of occurrence of each 'keyword' in the meaningful content of the Document frequency list, L forms the input to e-Clustering. Note that not only the frequency of each keyword, but also the list of keywords for each web document can be different.

$$L_k = [F_1^k, F_2^k, \dots, F_n^k] \quad (173.1)$$

The rules generated by e-Clustering are of the following form:

$$R^i : IF(L^k \text{ is similar to } L^{i*}) THEN (Group D^k \text{ together with } D^{i*}) \quad (173.2)$$

where R^i denotes the i th fuzzy rule; $i = [1, N]$; N is the number of fuzzy rules.

Fuzzy membership functions of Gaussian type are formed for each 'keyword' of the prototypes, L^{i*} , $[F_1^*, F_2^*, \dots, F_n^*]$.

$$\tau_k^i = e^{-\sum_{j=1}^n \left(\frac{d_{\cos ik}^j}{\sigma_j^i} \right)^2} \quad i = [1, N]. \quad (173.3)$$

where $d_{\cos ik}$ is the dissimilarity between a web document and the prototype of the i th group (the focal point of the fuzzy rule); σ^i is the spread of the membership function. We use so called 'cosine distance' as a measure for dissimilarity between the frequency lists, L [3]:

$$d_{\cos} = 1 - \frac{\sum_{j=1}^n F_j^k F_j^i}{\sqrt{\sum_{j=1}^n (F_j^k)^2 \sum_{j=1}^n (F_j^i)^2}} \quad (173.4)$$

e-Clustering assumes that the number of content categories is not pre specified, and the number of fuzzy rules evolves during the process of online clustering. *e-Clustering* is centered on the concept of data spatial density measured by so called 'potential' [12, 14]. Each web document is considered a potential cluster centre. A document surrounded by documents will have a higher potential value. Potential is inversely proportional to the dissimilarity between a document and all the previous documents [12, 14]:

$$P_k(L_k) = \frac{1}{1 + \left(\sum_{i=1}^{k-1} d_{\cos}^2(L_k, L_i)\right)/(k-1)}; \quad k = 1, 2, \dots \quad (173.5)$$

where k is the index of the current document.

To enable the online one-pass ability of the approach, the processed web documents will only be processed once and discarded immediately when processed. This requires all the calculations for the leaning procedure to be recursive to accumulate the history information without memorizing the history data. An online clustering procedure starts with the first document established as a prototype for the first cluster. Its frequency list is used to form the first fuzzy rule. Starting from the next data document onwards the potential of the new documents is calculated recursively by [14]:

$$P_k(L_k) = \frac{1}{2 - \frac{1}{\sqrt{\sum_{i=1}^n (F_k^i)^2}} \sum_{i=1}^n F_k^i a_{k-1}^i}; \quad P_1(L_1) = 1 \quad (173.6)$$

Where

$$a_k^i = a_{k-1}^i + \sqrt{\frac{(F_k^i)^2}{\sum_{l=1}^n (F_k^l)^2}}; \quad a_1^i = \sqrt{\frac{(F_1^i)^2}{\sum_{l=1}^n (F_1^l)^2}}; \quad i = [1, n] \quad (173.7)$$

In this way, the spatial density at each new document, P_k in respect to all previous documents can be recursively calculated using n accumulated values in a single auxiliary variable. This makes possible learning the information of contextual density representing the whole previous history which is the distinctive feature of the proposed algorithm. Density in the global data space is affected whenever a new web document enters the space; therefore the potentials of all existing prototype keyword frequency vector needs to be updated. This update is also done in a recursive way:

$$P_k(L^*) = \frac{(k-1)P_{k-1}(L^*)}{k-2 + P_{k-1}(L^*) + P_{k-1}(L^*)d(L^*, L_k)} \quad (173.8)$$

Updating of (173.8) is only based on the current inputs and the current potential $P_{k-1}(L^*)$, therefore, no additional variable needs to be stored/memorized.

The proposed online self-adaptive web documents clustering approach starts ‘from scratch’, without any pre-defined prototype. Frequency vector of each document are used to upgrade the rule-base. Potential of each document, $P_k(L_k)$ is updated recursively by (173.6) and (173.7) from the second document received and onwards. The potential of all existing prototypes, $P_k(L^*)$ is also updated using (173.8). Comparing the potential of a new document with the potential of each of the existing prototypes:

$$P_k(L_k) < \min(P_k(L^{i*})) \quad (173.9)$$

If condition (173.9) occurs that means the new document is very distinctive and potentially a new group can be formed around it as prototype. We also check whether any of the already existing prototypes is similar enough to (also means well described by) the new prototype. The degree of the membership is therefore tested:

$$\exists i, i = [1, N]; \quad \tau_k^j > 1/3 \quad \forall j, j = [1, n] \quad (173.10)$$

If both (173.9) and (173.10) stands, which means the new web documents is a candidate prototype and very similar to the existing prototype web document, therefore, it can replace this similar prototype; otherwise, if only (173.9) is true, a new group is formed with the new web document as the prototype.

The procedure of e-Clustering can be described by the following pseudo code:

BEGIN eClass

Initialize (*get the first document, D₁; define the keyword frequencies, F₁ and generate its list L₁; initialize the rule base and parameters*);

DO for any next document, F_k, k = 2, 3, ...

Pre-process the document, Extract frequency list, L_k;

/---Evolve Rule-base---*/*

Calculate P_k(L_k) using (6) - (7);

Update P(L*) using (8)

Compare the potential of new document and that of the existing prototypes (9)

IF ((9)) **THEN add** a new prototype formed around the new document, D_k;

IF ((10)) **THEN remove** the prototype(s), L_{i*} for which this holds;

WHILE More documents are available

END DO

173.3 Experimental Study

We have performed extensive experiments and development of the software application, *SmartSearch* (Fig. 173.1) available at www.onlineclustering.com. In order to compare the results we have used four other previously existing approaches to clustering Web-based documents, namely:

- Offline matrix based approach (clumping) comparing all pair wise distances with Euclidean distance.

- Online agglomerative approach with Cosine distance measures.

- Offline kNN where K = 4, using Euclidean distance measures.

Fig. 173.1 The online entrance of SmartSearch



Table 173.1 Search term = Lancaster

	Clumping	Agglomerative	<i>k</i> NN (k = 4)	<i>e</i> -Clustering	Subjective clustering
No of clusters	21	15	4	6	5
Online/Offline	Offline	Online	Offline	Online	Offline
Total time, s	~ 30	~ 3	~ 1 per iteration	0.75	> 1000
Threshold	0.5	0.80	NO	NO	NO
Notes/Clusters formed	Significant cluster overlap on all clusters.	Universities, Businesses City of Lancaster Pennsylvania, USA (10 noise clusters)	University, Lancaster city, Businesses, Newspapers	Governments, Educations, Tourism, Businesses	Tourism Local government, Companies, America, News & University

Subjective clustering.

The software application SmartSearch is a server side PHP application with a web browser based front end. The user is able to choose between the matrix, kNN, agglomerative and online self-adaptive methods to cluster web documents. The tests were based around three search terms; “Lancaster” (Table 173.1), “Ball” (Table 173.2) and “Cold” (Table 173.3). The success of the results was determined by the total time taken from submitting the query to receiving the results, and the minimum number of clusters formed. Manual subjective clustering was also compared. The results are given in the tables below for grouping 30 documents for each query.

Table 173.2 Search term = Ball

	Clumping	Agglomerative	<i>kNN</i> (k = 4)	<i>e-Clustering</i>	Subjective clustering
No of clusters	15	15	4	5	4
Online/Offline	Offline	Online	Offline	Online	Offline
Total time, s	~ 30	~ 3	~ 1 per iteration	0.73	> 1000
Threshold	0.5	0.90	NO	NO	NO
Notes/Clusters formed	Significant cluster overlap on all clusters.	Games, News, People (11 noise clusters)	Ball games, Companies, News, People.	Ball Games Company University	Companies University People Games.

Table 173.3 Search term = Cold

	Clumping	Agglomerative	<i>kNN</i> (k = 4)	<i>e-Clustering</i>	Subjective clustering
No of clusters	13	15	4	2	5
Online/Offline	Offline	Online	Offline	Online	Offline
Total time, s	~ 30	~ 3	~ 1 per iteration	0.42	> 1000
Threshold	0.5	0.85	NO	NO	NO
Notes/Clusters formed	Cold war Music Cold fusion (10 noisy clusters)	Cold war, common cold (13 noise clusters)	Common cold, Cold fusion, Weather, Cold war	Common cold Cold war	Cold fusion Cold weather Common cold Cold war Music

173.4 Results Analysis and Conclusions

The matrix based approach was the least efficient technique and its use of un-weighted keywords resulted in significant cluster overlap. In the agglomerative method, the order in which the documents were presented affected the final clustering and many atomic (noise) clusters were formed. The *kNN* approach required all the training data to be available beforehand. Finally, clumping and agglomerative clustering methods required a predefined threshold value which problem specific.

To automatically group an unknown number of the web documents online, without predefining the clusters, a novel approach called *e-Clustering* has been developed and implemented as a software on-line application called *SmartSearch*. The proposed approach is a fully unsupervised fuzzy clustering approach that

process the web documents incrementally (in one pass), which avoids memorizing the history data and thus makes possible to apply the approach in on-line. Additionally, the length of the feature vectors can be different using ‘cosine distance’, which is not the case with Euclidean distance.

The proposed novel approach will revolutionise the next generation of search engines and will also be useful in data analysis to automatically discover the nature or structure of the data. Application to clustering desk-based documents, e-mail messages etc. is also under intensive consideration and will constitute a further step of investigations.

References

1. Fayyad G, Shapiro P, Smyth P (1996) From data mining to knowledge discovery: an overview. *Advances in knowledge discovery and data mining*, MIT Press, Cambridge
2. Domingos P, Hulten G (2001) Catching up with the data: Research issues in mining data streams, Workshop on research issues in data mining and knowledge discovery, Santa Barbara, CA
3. Duda RO, Hart PE, Stork DG (2000) *Pattern classification*, 2nd edn. Wiley, Chichester
4. Salton G, Lesk ME (1965) The smart automatic retrieval system an illustration. *ACM* 8:6
5. Runkler TA, Bezdek JC (2003) Web mining with relational clustering. *Int J Approx Reason* 32(2–3):217–236
6. Fox CJ (1990) A stop list for general text. *SIGIR Forum* 24(1–2):19–35
7. Salton G, Wong A, Yang CS (1975) A vector space model for automatic indexing. *ACM* 18(11):613–620
8. Angelov P, Evans T (2004) Semantic categorization of web-based documents. In: *Proceedings of 5th international conference recent advances in soft computing (RASC)*, Nottingham, UK, pp 500–505
9. Yang B, Song W (2005) A SOM-based web text clustering approach. In: *Proceedings of IFSA*, pp 618–621
10. Suryavanshi BS, Shiri N, Mudur SP (2005) Incremental relational fuzzy subtractive clustering for dynamic web usage profiling. In: *Proceedings of WebKDD*, Chicago
11. Chai KMA, Ng HT, Chieu HL (2002) Bayesian online classifiers for text classification and filtering. In: *Proceedings of SIGIR’02*, 11–15 Aug, Tampere, Finland, pp 97–104
12. Angelov P (2004) An approach for fuzzy rule-base adaptation using on-line clustering. *Int. J Approx Reason* 35(3):275–289
13. Angelov P, Zhou X (2006) Evolving fuzzy systems from data streams in real-time. 2006 international symposium on evolving fuzzy systems. IEEE Press, Ambleside, pp 26–32
14. Angelov P, Zhou X, Klawonn F (2007) Evolving fuzzy rule-based classifiers. In: *IEEE international conference on computational intelligence applications for signal and image processing*, 1–5 April, Honolulu, Hawaii, USA

Chapter 174

An Order-Based Taxonomy for Text Similarity

Yi Feng

Abstract Text similarity is a common and basic issue to consider in many fields. This paper proposes a new order-based taxonomy for text similarity. Based on the different consideration on the order of text comparison unit, we classify text similarities into three categories: order-sensitive similarity, order-insensitive similarity, and order-semi-sensitive similarity. For order-sensitive similarity, each text is considered as a string of items, and text matching is carried out item by item as a pairwise alignment process. For order-insensitive similarity, each text is considered as a set of distinct items, and only item co-occurrence is considered during comparison. For order-semi-sensitive similarity, block of items with dynamically determined length is used as comparison unit, and only local order (the item order within each block) is preserved during matching. The taxonomy presented in this paper provides us an insight into the text similarity issue in an order perspective, which could be beneficial in understanding and developing this basic element in many disciplines.

Keywords Text similarity · Text similarity taxonomy · Order-sensitive similarity · Order-insensitive similarity · Order-semi-sensitive similarity · Text matching · Information retrieval

Y. Feng (✉)
College of Computer Science and Information Engineering,
Zhejiang Gongshang University, Hangzhou 310018, China
e-mail: yfeng@mail.zjgsu.edu.cn

174.1 Introduction

In the current era of information explosion, textual materials are rapidly generated, stored and updated, both on and off the web. How to effectively organize and utilize such tremendous volume of textual information has attracted intensive interest in many related fields, such as text classification, text summarization, information retrieval, information extraction, natural language processing, etc.

Among these applications, text similarity has become a common and basic issue to consider. One typical application of text similarity is the vector space model in information retrieval, where the retrieved documents for a given query are determined by ranking candidate documents in reversed order of their similarity to the input query [1, 2]. Further Extensions of traditional information extraction have also involved text similarity in their processing, such as query expansion (query–query similarity needed) [3], question answering (question–answer sentence similarity needed) [4], and sponsored search (query–ad similarity needed) [5]. The classification of textual materials is also conducted using the text representation model based on text similarity, particularly the vector space model [2]. Text similarity has also been widely used as basic elements in text summarization [6], word sense disambiguation [7], text copy detection [8], and so on.

As a common and crucial issue in such various applications, the topic of text similarity is worthy of careful study in a systematic framework. However, seldom taxonomy for text similarity has been established up to now. As an exploring work in this aspect, this paper presents a new taxonomy of text similarity based on the perspective of order. According to the different consideration on the order of text comparison unit (we call such comparison unit as item, typically word) within the text, we classify text similarities into three categories: order-sensitive similarity, order-insensitive similarity, and order-semi-sensitive similarity. The idea of order perspective of text similarity is firstly briefly mentioned in [9]. In this paper, this idea is extended and presented as a more systematic and detailed taxonomy for text similarity.

The rest of the paper is organized as below. A basic order-based taxonomy of text similarity is first defined and proposed in Sect. 2. Under such taxonomy, the three categories of text similarity are then presented and discussed one by one, with order-sensitive similarity in Sect. 3, order-insensitive similarity in Sect. 4, and order-semi-sensitive similarity in Sect. 5. Finally a conclusion is provided as Sect. 6.

174.2 An Order-based Taxonomy for Text Similarity

In this paper, we present an order-based taxonomy for text similarity. To measure the closeness between two texts, we could consider this problem in three categories, based on the role of item order within text on comparison. We define an

item as the basic unit of text comparison. The granularity of item can be determined according to language and application. For languages separated by space such as English, French and German, each word, phrase (need phrase segmentation), or character (seldom) can be considered as one item, according to application. For languages that need word segmentation such as Chinese, Korean and Japanese, we could treat each character as one item. If word segmentation is effectively carried out, however, we could also extend the granularity of item to word level. Besides, some preprocessing procedures could be carried out to achieve better comparison result, depending on language and application. Take English, for example. Typical preprocessing steps include stop-word elimination and stemming. For texts rich of synonymy and polysemy, synonymy replacement and word sense disambiguation (WSD) might be helpful. After these preprocessing steps, three categories of text similarity could be calculated, as presented below.

174.3 Category 1: Order-Sensitive Similarity

In this category, text is usually considered as string, more precisely, string of items. To derive string similarity, a process named string alignment is carried out. As the literal meaning of string alignment implies, the item order within text is strictly preserved during text matching in this category. That is, matching is carried out item by item. Text similarity derived from this procedure of pairwise alignment can be classified into the category of order-sensitive similarity. Although its close relationship with traditional string matching, order-sensitive comparison might differ from it in the granularity of comparison. Basically, traditional string matching compares strings at character level, whereas text comparison can also be carried out at other item level, such as word level. Similarities based on edit distance and its most variants can fall into this category of text similarity.

Levenshtein [10] firstly presented edit distance to measure dissimilarity of strings, which is based on the idea of dynamic programming. This distance allows insertions, deletions and substitutions of items during string matching. The number of operations need to change one string to another is treated as the dissimilarity of two strings. Hamming distance was firstly proposed by Sankoff et al. [11], which allows only substitutions. Das et al. [12] presented episode distance in data mining background. Only insertion operation is allowed during episode matching. Needleman et al. [13] firstly proposed longest common subsequence distance (LCS). Deletion of zero or more symbols from each of the two given string is carried out to obtain LCS. Many other variants of edit distance exist to improve its time complexity and space complexity. Most of above distances used in string alignment mainly focus on character in string. However, we could easily extend these similarities to different item level, for example, word level. All of these comparisons share the same feature: the order of each item in text is preserved during comparison.

174.4 Category 2: Order-Insensitive Similarity

The comparison in category two does not focus on matching at string level. Instead, a text is treated as a set of distinct items. Neglecting item order within text, only item co-occurrence is considered during comparison. Such kind of text similarity can be termed as order-insensitive one.

Two subcategories can be defined within this group, based on the characteristic of text. The first subcategory of order-insensitive similarity is applicable to texts with few item duplications. To simplify the comparison between two texts here, only two situations are considered due to limited duplications: for each item within one text, it might does not occur in another text, or it occurs in both texts. To evaluate the equivalence between two texts in this subgroup, we count the number of co-occurrence and unilateral occurrence. Then some mathematic coefficients, such as Jaccard coefficient [14], can be calculated from these numbers to measure the degree of overlap between two sets of items. This kind of coefficient is usually utilized as similarity for texts with short length. Typically, when the number of distinct items reaches to a considerable figure, some models, such as vector space model (VSM), could be applied to reduce the time and space complexity. In VSM text is represented by a vector of terms [2]. Terms are typically words and phrases in VSM, corresponding to the comparison level of word and phrase in our definition of item, respectively. Due to few item duplications, binary vector is enough to evaluate the closeness of two texts. The angle between two vectors (texts) could be used as a measure of divergence between the vectors, and cosine of the angle is used as the numeric similarity.

The similarities used to measure the closeness of texts that are rich of item duplications constitute the second subcategory. Here, the objects to compare are always texts of large length and typically documents. Thus, considering whether an item occurs in both texts or not is not enough, since this cannot distinguish between texts consisted of same items but with different item frequency. To effectively discriminate texts in this background, term frequency (or tf) is incorporated into VSM. Moreover, some scoring schemes combining VSM are proposed to measure the similarity between documents. Among them, Okapi weighting presented by Robertson is a method widely accepted in information retrieval, based on tf and document frequency (or df) [2]. Motivated by Robertson's work, Researches at Cornell University developed pivoted normalization weighting which considers not only tf, df, but also document length (or dl) [2]. These two methods and other variants using both tf and df are generally termed tf-idf weighting and are widely applied in text mining, document query, text classification/clustering, document copy detection, and so on.

The common characteristic residing in these two subcategories can be summarized as below: the sequence in which the items occur is disregarded during text comparison, according with the literal meaning of order-insensitive.

174.5 Category 3: Order-Semi-Sensitive Similarity

In this category, text is still treated as string of items. Thus, this group of similarity differs from order-insensitive ones on the partial consideration of item sequence within text, on which point it is similar with order-sensitive ones. But where does the prefix “semi” come from? In definition, a similarity belongs to order-semi-sensitive one only if it satisfies the following three conditions simultaneously:

(1) *Block of items is used as comparison unit.*

The distinction between order-sensitive and order-semi-sensitive similarities lies firstly in the granularity of comparison. The former compare texts at single item level, whereas the granularity of order-semi-sensitive comparisons extends to block of items. That is, we regard a block of continuous matching of items as comparison unit. Note that, if we only larger the granularity of certain order-sensitive comparison, we do capture continuous matching in sense of original item level, then how can we distinguish this level of order-sensitive comparison from order-semi-sensitive one? This ambiguity can be removed by the second condition presented below.

(2) *Only local order is preserved during matching.*

During order-sensitive comparison, the global sequence wherein each item occurs is preserved. However, during order-semi-sensitive one, only the item order within each block (local order) is kept, while the position of this block in text is totally ignored. Based on the feature of local order, we can eliminate above ambiguity: by increasing the item level of order-sensitive comparison, we do capture continuous matching in sense of original item level, but the sequence wherein matching blocks occur (global order) is still important on calculation of order-sensitive similarity. Thus, the order-semi-sensitive measure is not simply a version of order-sensitive one that increases the item level.

(3) *The size of each item block is dynamically determined.*

This feature distinguishes order-sensitive similarities from phrase-based VSM[15] which belongs to the category of order-insensitive. In phrase-based VSM, we treat each phrase, i.e. a string of words, as comparison unit, and we also do not care the position of each phrase in a given text. However, in phrase-based VSM, each phrase is determined beforehand by some knowledge sources, such as UMLS in [15]. That is, the set of phrase is static during one run of text comparison. This method exhibits desirable efficiency for texts of large length, but is restricted by the completeness of knowledge source. When the texts to compare are of short or medium length, we could allow the size of item block to dynamically change, trading some efficiency for more effective usage of sequence information. Moreover, in cases lack of knowledge sources or many applications with dissatisfactory knowledge sources, phrase segmentation becomes a difficult problem. Thus, the size of each item block could be dynamically determined as a solution.

Since global order could be disregarded, to allow matching at different offsets within texts, one straightforward idea is to introduce swapping operation into text matching. This idea helps to solve the problem brought by the example in category one. The order within swapping unit is preserved, while the global order is not kept during comparison. Lowrance and Wagner [16] firstly extended edit operations to include swap operation. They also presented polynomial-time algorithms [17] for some restricted cases of extended string-to-string correction problem (ESSCP) where the swap edit operation is added. Amir et al. [18] formulated this problem as “pattern matching with swap”. However, the swap operation in [16–18] allows only two adjacent characters to interchange, which is still at single item level. This version of swap operation is termed “local swap” by Amir et al. [18]. Tichy [19] firstly applied block-moving approach to handle string-to-string correction problem, which extended to allow blocks to swap. Lopresti et al. [20] presented block edit models based on block edit distance for approximate string matching.

Although differing in allowed type of operations, almost all of these distances mentioned above are rooted on the idea of edit distance. The less edit operations to take from one string to another, the more similar for these two strings. This idea is reasonable and can adapt to many real-world applications. However, in a large number of real-life situations, such as text categorization, the methods based on edit operations seem a little bit complicated and not straightforward. As an alternative approach, Feng et al. [9] presented another order-semi-sensitive similarity M-similarity, which considers single item matching, maximum sequence matching and potential matching.

174.6 Conclusion

Text similarity is a common and basic issue to consider in many fields. This paper proposes a new order-based taxonomy for text similarity. Based on the different consideration on the order of text comparison unit, we classify text similarities into three categories: order-sensitive similarity, order-insensitive similarity, and order-semi-sensitive similarity. For order-sensitive similarity, each text is usually considered as a string of items, and text matching is carried out item by item as a pairwise alignment process, with swap-based matching not allowed. For order-insensitive similarity, each text is considered as a set of distinct items, and only item co-occurrence is considered during comparison, neglecting item order within text. For order-semi-sensitive similarity, block of items with dynamically determined length is used as comparison unit, and only local order (the item order within each block) is preserved during matching. Thus, text comparison using this kind of similarity is a process of swap-allowed dynamic block matching. The taxonomy presented in this paper provides us an insight into the text similarity issue in an order perspective, which could be beneficial in understanding and developing this basic element in many disciplines.

Acknowledgments This work is supported by National Natural Science Fund of China (No. 60905026, 71071141), Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20093326120004), Natural Science Fund of Zhejiang Province (No. Y1091164, Z1091224), and Zhejiang Science and Technology Plan Project (No. 2010C33016).

References

1. Salton G, Lesk M (1971) Computer evaluation of indexing and text processing. Prentice Hall, Englewood Cliffs, pp 143–180
2. Singhal A (2001) Modern information retrieval: a brief overview. *Bullet IEEE Comput Soc Tech Committee Data Eng* 24(4):35–43
3. Sahami M, Heilman T (2006) A web-based kernel function for measuring the similarity of short text snippets. In: *Proceedings of WWW*, pp 377–386
4. Murdock V, Croft WB (2005) A translation model for sentence retrieval. In: *Proceedings of HLT/EMNLP 2005*, pp 684–691
5. Hillard D, Schroedl S, Manavoglu E, Raghavan H, Leggetter C (2010) Improving ad relevance in sponsored search. In: *Proceedings of WSDM 2010*, pp 361–369
6. Lin C, Hovy E (2003) Automatic evaluation of summaries using N-gram co-occurrence statistics. In: *Proceedings of HLT/NAACL 2003*, pp 150–156
7. Schutze H (1998) Automatic word sense discrimination. *Computat Linguist* 24(1):97–124
8. Brin S, Davis J, Garcia-Molina H (1995) Copy detection mechanisms for digital documents. In: *Proceedings of ACM SIGMOD 1995*
9. Feng Y, Wu ZH, Zhou ZM (2005) Combining an order-semi-sensitive text similarity and closest fit approach to textual missing values in knowledge discovery. In: *Proceedings of KES 2005*, pp 943–949
10. Levenshtein VI (1965) Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR* 163(4):845–848
11. Sankoff D, Kruskal JB (1983) Time warps, string edits, and macromolecules: the theory and practice of sequence comparison. Addison-Wesley, Reading, MA
12. Das G, Fleischer R, Gasieniec L, Gunopulos D, Karkkainen J (1997) Episode matching. In: *Proceedings of CPM 1997, LNCS*, vol 1264. Springer, Berlin, pp 12–27
13. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48:444–453
14. Jaccard P (1908) *Nouvelles Recherches Sur La Distribution Florale*. *Bull Soc Vaud Sci Nat* 44:223–270
15. Mao W, Chu W (2002) Free-text medical document retrieval via phrase-based vector space model. In: *Proceedings of AMIA Annual Symposium 2002*
16. Lowrance R, Wagner R (1975) An extension of the string-to-string correction problem. *J ACM* 22(2):177–183
17. Wagner R (1975) On the complexity of the extended string-to-string correction problem. In: *Proceedings of seventh annual ACM symposium on theory of computing* pp 218–223
18. Amir A, Aumann Y, Landau GM, Lewenstein M, Lewenstein N (2000) Pattern matching with swaps. *J Algorithms* 37(2):247–266
19. Tichy F (1984) The string-to-string correction problem with block moves. *ACM Trans Comput Syst* 2(4):309–321
20. Lopresti D, Tomkins A (1997) Block edit models for approximate string matching. *Theor Comput Sci* 181(1):159–179

Chapter 175

Distributed Intrusion Detection System Using Autonomous Agents Based on DCOM

Yujia GE, Xiaofei Feng and Aijun Zhang

Abstract This chapter reports on the design and implementation of a Distributed intrusion detection system using autonomous agents based on Distributed Component Object Model (DCOM). First, we introduce our design of the architecture. It is followed by the implementation of the architecture on the DCOM. A login fail example is given by using the system. We also discuss the advantages and limitations of the architecture.

Keywords Intrusion detection · Autonomous agents · Distributed architecture

175.1 Introduction

Intrusion detection is to detect inappropriate incorrect or anomalous activity within security space. Intrusion detection system (IDS) is used to determine whether the systems have experienced these kinds of intrusion. Many researches are going on intrusion detection for improving the technologies to protect the systems from hackers and viruses. Algorithms for building detection models are broadly classified into two categories, Misuse Detection and Anomaly Detection [1].

Y. GE (✉) · X. Feng · A. Zhang
College of Computer and Information Engineering, Zhejiang Gongshang University,
Hangzhou, China

e-mail: yge@mail.zjgsu.edu.cn

X. Feng

e-mail: xffeng@mail.zjgsu.edu.cn

A. Zhang

e-mail: zaj@mail.zjgsu.edu.cn

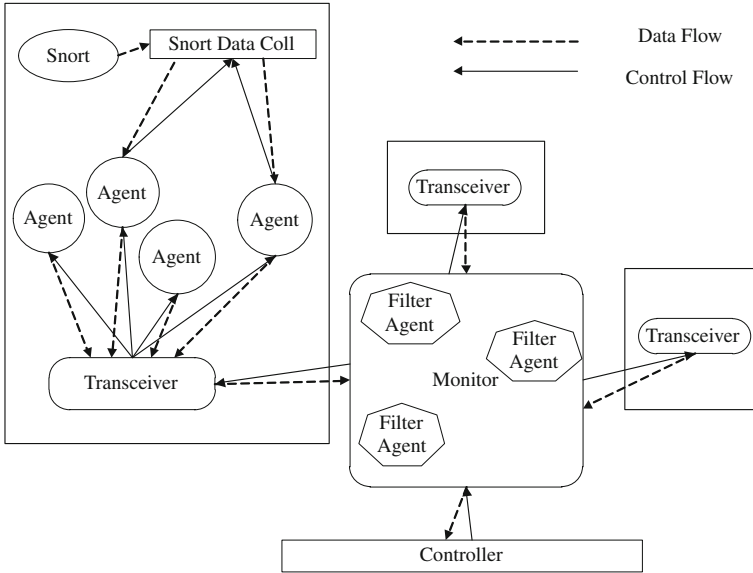


Fig. 175.1 Components in our general architecture

Distributed Intrusion Detection systems are widely used because they allow detection software modules to be placed throughout the network with a central controller collecting and analyzing the data from all the modules. This provides a robust mechanism for detecting intrusions across several subnets and several hosts. There are several methods for building intrusion detection system, including the distributed ID systems without agents, such as Distributed Intrusion Detection System (DIDS), Collaborative Intrusion Detection System (CIDF) and agent-based ID system, such as AAFID2 [2], MAIDS [3].

Our goal of the project is to implement an agent-based distributed intrusion detection system which has layered architecture. The requirements of the system include: extensibility and flexibility, reliability and efficiency, scalability and concurrency.

175.2 System Architecture

With references to the architecture proposed in “An Architecture for Intrusion Detection using Autonomous Agents” [2], the agent-based system has the following architecture. The main idea to choose this architecture is for scalability and flexibility as supposed in AAFID.

Figure 175.1 shows the components and the communication of the whole architecture respectively. There are five following types of components.

Snort. Snort is a lightweight intrusion detection system that can log packets coming across your network. It runs quite well on smaller networks and it is free. Snort helps to generate alert data by rules and store it to the local database. Snort is installed on every host on the network as alert data source in our project.

Agents. An agent is an entity which runs on some host independently and reports abnormal behavior to the appropriate control components. There are different types of agents who are interested in certain events. For example, an agent is interested in “login error” event.

Basically, there are two categories of agents:

1. Agents who collect data from database

Agents who communicate with the monitor on central server: The task of this category of agents is to send the alerts to the central monitor because the alerts generated by Snort do not need to be correlated with other alerts.

Agents who communicate with the monitor on local host: The task of this category of agents is to send the alerts to the local monitor because the alerts it concerns still need to be correlated with other alerts through transceiver.

2. Filter agents who do the correlation

Transceivers. Transceivers deal with the task of communications for each host: manage agents, communicate with the monitors, process the information sent by agents, distribute information.

Monitors. Monitors receive the processed data from transceiver. They communicate with transceivers, process the information sent by transceivers and do high-level correlation and distribute information.

Controller. Controller is the central console to deal with the highest-level task:

1. Interact with users and deal with visualization
2. Collect all the data correlated by monitors

175.3 Agent Design with Alert Correlation

Because agents are independent running entities, it can help to improve the performance in the intrusion detection environment. Actually all the analysis is done by those agents.

Usually intrusion has two categories which include Misuse and Anomalous Behavior types.

1. Misuse behavior: existence of changes in system, regular expression patterns of events, sequential patterns of events, patterns that require embedded negation and generalized selection
2. Anomaly behavior: activity intensity rate, audit record distribution of particular activities, categorical distribution of activities, ordinal measurable activities

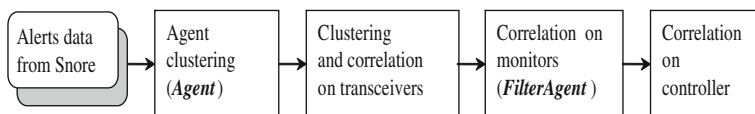


Fig. 175.2 Alert correlation hierarchy

Because we use layered architectures, alert level can be divided to different levels. We use agents to complete every specific correlation in different levels. There has been several comprehensive researches on alert coloration [4]. In our architecture, there will be 3 levels' correlation as illustrated in Fig. 175.2. In the transceiver, the task of correlation agents is to correlate the alerts from the local host. In the monitors, the task of correlation agents is to correlate the alerts from the hosts which the monitors are managing. In the central controller, the task of correlation agents is to do the global correlation for the whole network. This is the highest level.

The detection of certain attacks requires information from multiples sources. In our prototype, we focus on the design of Filter Agent which is on the monitor level correlation. A simple example is that the intruder's goal is to get the password of a certain user id. The scenario is that an intruder may telnet a target machine and try several user id and password several times in a short period of time continuously. Several different machines can be used to accomplish the attack. And it is also possible that the attack is to a bunch of hosts. The "login error" activity level on any single host may not be sufficiently high enough to be any intrusion alert by itself. However, the "login error" in a group of machines may be high enough to be treated as an intrusion. Each host reports the alerts to the monitors and the intrusion alerts can be raised in the monitors by the corresponding filter agents.

175.4 Implementation

System environment. We use Windows as our implementing environment and the Distributed Component Object Model (DCOM) as distributed objects communication technologies. DCOM is a protocol that enables software components to communicate directly over a network in a reliable, secure, and efficient manner. DCOM is an inherently symmetric network protocol and programming model and provides high, interactive communication between clients and servers and among peers.

Programming tool in our project is Visual C ++. We also use ATL lib [5] to implement COM and also Snort 2.8.0.

Classes. Main classes include SnortDataSource, Transceiver, Monitor, Agent and FiterAgent. SnortDataSource is to collect data from and provide to Agents. Transceiver is to manage all the Agents on the host, collect data from agents and

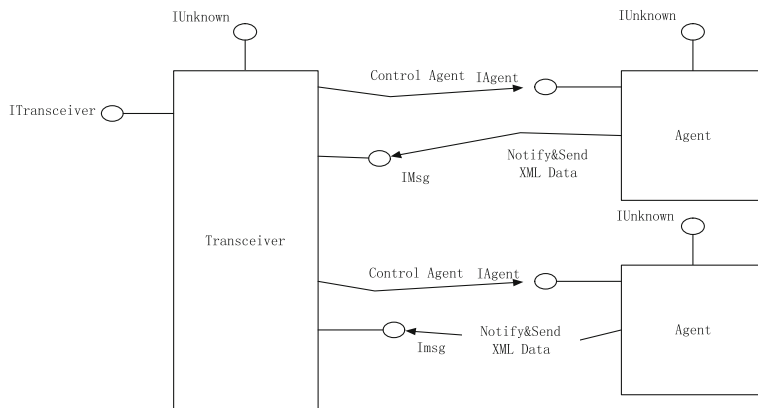


Fig. 175.3 Communication between agents and transceivers

reported to Monitors. Monitor is to manage different Transceivers. Transceiver can manage specific agents and report to up level Monitor or Controller. Agent is to collect the data which it is interested, analyze and send to Transceiver or send to Transceiver directly. FilterAgent is to run on Monitors with the task to correlate data from several transceivers and produce higher level Alerts and cluster the alerts.

Communication. We use XML alert using standard IDMEF (draft-ietf-idwg-idmef-xml-10.txt) and modify snort source by adding an output-module to output alerts in XML and output to a windows named-pipe.

For example, the communication between Agents and Transceivers is shown by Fig. 175.3. Agent is COM object according to users’ needs. Each Agent must implement the interface IAgent. Each FilterAgent must implement the interface IFilterAgent. Objects report to up level objects by IMsg.

175.5 Results

In our sample agent (LoginFAgent), we are detecting the intrusion from one IP address who is trying to login to several machines under one monitor. If more than three alerts are generated from Snort within 60 s, a correlated alert is generated and sent to controller to display.

For example, we install transceiver and agent on 129.186.157.123 and 129.186.93.230, telnet twice on 129.186.93.230 and telnet once on 129.186.157.123 within 60 s. A message “TELNET login incorrect” was popped on the intrusion information screen. Fig. 175.4 shows the result from the controller.

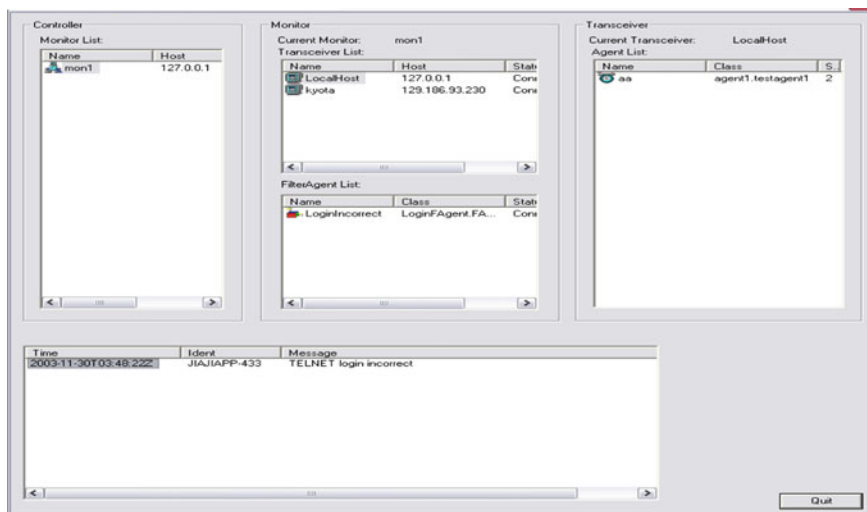


Fig. 175.4 User interface with login fail alert

175.6 Conclusion

DCOM is a mature technology for distributed system and our distributed intrusion detection system is suitable to be based on the DCOM.

Our intrusion detection system satisfies some intrusion-detection requirements compared to other architecture:

1. Easily to extend and flexibility: the main advantage of this architecture is its extensibility and flexibility. Although the whole architecture is easily to break down if there are some problems on monitors.
2. Quick response and efficiency: no database retrieval operation is required. All the data are put into a named pipe. Without mobile agents, the efficiency of agents are obvious
3. Scalability: our system can scale as the network grows. More monitors and transceivers can be added when more hosted are added. And the agents run independently of each other and they can be added without affecting the performance.
4. Manual control: the user interface allows for direct manipulation of each object in the model, including transceivers and its agents, monitors and related transceiver with filter agents.

Although, there are still some limitations of the architectures:

1. Security and access control: does not have own security access control. It depends on windows user id management.

2. Heterogeneous support: the current system supports well on the windows system. Although DCOM is also supported on other operating system, the Microsoft technologies are still best suitable on Windows platform.

Future work includes that the realizations of the agents still need other technologies to support, such as AI technologies.

References

1. Sadiq Ali Khan M (2010) Revealing method for the intrusion detection system. *J Telecommun* 2(1):36–41
2. Balasubramanian J, Garcia-Fernandez J (1998) An architecture for intrusion detection using autonomous agents. In: 14th Computer Security Applications Conference, 13–24 Dec 1998
3. Mark S (2003) The design and implementation of MAIDS [J]. *Microcomput Inf* 17:053–059
4. Valeur F, Vigna G, Kruegel C, Kemmerer RA (2004) A comprehensive approach to intrusion detection alert correlation. *IEEE Trans Dependable Secur Comput* 1(3):1–23
5. Grimes R (1999) *Beginning ATL 3 COM programming*. Wrox Press Inc, Chicago

Chapter 176

A Method for Uncertain Linguistic Multiple Attribute Decision Making and Its Application

Xiao-yi Liang

Abstract The aim of this chapter is to investigate the multiple attribute decision making problems to deal with the supplier selection in supply chain management with uncertain linguistic information, in which the information about attribute weights is incompletely known, and the attribute values take the form of uncertain linguistic variables. We developed a multiple attribute decision making method to select supplier by similarity to ideal supplier alternative in uncertain linguistic setting, by which the attribute weights can be determined. We utilize the uncertain linguistic weighted geometric operator to aggregate the uncertain linguistic variables corresponding to each alternative, and then rank the alternatives by means of the aggregated linguistic information. Finally, an example about supplier selection is shown to highlight the procedure of the proposed algorithm.

Keywords Multiple attribute decision making · Uncertain linguistic variables · Uncertain linguistic weighted geometric (ULWG) operator · Similarity to ideal point · Supplier selection

176.1 Introduction

In today's global market economy, with the development in the use of just-in-time manufacturing by a lot of companies, supply chain management (SCM) has become one important part of a company's overall strategy [1, 2]. Once a supplier

X. Liang (✉)
Shaanxi University of Science and Technology, Northern Suburb of University Park
Weiyang, Xi'an, 710021 Shaanxi, China
e-mail: liangxiaoyi99@126.com

becomes part of a well-managed and established supply chain, the relationship between the buyer and supplier will have a lasting effect on the competitiveness of the entire supply chain. Therefore, the supplier selection has become one of the most important issues for establishing an effective SCM system. The overall objective of supplier selection process is to maximize overall value to reduce purchase risk, and establish the closeness and long-term collaboration of relationships. A suitable supplier must supply high quality materials, deliver proper quantities of materials at proper time, reduce cost, and provide excellent services in order to satisfy what the buyer requires. Hence, the supplier selection is a multiple attribute decision-making (MADM) problem that includes both qualitative and quantitative attributes such as quality, delivery [3], performance history, warranties, price, technical capability, and financial position, etc.

The aim of this chapter is to develop a new method to solve uncertain linguistic MADM with incomplete weight. The remainder of this chapter is set out as follows. In the next section, we introduce some basic concepts and operational laws of uncertain linguistic variables. In Sect. 176.3 we introduce the MADM problem to deal with the supplier selection in supply chain management with uncertain linguistic information, in which the information about attribute weights is incompletely known, and the attribute values take the form of uncertain linguistic variables. In Sect. 176.4, an illustrative example is pointed out. In Sect. 176.4 we conclude the chapter and give some remarks.

176.2 Preliminaries

- (1) Let $S = \{s_i | i = 1, 2, \dots, t\}$ be a linguistic term set with odd cardinality. Any label, s_i represents a possible value for a linguistic variable, and it should satisfy the following characteristics 5 [4], 29:(1) The set is ordered: $s_i > s_j$, if $i > j$; (2) There is the negation operator: $neg(s_i) = s_j$ such that $j = t + 1 - i$; (3) Max operator: $\max(s_i, s_j) = s_i$, if $s_i \geq s_j$; (4) Min operator: $\min(s_i, s_j) = s_i$, if $s_i \leq s_j$. For example, S can be defined as

Definition 1 Let $\tilde{s}_1 = [s_{\alpha_1}, s_{\beta_1}]$ and $\tilde{s}_2 = [s_{\alpha_2}, s_{\beta_2}]$ be two uncertain linguistic variables, and let $len(\tilde{s}_1) = \beta_1 - \alpha_1, len(\tilde{s}_2) = \beta_2 - \alpha_2$, then the degree of possibility of $\tilde{s}_1 \geq \tilde{s}_2$ is defined as [5]

$$p(\tilde{s}_1 \geq \tilde{s}_2) = \frac{\max(0, len(\tilde{s}_1) + len(\tilde{s}_2) - \max(\beta_2 - \alpha_1, 0))}{len(\tilde{s}_1) + len(\tilde{s}_2)} \tag{176.1}$$

From Definition 1 we can get the following results easily:

- (1)

$$0 \leq p(\tilde{s}_1 \geq \tilde{s}_2) \leq 1, 0 \leq p(\tilde{s}_2 \geq \tilde{s}_1) \leq 1;$$

(2) $p(\tilde{s}_1 \geq \tilde{s}_2) + p(\tilde{s}_2 \geq \tilde{s}_1) = 1$. Especially, $p(\tilde{s}_1 \geq \tilde{s}_1) = p(\tilde{s}_2 \geq \tilde{s}_2) = 0.5$.

Definition 2 Let $\tilde{R} = (\tilde{r}_{ij})_{m \times n}$ be an uncertain linguistic decision matrix, then we call $\tilde{r}^* = (\tilde{r}_1^*, \tilde{r}_2^*, \dots, \tilde{r}_n^*)$ an negative point of attribute values, where $\tilde{r}_j^* = [\tilde{r}_j^{*L}, \tilde{r}_j^{*U}]$, $\tilde{r}_j^{*L} = \min_i \{ \tilde{r}_{ij}^L \}$, $\tilde{r}_j^{*U} = \min_i \{ \tilde{r}_{ij}^U \}$, $j = 1, 2, \dots, n$.

Definition 3 Let $\tilde{s}_1 = [s_{\alpha_1}, s_{\beta_1}]$ and $\tilde{s}_2 = [s_{\alpha_2}, s_{\beta_2}]$ be two uncertain linguistic variables, then we call

$$d(\tilde{s}_1, \tilde{s}_2) = |\alpha_1 - \alpha_2| + |\beta_1 - \beta_2| \tag{176.2}$$

the distance between \tilde{s}_1 and \tilde{s}_2 .

176.3 A Method for Uncertain Linguistic Multiple Attribute Decision Making and its Application to Supplier Selection

The following assumptions or notations are used to represent the MADM problems with incomplete weight information in uncertain linguistic setting [6]:

- (1) The alternatives are known. Let $A = \{A_1, A_2, \dots, A_m\}$ be a discrete set of suppliers;
- (2) The attributes are known. Let $G = \{G_1, G_2, \dots, G_n\}$ be a set of attributes;
- (3) The information about attribute weights is incompletely known. Let $w = (w_1, w_2, \dots, w_n) \in H$ be the weight vector of attributes, where $w_j \geq 0$, $j =$

$1, 2, \dots, n$, $\sum_{j=1}^n w_j = 1$, H is a set of the known weight information, which can

be constructed by the following forms, for $i \neq j$: *Form 1* A weak ranking: $w_i \geq w_j$; *Form 2* A strict ranking: $w_i - w_j \geq \alpha_i$, $\alpha_i > 0$; *Form 3* A ranking of differences: $w_i - w_j \geq w_k - w_l$ for $j \neq k \neq l$; *Form 4* A ranking with multiples: $w_i \geq \beta_i w_j$, $0 \leq \beta_i \leq 1$; *Form 5* An interval form: $\alpha_i \leq w_i \leq \alpha_i + \varepsilon_i$, $0 \leq \alpha_i < \alpha_i + \varepsilon_i \leq 1$.

In the real life, there always exist some differences between the negative supplier alternative of attribute values and the vector of attribute values corresponding to the supplier alternative $A_i (i = 1, 2, \dots, m)$. In what follows we define the distance $d(\tilde{z}(w), \tilde{z}_i(w))$ between the overall value $\tilde{z}(w)$ of negative ideal supplier alternative and the overall value $\tilde{z}_i(w)$ of the supplier alternative A_i :

$$d(\tilde{z}(w), \tilde{z}_i(w)) = \sum_{j=1}^n w_j d(\tilde{r}_j^*, \tilde{r}_{ij}) \tag{176.3}$$

Obviously, the larger $d(\tilde{z}(w), \tilde{z}_i(w))$, the better the supplier alternative A_i will be. Thus, a reasonable weight vector $w^* = (w_1^*, w_2^*, \dots, w_n^*)$ should be determined

so as to make all the distances $d(\tilde{z}(w), \tilde{z}_i(w))$ ($i = 1, 2, \dots, m$) as larger as possible, which means to maximize the following distance vector:

$$d(w) = (d(\tilde{z}(w), \tilde{z}_1(w)), d(\tilde{z}(w), \tilde{z}_2(w)), \dots, d(\tilde{z}(w), \tilde{z}_m(w))) \tag{176.4}$$

under the condition $w \in H$, where H is the set of the known weight information defined as in Sect. 176.3. In order to do that, we establish the following multiple objective optimization model:

$$(M - 1) \begin{cases} \text{maximize } d(w) = (d(\tilde{z}(w), \tilde{z}_1(w)), d(\tilde{z}(w), \tilde{z}_2(w)), \dots, d(\tilde{z}(w), \tilde{z}_m(w))) \\ \text{subject to } w \in H, \sum_{j=1}^n w_j = 1, w_j \geq 0, j = 1, 2, \dots, n \end{cases}$$

By linear equal weighted method, the above model can be transformed into a single-objective programming model:

$$(M - 2) \begin{cases} \text{maximize } d(w) = \sum_{i=1}^m \frac{1}{m} \left(\sum_{j=1}^n d(\tilde{r}_j^*, \tilde{r}_{ij}) w_j \right) \\ = \sum_{j=1}^n w_j \sum_{i=1}^m \frac{1}{m} (d(\tilde{r}_j^*, \tilde{r}_{ij})) = \sum_{j=1}^n d(\tilde{r}_j^*, \tilde{r}_j) w_j \\ \text{subject to } w \in H, \sum_{j=1}^n w_j = 1, w_j \geq 0, j = 1, 2, \dots, n \end{cases}$$

where $\tilde{r}_j = \frac{1}{m} \sum_{i=1}^m \tilde{r}_{ij}, j = 1, 2, \dots, n$.

By solving the model(M-2), we get the optimal solution $w^* = (w_1^*, w_2^*, \dots, w_n^*)$, which can be used as the weight vector of attributes.

Based on the above models, we develop a practical MADM problem to deal with the supplier selection, in which the information about attribute weights is incompletely known, and the attribute values take the form of uncertain linguistic variables. The method involves the following steps:

Step 1. Establish the single-objective programming model (M-2);

Step 2. Solve the model (M-2) to obtain the attribute weights;

Step 3. Utilize the weight vector $w^* = (w_1^*, w_2^*, \dots, w_n^*)$ and by Eq. 176.3, we obtain the overall values $\tilde{z}_i(w^*) (i = 1, 2, \dots, m)$ of the supplier alternative $A_i (i = 1, 2, \dots, m)$;

Step 4. Rank these overall preference values $\tilde{z}_i(w^*) (i = 1, 2, \dots, m)$, we first compare each $\tilde{z}_i(w^*)$ with all the $\tilde{z}_j(w^*) (j = 1, 2, \dots, m)$ by using Eq. (176.1). For simplicity, we let $p_{ij} = p(\tilde{z}_i(w^*) \geq \tilde{z}_j(w^*))$. Then we develop a complementary matrix as $P = (p_{ij})_{m \times m}$, where $p_{ij} \geq 0, p_{ij} + p_{ji} = 1, p_{ii} = 0.5, i, j = 1, 2, \dots, n$

Summing all the elements in each line of matrix P , we have

$$p_i = \sum_{j=1}^m p_{ij}, i = 1, 2, \dots, m.$$

Then we rank the collective overall preference values $\tilde{z}_i(w^*)(i = 1, 2, \dots, m)$ in descending order in accordance with the values of $p_i(i = 1, 2, \dots, m)$.

Step 5. Rank all the supplier alternative $A_i(i = 1, 2, \dots, m)$ and select the best one(s) in accordance with $p_i(i = 1, 2, \dots, m)$.

Step 6. End.

176.4 Conclusion

With respect to the MADM problems to deal with the supplier selection in supply chain management, in which the information about attribute weights is incompletely known, and the attribute values take the form of uncertain linguistic variables. We developed a new method to select supplier by similarity to negative ideal supplier in uncertain linguistic setting, by which the attribute weights can be determined. Based on this model, we develop a method to rank alternatives and to select the most desirable one(s). At last, a practical example about supplier selection is provided to illustrate the proposed method.

References

1. Xu ZS (2004) Uncertain multiple attribute decision making: methods and applications. Tsinghua University Press, Beijing
2. Xu ZS (2006) A method for multiple attribute decision making with incomplete weight information in linguistic setting. Knowledge-based systems 20(8):719–725
3. Xu ZS (2007) An interactive procedure for linguistic multiple attribute decision making with incomplete weight information. Fuzzy Optim Decis Mak 6(1):17–27
4. Wu ZB, Chen YH (2007) The maximizing deviation method for group multiple attribute decision making under linguistic environment. Fuz Sets Syst 158(14):1608–1617
5. Xu ZS (2004) Uncertain linguistic aggregation operators based approach to multiple attribute group decision making under uncertain linguistic environment. Info Sci 168:171–184
6. Xu ZS (2006) Induced uncertain linguistic OWA operators applied to group decision making. Info Fus 7:231–238

Chapter 177

Web2.0 Environment of Personal Knowledge Management Applications

Jie Fang and Liqun Gong

Abstract Web2.0 with its superior technology platform and outstanding features for personal knowledge management provides powerful tools. This chapter discusses the application of personal knowledge management, from the knowledge acquisition, knowledge storage, knowledge sharing and knowledge using four aspects of proposed Web2.0 environment personal knowledge management skills, and puts forward the Web2.0 environment of individual knowledge management related tools.

Keywords Web2.0 • Personal knowledge management • Tools

177.1 The Concept of Personal Knowledge Management

Personal Knowledge Management The concept was introduced by the United States first proposed by Prof. Paul Dorsey's [1, 2]. For this concept, there are many different definitions. I think that personal knowledge management includes three meanings: First, individuals have the knowledge to manage; Second, through a variety of ways to learn new knowledge, learn from and draw on the experience of others, the standing views and ideas with

J. Fang (✉)
Nan Jing University, Nan Jing, China
e-mail: fangjie0701@sina.com

J. Fang
Urumqi Vocational University, Urumqi, China

L. Gong
Chang ji College Chang ji, Xinjiang, China
e-mail: gonglq1010@sina.com

other people's ideological essence, achieve implicit knowledge to explicit, stimulate innovation and ne advantages and strengths to make up their own thinking and knowledge of defects, characteristic constantly construct their own knowledge; Third, the use of his own knowledge and the long-w knowledge.

For personal knowledge management, we must first analyze what we have knowledge, how to acquire knowledge, the structure of this knowledge, implementation of the mobilization of this knowledge to accomplish our goals, but also in the process, including how to create knowledge. Personal knowledge management can be divided into the following: interpersonal communication resources (such as address book contacts, the characteristics and strengths of each person, etc.); Communication Management (correspondence, e-mail, fax, etc.); personal time management tool (transaction reminders, to-do, and personal memo); network resource management (site management and connection); document file management.

Visible attention to personal knowledge management is the integration of their information resources to enhance the core competitiveness of the individual.

177.2 Personal Knowledge Management in Web2.0 Times

Web2.0 is presented by Flickr, Craigsliis, Tlinkedin, Tribes, Ryze, friendster, De.licio.us, 43Things.com site [3], In Blog, TAG, SNS, RSS, Wiki, Web office applications and other social software as the core, according to six degrees of separation, XML, Ajax and other new theory and technology of the Internet generation mode. The most prominent feature of Web2.0 is a strong user participation, interaction, personalization, community, sharing, so that ordinary users into the Internet truly become the masters of the Internet.

Different communities, and enhance communication skills, especially web-based is more conducive to the exchange.

177.3 Personal Knowledge Management Skills

In web2.0 personal knowledge management environment, we master the knowledge acquisition, knowledge storage, knowledge sharing and knowledge utilization skills.

177.3.1 Knowledge Acquisition

Knowledge has both explicit knowledge and tacit knowledge, knowledge of different types, and its access mechanism. Tacit knowledge deep in the minds of individuals's experience, it is difficult with the language, communication.

Management guru Peter Drunker once said tacit knowledge can only be demonstrated to prove its existence [4], the only way to learn this knowledge is the understanding and practice. Therefore, it can carry through with others Line communication and learning in order to gain their tacit knowledge. Explicit knowledge under the recording. Recording different vector into different categories. The knowledge resources on the network, you can visit the site, search engines, etc. Knowledge of personal information in your organization, you can find the knowledge Database to obtain the required knowledge and information; for published books, newspaper article Chapter, you can purchase books, subscribe to the newspaper and access to relevant knowledge.

177.3.2 Knowledge Storage

It can record and express a personal Blog to learn bit by bit life feelings, so that usually sparks flashes of thought and intelligence are preserved. After a period of knowledge management, people through their own Blog and web browsing pick better able to see their mental models and self-evaluation and reflection, self-transcendence.

177.3.3 Knowledge Sharing

Web2.0 provides a wealth of knowledge-sharing platform, such as Blog, SNS, P2P, Wiki, etc. These technology platforms to accelerate knowledge sharing, allowing users to maximize the personality are reflected, in addition to both the social characteristics of the users to easily find like-minded. Personal Portal also provide people with a variety of interaction and feedback, where people can pick through the Blog and web platforms such as building self-organized learning of the different communities, and enhance communication skills, especially web-based is more conducive to the exchange.

177.3.4 Knowledge Utilization

Through the network across time and space, to communicate, explore, collaboration, inquiry, to solve problems in learning, in order to achieve the experience and wisdom to share. To optimize the learning process, it is also fully tap people's creative thinking, to produce new technologies and new services, innovative thinking, more conducive to the realization of knowledge value and added value.

177.4 Personal Knowledge Management Tools.

Web2.0 platform of RSS, BLOG, TAG, etc. for personal knowledge provides a convenient channel for.

177.4.1 Social Book Mark

Social book mark provides a kind of collection, classification, sorting, sharing Internet information resources. Use it to store a list of URLs and related information, the use of labels (Tag) to index the URL of the Web site of resources and orderly classification and indexing, the website and the social sharing of information possible, human participation in the sharing of the value of the process of site assessment is given. Participation by groups of people mining costs under control and effective information, through knowledge classification mechanism that has the same interest it easier for users to share information and exchange network to pick the site. Social book mark site showing a picture of community knowledge classification. Abstract effective network make up the traditional browser favorites is not easy to move short depression can be placed in personal favorites on the network, anytime, anywhere visit view. Pick through the Social book mark, users can easily view; manage their useful pages and pages through the form of excerpts to share their collections with others, to experience the fun of sharing. Its value is webmaster pick or other users will think the most valuable content of their recommendations to the majority of Internet users, thus saving vast amounts of information users choose their own in the interest of time consuming content. Despite being a social behavior, social book mark for users to pick a real sense of personal knowledge management provides a possibility. Pick through the social book mark, the user can save the Internet to have a collection of mechanisms for reading the information, and make the necessary descriptions and notes, accumulated to form personal knowledge. Meanwhile, the social book mark allows users to pick between the collections of information sharing. This will greatly reduce the user access to information for all the costs involved can make it easier for users to get more volume, more point of information. Through knowledge classification, you can quickly make to have the same interests and specific skills of people to form a communication group, enhance each other through the exchange and sharing of knowledge, to meet the communication, presentation and other social needs.

In addition, the social book mark pick up to date, so the user can timely and effectively obtain the required information; it exist in the network, allowing the user access to the Internet, no matter when and where they can live for. It can see from social bookmark-oriented classification of network, system, update, exchange, innovation and good personal knowledge management embodies the spirit and the inner core requirements.

177.4.2 RSS

Also known as Rss [5], is quickly given an effective way to access information. It can automatically view the user's needs and monitoring of certain specified content of the site, the content will be pushed to the intended users and Automatic updates. With RSS, users no longer have a website a website, a page of a web page to gather information, as long as the content subscription required In an RSS reader, once with updates, RSS reader will automatically updates t, for the user to save a lot of time.

177.4.3 TAG

TAG is processed to obtain personal knowledge, TAG was formed under Web2.0 [6], a new classification method, the function of a keyword in the form of content classification Complex mass of information on the Internet, only to classify and organize them for their own use. TAG keyword classification and the past provide the key words are different, the user can select according to their own understanding of key words, and this feature is conducive to individual personality of their own knowledge management. Parallel between the various TAG relationships, individuals may be related to the analysis of TAG, resulting in the classification of correlation can be effective management of associated information.

177.4.4 Personal Portal (Personal Information Portal)

Through a personal portal, users can become their own Internet portal and center, "the house and do not know everything about the world"; through a personal portal, users can share the information; through a personal portal, users can optimize information acquisition, information transmission, Information processing, information applications; through the personal portal, users can be learning and innovation; users as their own personal portal on the Internet as representative.

A complete personal portal at least should have the following modules:

177.4.4.1 Network Summary and Article Collection Network

The formation of such a truly integrated approach Knowledge base, continuously integrating personal knowledge, and the formation of orderly system;

BLOG

Demonstrate personal knowledge and ideas, to communicate and interact with others;

RSS

Constantly updated network information, real-time access to information resources;

Network Storage

Integrated image storage (Flicker), file storage, Music and other multimedia content network storage;

E-mail

Online read and send personal e-mail of all E-mail;

IM

Chat in real time exchanges;

SEARCH

Integrated search engine technology.

177.4.5 BLOG

Blog is time for individuals or groups to a variety of records made in order, and constantly updated. Blog there are many, such as personal blog, corporate blog, product blog, knowledge base, blog, public community blog, etc., Individuals can learn according to their daily needs or for a particular type or a combination of several blog types of information they need to collect.

The value of personal knowledge management is not purely the accumulation of personal knowledge, but to realize the value of knowledge, knowledge about individual creativity into the process of organizational knowledge management. In Personal Knowledge Management and Knowledge Management between the need to undertake an individual knowledge and organizational knowledge [7, 8], personal knowledge into organizational knowledge system platform, the platform must be able to promote personal knowledge management in the organization under the positive environment for personal knowledge management, that is, on the initiative of individual emphasis on knowledge management.

177.4.5.1 Blogs in the Application of Individual Knowledge Management

Many features of the Blog provide more effective space for personal knowledge management platform. Blog, can be expressed as follows in Personal Knowledge management:

A blog is very personal network platform, free and convenient, BBS and other network communication tools than to better reflect the individual's characteristics.

Blog provides free programs page template, personal preferences can be selected based on templates and template for planning, this process requires the full participation of individuals and can reflect the individual needs and characteristics.

Blog written by a simple log of operations management background in the blog category of personal knowledge management. This classification is mainly embodied in the professional: First, record their own classification are thinking, easy to sort out the brain of knowledge, but also a process of self-display; The second is that interest us the aggregation classification, the source from the knowledge level of the formation of personal knowledge system, and increasing source of knowledge through learning the link capacity.

It can increase the industry blog, websites and other links in order to facilitate mutual visits, comments and exchanges. by gathering to communicate with their people, to get knowledge about his own personal benefit under the branches of the knowledge of knowledge, to make up their own shortcomings, and make implicit of explicit knowledge can be said that this process can best embody the blog AC characteristics.

Knowledge of the blog presented by way of opening the blog site, to facilitate knowledge sharing. Sharing is one of the salient features of the blog is popular blog network reasons.

Knowledge of the blog writing is a systematic process that can produce subtle effects on individuals.

Blog help organizations manage through the personification of the platform, its openness and freedom determine the level of organizational knowledge management level, and so on.

177.5 Conclusion

In the development of the times, the method of organize knowledge and use of knowledge, creation of knowledge is developing and changing. Pick either social book mark, personal library or personal portal, as demand changes and network development will continue to improve in function and development. However, ordering, sharing, communication, innovation will not change the subject, and personal knowledge management is also a prerequisite for success.

With its wealth of technology and knowledge management platform Web2.0 provide support to individuals. In the Web2.0 environment, personal knowledge management more efficient, fast. Personal knowledge and social knowledge-the cycle is updates, so as to promote the continuous development of the individual and society. We believe that the future of personal knowledge management will be more convenient, efficient and closer to the lives of ordinary people.

References

1. Jones R (2009) Personal knowledge management through communicating [J]. *Online Inform Rev* 33(2):225–236
2. Xu H (2008) Personal knowledge management application [J]. *Mod Inf* 9:21–23

3. Shi XW (2009) Personal Portal 3.0 personal knowledge management [J]. *Inf Doc* 2:61–62
4. Wei B (2010) Based on personal knowledge management Web 2.0 [J]. *Mod Inf* 5:170–173
5. Dong Y (2010) Web personal knowledge management network [J]. *Res Libr Sci* (9):32–35
6. Zhihui Z, Chao C (2008) Discussion on personal knowledge management [J]. In: Twenty second national conference proceedings of computer information management, pp 160–163
7. Cai xu H (2009) Two of personal knowledge management trends [J]. *Mod Educ Tech* 4:56–58
8. Liu Z, Nu N, Zhang Z (2009) Analysis of knowledge management research abroad [J]. *Libr Inf Serv* 5:91–94

Chapter 178

Semantic Web and Its Applications

WenYing Guo

Abstract The World Wide Web has changed the way people communicate with each other and the way business is conducted. The Semantic Web is a “web of data” that enables machines to understand the semantics or meaning of information on the World Wide Web. This paper provides the basic knowledge required to the Semantic Web and gives an overview of some semantic web applications. It firstly describes the Extensible Markup Language (XML), introduces the basic concepts of Resource Description Framework (RDF) and Resource Description Framework Schema (RDFS), describes the contents and purpose of ontologies and owl specification documents, then gives the brief descriptions about logic and rules used, finally shows some typical semantic applications.

Keywords Semantic web · Extensible markup language (XML) · XML schema · Resource description framework(RDF) · RDF schema · Ontologies · Web ontology language (OWL) · Simple protocol and RDF query language (SPARQL) · Rule interchange format (RIF)

178.1 Introduction

The World Wide Web has changed the way people communicate with each other and the way business is conducted, it has also changed the way people think of computers. Most of today’s Web content is suitable for human consumption. Even

W. Guo (✉)

College of Computer Science and Information Engineering, Zhejiang GongShang University, Hangzhou 310035, China

e-mail: gwy@mail.zjgsu.edu.cn

W. Guo

Zhejiang Gongshang University, Xiasha University Town, No.18, Xuezheng Street, Hangzhou 310018, China

the web content that is generated automatically from databases is usually presented without the original structural information found in databases. Typical uses of the Web today involve people's seeking and making use of information, searching for and getting in touch with other people, reviewing catalogs of online stores and ordering products by filling out forms, and viewing adult material.

The main obstacle to providing better support to Web users is that, at present, the meaning of Web content is not machine-accessible [1, 2]. Of course, there are tools that can retrieve texts, split them into parts, check the spelling and count their words. But when it comes to interpreting sentences and extracting useful information for users, the capabilities of current software are still very limited. However, if the interaction between person and hypertext could be so intuitive that the machine-readable information space gave an accurate representation of the state of people's thoughts, interactions, and work patterns, then machine analysis could become a very powerful management tool.

The Semantic Web is a "web of data" that enables machines to understand the semantics or meaning of information on the World Wide Web. It extends the network of hyperlinked human-readable web pages by inserting machine-readable metadata about pages and how they are related to each other, enabling automated agents to access the web more intelligently and perform tasks on behalf of users. It provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries.

The Semantic Web is about two things. It is about common formats for integration and combination of data drawn from diverse sources, where on the original Web mainly concentrated on the interchange of documents. It is also about language for recording how the data relates to real world objects. That allows a person, or a machine, to start off in one database, and then move through an unending set of databases which are connected not by wires but by being about the same thing.

The Semantic Web technologies include the Resource Description Framework (RDF) [2, 3], a variety of data interchange formats and notations such as RDF Schema (RDFS) and the Web Ontology Language (OWL), all of which are intended to provide a formal description of concepts, terms, and relationships within a given knowledge domain.

178.2 Key Technologies

Many files on a typical computer can be loosely divided into documents and data. Documents like mail messages, reports, and brochures are read by humans. Data, like calendars, address books, playlists, and spreadsheets are presented using an application program which lets them be viewed, searched and combined in many ways.

Currently, the World Wide Web is based mainly on documents written in Hypertext Markup Language (HTML) [4], a markup convention that is used for

coding a body of text interspersed with multimedia objects such as images and interactive forms Metadata tags. The Semantic Web takes the solution further. It involves publishing in languages specifically designed for data: RDF, OWL, and Extensible Markup Language (XML) [4–7]. HTML describes documents and the links between them. By contrast, RDF, OWL and XML can describe arbitrary things such as people, meetings, or airplane parts. These technologies are combined in order to provide descriptions that supplement or replace the content of Web documents. Thus, content may manifest itself as descriptive data stored in Web-accessible databases, or as markup within documents. The machine-readable descriptions enable content managers to add meaning to the content, In this way, a machine can process knowledge itself, instead of text, using processes similar to human deductive reasoning and inference, thereby obtaining more meaningful results and helping computers to perform automated information gathering and research.

Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data. The semantic web comprises the standards and tools of XML, XMLS, RDF, RDFS and OWL that are organized in the Semantic Web Stack. OWL adds more vocabulary for describing properties and classes: among others, relations between classes, cardinality, equality, richer typing of properties, characteristics of properties, and enumerated classes. SPARQL is a protocol and query language for semantic web data sources. Rule Interchange Format (RIF) is the Rule Layer of the Semantic Web Stack.

178.2.1 XML and XMLS

Extensible Markup Language is a set of rules for encoding documents in machine-readable form. It is a universal metalanguage for defining markup. It provides a uniform framework and a set of tools like parsers, for interchange of data and metadata between applications. It provides an elemental syntax for content structure within documents, yet associates no semantics with the meaning of the content contained within. XML is not at present a necessary component of Semantic Web technologies.

XML's design goals emphasize simplicity, generality, and usability over the Internet. It is a textual data format with strong support via Unicode for the languages of the world. Although the design of XML focuses on documents, it is widely used for the representation of arbitrary data structures, for example in web services.

Many application programming interfaces (APIs) have been developed that software developers use to process XML data [8], and several schema systems exist to aid in the definition of XML-based languages.

XML Schema is a language for providing and restricting the structure and content of elements contained within XML documents. It is published as a W3C recommendation in May 2001, is one of several XMLS languages [9, 10].

178.2.2 RDF and RDFS

Extensible Markup Language does not provide any means of talking about the semantics (meaning) of data. However, The RDF is a language for representing information about resources and their relationships in the World Wide Web. An RDF-based model can be represented in XML syntax. It is particularly intended for representing metadata about Web resources and modification date of a Web page, copyright and licensing information about a Web document, or the availability schedule for some shared resource. RDF can also be used to represent information about things that can be identified on the Web, even when they cannot be directly retrieved on the Web. Examples include information about items available from on-line shopping facilities (e.g., information about specifications, prices, and availability), or the description of a Web user's preferences for information delivery.

RDF is intended for situations in which the information needs to be processed by applications, rather than being only displayed to people. RDF provides a common framework for expressing this information so it can be exchanged between applications without loss of meaning. Since it is a common framework, application designers can leverage the availability of common RDF parsers and processing tools.

RDF is based on the idea of identifying things using Web identifiers (called Uniform Resource Identifiers, or URIs), and describing resources in terms of simple properties and property values. This enables RDF to represent simple statements about resources as a graph of nodes and arcs representing the resources, and their properties and values. It is a universal language that lets users describe resources using their own vocabularies. RDF does not make assumptions about any particular application domain, nor does it define the semantics of any domain. It is up to the user to do so in RDFS.

RDFS extends RDF and is a vocabulary for describing properties and classes of RDF-based resources, with semantics for generalized-hierarchies of such properties and classes. RDFS defines the vocabulary used in RDF data models. RDFS can define the vocabulary, specify which properties apply to which kinds of objects and what values they can take, and describe the relationships between objects. The fundamental concepts of RDF are resources, properties and statements. RDF Schema is an extensible knowledge representation language, providing basic elements for the description of ontologies. The first version was published by the World-Wide Web Consortium (W3C) in April 1998 [11], and the final W3C recommendation was released in February 2004 [12]. Many RDFS components are included in the more expressive language OWL.

178.2.3 Ontologies

Ontologies have proved to be a key technology for effective information access because they help to overcome some of the problems of free-text search by relating

and grouping relevant terms in a specific domain as well as providing a controlled vocabulary for indexing information.

Ontology languages allow users to write explicit, formal conceptualizations of domain models. The main requirements are a well-defined syntax, efficient reasoning support, a formal semantics, sufficient expressive power, convenience of expression.

178.2.4 OWL

The importance of a well-defined syntax is clear and known from the area of programming languages. A formal semantics describes the meaning of knowledge precisely. Precisely here means that the semantics does not refer to subjective intuitions, nor is it open to different interpretations by different people (or machines). The importance of a formal semantics is well-established in the domain of mathematical logic.

The Web Ontology Working Group of W3C identified a number of characteristic use-cases for the Semantic Web, and also identified the need for a more powerful ontology modeling language. This led to a joint initiative to define a richer language OWL which is the language that is aimed to be the standardized and broadly accepted ontology language of the Semantic Web. The OWL is a family of knowledge representation languages for authoring ontologies. The languages are characterised by formal semantics and RDF/XML-based serializations for the Semantic Web.

OWL build upon RDF and RDFS and have the same kind of syntax, instances are defined using RDF descriptions and most RDFS modeling primitives are used, in the sense that OWL would use the RDF meaning of classes and properties (rdfs:Class, rdfs:subClassOf, etc.) and would add language primitives to support the richer expressiveness required. OWL is the proposed standard for Web ontologies. It allows us to describe the semantics of knowledge in a machine-accessible way. Formal semantics and reasoning support is provided through the mapping of OWL on logics. Predicate logic and description logics have been used for this purpose.

178.2.5 Logic and Inference: Rules

Simple Protocol and RDF Query Language (SPARQL) is an RDF query language. It allows for a query to consist of triple patterns, conjunctions, disjunctions, and optional patterns. It was standardized by the RDF Data Access Working Group (DAWG) of the World Wide Web Consortium, and is considered a key semantic web technology.

The Rule Interchange Format (RIF) is a W3C Recommendation. RIF is part of the infrastructure for the semantic web, along with RDF and OWL. Although originally envisioned by many as a “rules layer” for the semantic web, in reality the design of RIF is based on the observation that there are many “rules languages” in existence, and what is needed is to exchange rules between them.

178.3 Applications

178.3.1 Business-to-Business Electronic Commerce

Many manufacturing companies today interact with hundreds of suppliers in order to obtain all the parts that go into making a single product. Online procurement has been identified as a major potential cost saver. Static, long-term agreements with a fixed set of suppliers can be replaced by dynamic, short-term agreements in a competitive open marketplace. Whenever a supplier is offering a better deal, the manufacturing company wants to be able to switch rather than being locked into a long-term arrangement with another supplier.

This online procurement is one of the major drivers behind B2B e-commerce. Current efforts in B2B e-commerce rely heavily on a priori standardization of data formats, that is, off-line industry wide agreements on data formats and their intended semantics. Electronic Components, Semiconductor Manufacturing and Telecommunications companies work to create and implement industrywide, open e-business process standards. These standards form a common e-business language, aligning processes between supply chain partners on a global basis.

Since such data formats are specified in XML, no semantics can be read from the file alone, and partners must agree in time-consuming and expensive standards negotiations, followed by hard-coding the intended semantics of the data format into their code.

A more attractive road would use formats such as RDF Schema and OWL, with their explicitly defined formal semantics. This would make product descriptions opening the way for much more liberal online B2B procurement processes than currently possible.

178.3.2 E-Learning

Traditional learning processes have not been suitable for every potential learner. The emergence of the Internet has paved the way for implementing new educational processes. Compared to traditional learning, e-learning is not driven by the instructor. In particular, learners can access material in an order that is not pre-defined, and can compose individual courses by selecting educational material. For

this approach to work, learning material must be equipped with additional information to support effective indexing and retrieval.

The use of metadata is a natural answer and has been followed, in a limited way, by librarians for a long time. In the e-learning community, standards such as IEEE LOM have emerged. They associate with learning materials information, such as educational and pedagogical properties, access rights and conditions of use, and relations to other educational resources. Although these standards are useful, they suffer from a drawback common to all solutions based solely on metadata, lack of semantics. As a consequence combining of materials by different authors may be difficult; retrieval may not be optimally supported, and the retrieval and organization of learning resources must be made manually. These kinds of problems may be avoided if the Semantic Web approach is adopted.

The key ideas of the Semantic Web, namely, common shared meaning (ontology) and machine-processable metadata, establish a promising approach for satisfying the e-learning requirements. It can support both semantic querying and the conceptual navigation of learning materials.

178.3.3 Semantic Web Search

Ontologies have proved to be a key technology for effective information access because they help to overcome some of the problems of free-text search by relating and grouping relevant terms in a specific domain as well as providing a controlled vocabulary for indexing information. A number of thesauri have been developed in different domains of expertise.

Semantic Web technology plays multiple roles in this architecture. First, RDF is used as an interoperability format between heterogeneous data sources. Second, an ontology is itself represented in RDF. Each of the separate data sources is mapped onto this unifying ontology, which is then used as the single point of entry for all of these data sources.

178.3.4 Data Integration

Normally a company operates thousands of databases, often duplicating and re-duplicating the same information, and missing out on opportunities because data sources are not interconnected. Current practice is that corporations rely on costly manual code generation and point-to-point translation scripts for data integration.

While traditional middleware improves and simplifies the integration process, it does not address the fundamental challenge of integration: the sharing of information based on the intended meaning, the semantics of the data.

Using ontologies as semantic data models can rationalize disparate data sources into one body of information. By creating ontologies for data and content sources

and adding generic domain information, integration of disparate sources in the enterprise can be performed without disturbing existing applications. The ontology is mapped to the data sources (fields, records, files, documents), giving applications direct access to the data through the ontology.

178.4 Future

The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation. One believable thing is that semantic technologies will bring a robust and extensible enhancement for emerging Web applications. Definitely, data reuse, distribution and aggregation can be greatly facilitated by the Semantic Web's infrastructures. Web proponents should encourage themselves to embrace those Semantic Web technologies. Such Web-like systems generate a lot of excitement at every level, from major corporation to individual user, and provide benefits that are hard or impossible to predict in advance.

References

1. Semantic web home page, <http://www.w3.org/>
2. The protégé homepage, <http://protege.stanford.edu/>
3. Guo W (2008) "Using semantic web technologies for ubiquitous computing", proceedings The first IEEE international conference on Ubi-Media computing and workshops, Jul 15 2008
4. Guo W (2008) Reasoning with semantic web technologies in ubiquitous computing environment, *The J Softw (JSW)* vol 3 (9) (Special issue on "mobile system, agent technology, and web services")
5. Guo W (2008) Ontology design for supporting matchmaking in e-commerce. international symposium on information science and engineering, vol 2, pp 410–413
6. Martinl D, Paolucci M, McIlraith S, Burstein M etc. (2005) "Bringing semantics to web services: The OWL-S approach", *SWSWPC 2004. LNCS 3387*, pp. 26–42
7. Trastour D, Bartolini C, Preist C (2003) Semantic web support for the business-to-business e-commerce pre-contractual lifecycle. *Comput Netw* 42:661–673
8. Bener B, Ozadali V, Ilhan E (2009) Semantic matchmaker with precondition and effect matching using SWRL. *Expert Sys with Appl* 36:9371–9377
9. Wombacher A, Fankhauser P, Mahleko B, Neuhold E (2004) Matchmaking for business processes based on choreographies
10. Colucci S, Noia T, Sciascio E, Donini F, Mongiello M (2005) Concept abduction and contraction for semantic-based discovery of matches and negotiation spaces in an e-marketplace, *Electron Commer Res and Appl* pp. 345–361
11. DAML home page, <http://www.daml.org/>
12. Ontology Inference Layer (OIL) home page, <http://www.ontoknowledge.org/oil/>

Chapter 179

A Semantics-Based Web Service Matching Framework and Approach

Bo Jiang, Tao He and Jiale Wang

Abstract Web service matching is an important component of web service system architecture. The recall and precision rate of traditional web service matching methods is usually not satisfactory. Semantic description of web service using ontology provides a way for service matching on semantic level. The chapter proposes an approach for semantic web service matching. Experimental results show that the proposed method can improve the recall and precision rate effectively.

Keywords Web service matching · Ontology · Semantic web service

179.1 Introduction

With the rapid development of Internet technologies and application, web service [1] is an emerging web application form and remote access criteria. Web service settles the problems of sharing data on heterogeneous platforms, and it also can support system seamless integration in different areas. It becomes more and more

B. Jiang (✉) · T. He · J. Wang
College of Computer and Information Engineering, Zhejiang Gongshang University,
Xiasha University Town, Hangzhou, 310018, China
e-mail: nancybjjiang@mail.zjgsu.edu.cn

T. He
e-mail: zjlahetao@126.com

J. Wang
e-mail: wjl8026@mail.zjgsu.edu.cn

difficult to find the service precisely and efficiently while the web service increases as well.

Traditional web service matching mechanism is always key words matching, lack of machine understandable semantic information [2], so that the recall and precision rate is on the low side. Therefore we can import machine understandable semantic information based on semantics to provide more precise and efficient web service matching approach, which is used to find relevant web service quickly.

Ontology is a formal presentation of knowledge as a set of concepts or terms within a domain. The knowledge presented by ontology marks web service, so that we can lead traditional web service based on syntactics to semantic level. The academia had already put forward the service description language based on ontology, such as OWL-S [3], WSMO [4]. Semantic web service matching [5] is the process of locating service request on semantic level. The chapter will propose a semantic web service matching framework, and a semantic web service matching approach based on the framework.

179.2 Web Service Matching Framework

In this chapter, the semantic web service is described by OWL-S structure [6], and separate the information which is provided by service profile in OWL-S into two aspects: service category, service function. The service category explains the category which the service belongs to, e.g. the service in transportation area. The service function describes the input and output of the service. Based on this analysis, this chapter proposes a total framework structure in Fig. 179.1 and its main modules and functions as follows:

- (1) Construct OWL-S service file: the service information provided by service provider and service requestor, combining with domain ontology base produces semantic web service description file OWL-S.
- (2) OWL-S/UDDI map [7]: this is used to implement UDDI semantic extension. The map mechanism of one to one map between OWL-S and UDDI can be applied to place service profile of OWL-S into UDDI.
- (3) UDDI and ontology base: UDDI center is used to store those published web service. Ontology Base: it consists of domain ontology, and service classification ontology. The domain ontology describes wordlist of a set of concepts within a specific field, and the relationships between those concepts. The service classification ontology is a di-partition of domain ontology.
- (4) Web service match module: this module is the core of the whole semantic web service matching framework. And it separates the map of semantic web service description information into two phases: firstly select web service by category matching, and then match the function, primarily input match and output match.

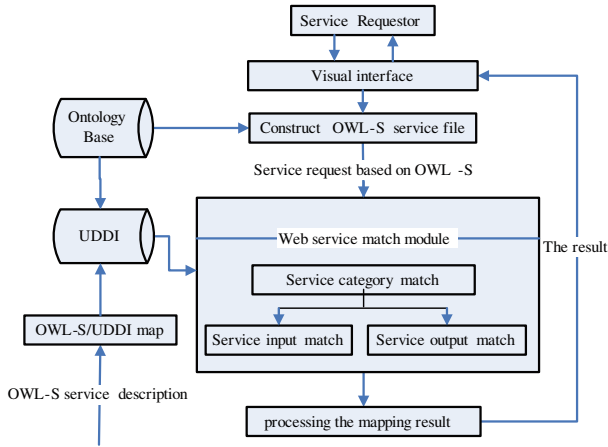


Fig. 179.1 Service matching framework based on ontology

179.3 Semantics-Based Web Service Matching Approach

179.3.1 Category Match of Semantic Web Service

The category match of semantic web service is the first phase of the whole matching process. The category is described by service category which is in the service profile of OWL-S. The service in this research uses the classification criteria of NAICS. I hypothesize that $A_{category}$ and $R_{category}$ are the category information of advertised service description and requested service description respectively. The relationship between the category similarity and matching level is summarized in Table 179.1:

Category matching can filter out that service independent or just intersect with the requested service category, and it is the first step to select the service, narrows the scope of service matching, saves matching time. Therefore it can lower the complexity of the matching.

179.3.2 Semantic Web Service Function Information Match

Considering the input and output in web service semantic description, we define the web service semantic description abstractly as follows:

Definition 1 Web service semantic description model $W = \langle I, O \rangle$. $I = \langle I_1, I_2, \dots, I_n \rangle$ is a concept vector, representing semantic description of N input parameters in W (web service). I_1, I_2, \dots, I_n represent their corresponding semantic concepts in the domain ontology base of N input parameters. $O = \langle O_1, O_2, \dots,$

Table 179.1 The relationship between the category similarity and matching level of service category

Category similarity	Matching level	Explanation
1.00	Exact	$A_{category} = R_{category}$
0.65	Plugin	$R_{category} \supseteq A_{category}$
0.35	Subsume	$A_{category} \supseteq R_{category}$
0.00	Other	Independent or intersect only

$O_m >$ is a concept vector, representing semantic description of M output parameters in W (web service). O_1, O_2, \dots, O_m represent their corresponding semantic concepts in the domain ontology base of M output parameters.

As a result we can change web service match into the match between requester’s ideal service semantic description model $W=<I, O >$ and service semantic description model in service base $W'=<I',O' >$. Furthermore it can transformed into the match between concept vectors I and I', O and O' in the unified domain ontology.

The service requestor is separated from the service provider before the service requestor finds the service which can meet the needs in a network environment. It is the reason why their semantic descriptions of web service are independent with each other. So the concept vectors $I, I', O,$ and O' may have different situations in aspects of the dimension and the elements order. We must define the match of any two concept vectors in the same domain ontology base before define the concept of web service match formally.

Definition 2 Define the calculation function of the similarity of any two concepts in the same domain ontology base as $Sim(C_1,C_2)$. C_1, C_2 are any two concepts in the domain ontology base, and the value of $Sim(C_1,C_2)$ is between 0 and 1(including 0 and 1). The bigger the value is, the more similar two concepts are. Based on definition 2, we can define the similarity of any two concept vectors A, B in the same domain ontology base.

Definition 3 Assume that $A = (A_1, A_2, \dots, A_n), B = (B_1,B_2,\dots, B_m)$ are any two concept vectors in the same domain ontology base, and then the correlation

matrix is $S_{AB} = \begin{pmatrix} Sim(A_1, B_1) & Sim(A_1, B_2) & \dots & Sim(A_1, B_m) \\ Sim(A_2, B_1) & Sim(A_2, B_2) & \dots & Sim(A_2, B_m) \\ \vdots & \vdots & & \vdots \\ Sim(A_n, B_1) & Sim(A_n, B_2) & \dots & Sim(A_n, B_m) \end{pmatrix}$, so that

we can get the formula of any two concept vectors in the same domain ontology base:

$$S(A, B) = \frac{1}{n} \sum_{i=1}^n (\max(\text{Sim}(A_i, B_j)), j \in [1, m]). \tag{179.1}$$

We can transform semantic web service match into the match of web service semantic description model by definition 1, and furthermore transform the match of concept vectors into the match of elements of concept vectors by definition 3,

that is, the match of any two concepts in the domain ontology base. In addition, the match of elements of concept vectors is calculated by similarity function which is defined in definition 2.

The researchers have already proposed various calculation approaches to the similarity of any two concept semantics in the ontology base of semantic web service match. According to the corresponding upper and lower (subClassOf) relationship of the input and output of web service, the chapter [8] divides semantic similarity into four types: extract, plugin, subsumes and fail, but this method seems too simple for the distinction of concept similarity. Based on SUMO, the paper [9] proposes an approach to calculate the semantic distance between two concept semantic similarity, but this approach did not consider the influence of the degree of semantic overlap. In general, the more same upper concepts which two concepts have, the bigger their semantic similarity is. A function which ranges within positive real numbers was defined to calculate the similarity of two concepts in the same domain ontology base in the paper [10]. But it just considered the calculation based on semantic-distance, paid no attention to the semantic similarity based on properties. People will usually compare the properties of different things in the process of discrimination or connection. If two things have a lot of properties in common, it shows that they are similar, whereas the opposite.

179.4 Improved Calculation of Semantic Similarity

Based on the analysis above, an improved calculation of semantic similarity which combines the characteristics of ontology is discussed in this chapter. The basic idea is that, according to constructing domain ontology, I combine two approaches: distance-based semantic similarity calculation and property-based semantic similarity calculation. With comprehensive consideration, the semantic similarity based on the domain ontology will implement a better web service match to meet users needs.

179.4.1 Calculation of Distance-Based Semantic Similarity

The calculation of distance-based semantic similarity gives the value of the semantic similarity of internal concepts in the domain ontology, primarily considering the semantic distance, the degree of semantic overlap and relevant regulatory factors etc.

- (1) Semantic distance: semantic distance is the length of the shortest relation-chain of all relation-chains between two concepts in the same ontology. $Dis(C_1, C_2)$ represents the shortest length between the concept C_1 and C_2 .

The semantic distance is an important factor to determine the semantic similarity. Usually the longer the semantic distance between two concepts is, the smaller the semantic similarity is.

Definition 4 The similarity formula for the semantic distance:

$$\text{Sim}^{\text{Dis}}(C1, C2) = \frac{1}{1 + \text{Dis}(C1, C2)}. \tag{179.2}$$

$\text{Dis}(C_1, C_2) = \sum_1^n \text{Weight}_i$. C_1, C_2 are any two concepts in the domain ontology base. Weight_i is the weight of the shortest path from C_1 to C_2 .

Definition 5 The formula for the weights between nodes:

$$\text{Weight}(C) = \begin{cases} \frac{1}{\text{Degree}(C)} & C \text{ is the root node} \\ \frac{1}{\text{Degree}(C)} * \frac{\text{Weight}(P(C))}{2} & C \text{ is not the root node} \end{cases}. \tag{179.3}$$

$\text{Degree}(C)$ represents the degree of the concept C , that is, the number of edges starting from C . $P(C)$ represents the parent node of the concept C . $\text{Wight}(C)$ represents the weights of the edge starting from the concept C .

(2) The degree of semantic overlap: that is the number of upper concepts which internal concepts have in common in the ontology. The degree of semantic overlap shows the same degree between two concepts. In the actual calculation, it can be simplified to the number of the public nodes.

Definition 6 The similarity formula for the degree of semantic overlap:

$$\text{Sim}^{\text{Coin}}(C1, C2) = \frac{P(C1) \cap P(C2)}{\text{Max}(P(C1), P(C2))}. \tag{179.4}$$

$P(C_1)$ represents the number of parent nodes of the concept C_1 , that is, the number of nodes from the concept C_1 dating back to the root node. $P(C1) \cap P(C2)$ represents the number of parent nodes which the concepts C_1 and C_2 share. $\text{Max}(P(C1), P(C2))$ represents the bigger number of parent nodes between the concepts C_1 and C_2 .

Considering factors of the semantic similarity based on semantic-distance, the formula for distance-based semantic similarity is defined as follows:

Definition 7 The formula for distance-based semantic similarity:

$$\text{Sim}^1(C1, C2) = \alpha * \text{Sim}^{\text{Dis}}(C1, C2) + (1 - \alpha) * \text{Sim}^{\text{Coin}}(C1, C2). \tag{179.5}$$

α is a regulatory factor. Regulatory factors can adjust according to different applications and requirements.

179.4.2 Calculation of Property-Based Semantic Similarity

In the real world, the process that people distinct different things is always by comparison the properties between the things, if the two things have many of the same properties, then the two things are very similar. Property-based semantic similarity calculation is based on this principle, which to determine the degree of similarity of the two concepts' corresponding properties.

Definition 8 The formula for calculating property-based semantic similarity:

$$Sim^2(C_1, C_2) = \frac{Prop(C_1) \cap Prop(C_2)}{Prop(C_1) \cup Prop(C_2)}. \quad (179.6)$$

$Prop(C_1)$ represents the number of properties about concept C_1 , $Prop(C_1) \cap Prop(C_2)$ represents the same number of properties owned by concepts C_1 and C_2 , $Prop(C_1) \cup Prop(C_2)$ represents the number of all properties owned by C_1 and C_2 .

179.4.3 Calculation of Integrated Semantic Similarity

Integrated the distance-based semantic similarity calculation and the property-based semantic similarity calculation, that is, from definition 2, definition 7 and definition 8, then we can get the formula for the similarity of any two concepts in the same domain ontology base:

$$Sim(C_1, C_2) = \beta * Sim^1(C_1, C_2) + (1 - \beta) * Sim^2(C_1, C_2). \quad (179.7)$$

β is a regulatory factor, because even for the same domain ontology, different experts will distribute different weight for both.

By the similarity between any two concepts ($Sim(C_1, C_2)$), then by definition 3, we can draw the similarity of any two concept vectors. Then we can define the formula for calculating the similarity of input and output:

Definition 9 The formula for calculating the similarity of input and output:

$$S_{ww} = \gamma * S(O, O') + (1 - \gamma) * S(I, I'). \quad (179.8)$$

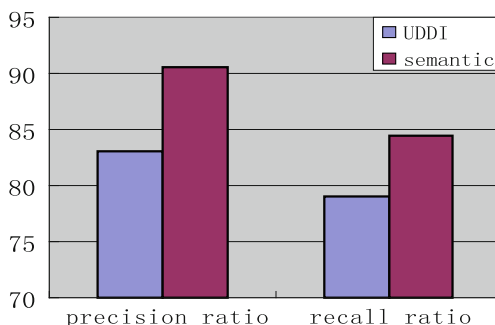
$\langle I, O \rangle$ and $\langle I', O' \rangle$ are the input/output of service W and W' ; $0.5 < \gamma < 1$, because the output is more important for users, and users have a greater control over input, so we want to increase the weight of calculating the similarity of output.

179.5 Experiments

It is better to find that the recall and precision of the web service discovery algorithm are as high as possible, so this chapter will use the recall and precision as the target data. It needs a lot of experimental data and multiple authentication to

Table 179.2 Test data

Numbers Approaches	Based on UDDI	Based on Semantic similarity
The number of returning service sets	158	169
The number of relevant service	131	153

Fig. 179.2 Recall and precision compared

improve the accuracy of experimental results. I downloaded 43 WSDL files describing web service and 157 OWL-S files from the Internet to collect test samples, and then formatted conversion the WSDL file. The service includes security market, train trips query, book query etc. The service requests are made based on UDDI and semantic similarity. The test data is shown in Table 179.2. In the experiment, the parameters are assigned as follows: $\text{matchDegree} = 0.6$, $\alpha = 0.5$, $\beta = 0.5$, $\gamma = 0.65$.

The experimental results show in Fig. 179.2, the recall and precision of the service based on UDDI (83%, 79%) are lower than those based on semantic similarity (90.5%, 84.5%).

179.6 Conclusion

This chapter proposes a semantics-based web service match framework, and then describes the matching approach based on the framework. First of all the service is filtered by service category, and then the web service functional match can be transformed into the calculation of the similarity between any concepts in the ontology base through different definitions. The experimental results are also show that the semantics-based web service match proposed in this chapter has better recall and precision than traditional UDDI match does.

Of course, in many ways, this work need to deepen, such as how to optimize the structure of domain ontology base to improve the efficiency of computing the

concept similarity; how to determine the found services can meet the user' requirements, can be invoked, etc. in particular, the experiment need a lot of experimental data to test.

References

1. Alonso G, Casati F, Harumi K et al (2004) Web services concept, architectures and applications[M]. Springer, New York
2. Berners-Lee T, Hendler J, Lassila O (2001) The semantic web[J]. *Sci Am* 284(5):34–43
3. Martin D, Burstein M, Hobbs J et al (2004) OWL-S:semantic markup for web service. W3C member submission 22 November
4. Roman D, Keller U, Lausen H (eds.) (2004) Web service modeling ontology(WSMO), available at <http://www.wsmo.org/2004/d2/v01/index.html>
5. Ma JG (2008) Web services discovery based on latent semantic approach[C] In: Proc. IEEE Press
6. Martin D et al. (2009) OWL-S: semantic mark up for web services[R]. Damlconsortium <http://www.daml.org/services/owl-s/1.0/owl-s.pdf>
7. Srinivasan N, Paolucci M, Sycara KP (2004) Adding OWL-S to uddi, implementation and throughput[C] In: Proceedings of the 1st international workshop on semantic web service and web process composition San Diego, USA
8. Paolucci M, Kawamura T, Payne TR (2002) Semantic Matching of web services capabilities. In: Proceedings of the first intl semantic web conference
9. Xu D-Z, Wang H-M (2007) Concept semantic similarity research based on ontology[J]. *Comput Engineer Appl* 43(8):154–156
10. Bramantoro A, Krishnaswamy S, Indrawan MA (2005) Semantic distance measure for matching web services. In: WWW2005, web service semantics workshop

Chapter 180

Using Integrated Technology to Achieve Three-Dimensional WebGIS System in Park Planning

Hu Lemin, Wu Qianhong, Yang Li and Cheng Chuanzhou

Abstract In the field of park planning often use CAD system to aided design. What CAD data has considered is model design and modeling, lack of property information. This paper discusses how to use GIS technology to manage and express that contain three-dimensional model data, property information and the surrounding terrain data. And use GIS technology for data management, spatial query and spatial analysis. WebGIS-based system has also becoming a trend in the field of park planning. This paper focuses on the integration of ArcGIS and SketchUp in three-dimensional modeling use at network, and through the Google Earth Web Plugin to publication at last to achieve three-dimensional WebGIS functions.

Keywords WebGIS · Three-dimensional GIS · SketchUp · ArcGIS · Park planning

180.1 Questions

CAD system in the park planning has been extensively, maturely used [1]. CAD focuses on model design and modeling, lacking of property information. I attempted to integrate with CAD system and GIS system. Under the premise of without changing the planners original work way. Using GIS technology to integrate and manage CAD planning data, model data, renderings, attribute data and

H. Lemin (✉) · W. Qianhong · Y. Li · C. Chuanzhou
School of Geosciences and Info-Physics, Center South University, Key Laboratory of Metallogenic Prediction of Nonferrous, Ministry of Education of P.R.China, Changsha 410083, Hunan, China
e-mail: hlm8610@163.com

engineering information. To achieve three-dimensional web publication, spatial query, spatial touring, measurement, and other three-dimensional WebGIS functions. There are three key technical issues below:

- (1) How to achieve CAD system and the three-dimensional WebGIS system data sharing [2, 3]. Although GIS was better than CAD at comprehensive data management and spatial analysis, but the powerful modeling and editing functions of CAD are becoming increasingly close to GIS. For example, you can provide a powerful technical support for GIS at data collection and edit. Park planning's CAD data usually includes general planning data and CAD modeling data. To achieve CAD system and the three-dimensional WebGIS system data sharing, not simply data conversion between the two types of data, but how to achieve organic integration of two types of data, to convenient calling, analysis, demonstration purposes. Need to build integrated spatial database to realize integrated management of planning data, models of spatial data, attribute data, terrain data in different formats, In order to relate associated data for developing advanced spatial analysis functions.
- (2) What is the method used to build and publish Web three-dimensional model, implement three-dimensional Web scene driven. Three-dimensional modeling method of WebGIS [4], must meet the requirements of the following points: First, the three-dimensional model can be show on the Web, visualization; second, to direct the three-dimensional expression and spatial analysis of space objects. The third, can save into the spatial database and be managed.
- (3) To achieve three-dimensional GIS web functions. How to use current Web-3D visualization technique, uses B/S internet mode and HTTP transmitting protocol to realize development of Web-3D [5] visualization to achieve the three-dimensional virtual scenes tour, custom flight path, aerial renderings and other functions in the web. And implement a custom three-dimensional solid object spatial query, advanced spatial analysis functions.

180.2 Problem-Solving

180.2.1 The Three-Dimensional WebGIS System and CAD System Data Sharing

Park Planning CAD system data including the functional structure planning, water supply planning, three-dimensional model diagram and other planning data, stored as dwg/dxf format. While all the existing GIS software provides a platform for migrating data from CAD to GIS, but if all the terminals are installed GIS platform will result in greatly increased planning costs, and requires the designer to control the operation of GIS software. Therefore, asked the system does not change planning staff's context of existing practices to achieve CAD and GIS data

sharing, difficult to solve the problem, including the following points: (1) The solution of differences in the way to describe and the organization of stored spatial object between CAD and GIS. (2) The coordinate system conversion. GIS in the coordinate system defined by the fixed datum and map projection two groups parameters, and AutoCAD support for the world coordinate system (WCS) and the user coordinate system (UCS) two sets of coordinates, the coordinate system is Cartesian coordinate system. (3) The lossless conversion between CAD and GIS data, CAD and GIS data conversion module can be built to convert, but often information is lost, and the expression of object is changed. To achieve data sharing between park planning CAD system and GIS system in the following ways.

- (1) By FME conversion tool, convert the CAD planning data to kml what support Google Earth's network and to shp format what support of ArcSDE spatial database storage. Due to the original data is small, for the loss of information and the object forms of expression changes occur in the conversion process, we can directly use kml editing software or ArcGIS to Manual revision to obtain the desired results.
- (2) Interacting ArcGIS with SketchUp to three-dimensional modeling. To achieve the data sharing between three-dimensional CAD model data and three-dimensional GIS model data.
- (3) Through ArcSDE + MSSQL Spatial database, to address the differences in the way to describe and the organization of stored spatial object between CAD and GIS

180.2.2 Build and Publish Web Three-Dimensional Model

Park planning three-dimensional WebGIS requested directly three-dimensional expression, spatial operations, analysis and management of the spatial objects. Through the three-dimensional GIS-based three-dimensional scene modeling and scene-driven, ultimately realized build and publish Web three-dimensional model.

Mainly with interacting ArcGIS with SketchUp to three-dimensional modeling, such as planning in CAD building model, terrain model, and other auxiliary model. For example, use SketchUp plug-ins provided by SketchUpESRI to achieve interact ArcGIS with SketchUp, And through the base platform for the high set in SketchUp to quickly create three-dimensional architectural model, while Model data will be made into a shape or Geodatabase data to save into ArcSDE spatial database; By calling SketchUp's Sandbox plugin to Automatically generate the terrain model. Google Earth Web plug-ins also can easily interact with SketchUp for modeling, model is main stored as the format kmz, to facilitate web publishing, rapidly virtual reality environment.

Web-based three-dimensional scene-driven adopted Google Earth Web Plugin, to ensure a variety of surface features' three-dimensional expression in the park

planning. Since Google Earth Web plug-in itself does not support the selection of three-dimensional model space objects, but to support the interaction with the kml. So through simulate Kml to three-dimensional space objects, can achieve inter-operate between users and three-dimensional scene. While taking the API that Google Earth Web Plugin provided can improve the efficiency of system development.

180.3 The Specific Process to Achieve Three-Dimensional Web-GIS Function

180.3.1 The Realization of Web-Driven Three-Dimensional Scene and Three-Dimensional Model Visualization

- (1) Environment to build: (1) Loading Google Earth Web Plugin API; (2) To create a DIV element that contains the plug-in, create a function to initialize; (3) After the page loads, call the initialization function.
- (2) Load the specified three-dimensional model data, terrain data, etc.; Including loading Google Earth provides the administrative boundary layer (LAYER_BORDERS) [6], three-dimensional architecture (LAYER_BUILDINGS); road and the name of the layer (LAYER_ROADS), make the scene realistic.
- (3) Load custom three-dimensional model data. At the same time using three methods to load custom data model:
 - (1) KmlNetworkLink: Loads a kml or kmz file from a specified URL. The fetched KML can then be added to the plugin just as any other object, using `ge.getFeatures().appendChild()`.
 - (2) FetchKml: Also loads kml from a url, but returns a KmlFeature object representing the root kml feature. The object's kml DOM can be accessed and updated before it is appended to the plugin's kml DOM.
 - (3) ParseKml: Takes a kml string, and also returns a KmlFeature object. As with `fetchKml`, the returned object's KML DOM can be accessed.

Each method is described in more detail below. Fig. 180.1

180.3.2 The Realizations on Multiple Three-Dimensional Scene Sightseeing Modes

The realization on ramble: Open the ramble toolbar by putting the code on navigation into service.

The code is: `ge.getNavigationControl().setVisibility(ge.VISIBILITY_AUTO);`



Fig. 180.1 Load custom three-dimensional model data, and terrain data rendering in the Google Earth plug

Customizing the sightseeing route: Google Earth Web plug-in function do not allow users to play by coding sightseeing freely [7]. First is to automatic generate the temporary kml file through the tour route customized by users, then play the users' freely customized sightseeing by `ge.getTourPlayer()`. `Play()`. At the same time, the system customized multiple the best tour routes available to the user to choose

The control on shot: Using API provided by Google Earth Web plug-into achieve the control function on shot, such as obtaining view, translation, inclination, flight. And the view of different locations is obtained by using `Camera way` to define the reviewers' space position dynamically. Using `LookAt` to realize specify the location's arbitrary angle views

The level limit on model fine display: Adopting nested regional method, appointing LOD (level of detail) limit, to realize model fine display by grade, according to the different distance between user's perspective and space object. When the distance is small, activating (LOD) area, selecting the higher display level automatically, so that the details of the model can be observed.

180.3.3 Three-Dimensional Spatial Query and Measurement

Three-dimensional Spatial Query [8, 9]: The main pattern is graphic-attribute mutual query. Users can obtain the corresponding space object through keyword, or the attribute information by clicking the existed space objects. Since Google Earth Web plug-in itself does not support the selection of three-dimensional model space objects, in order to interact, those steps can be taken one by one: First, Use

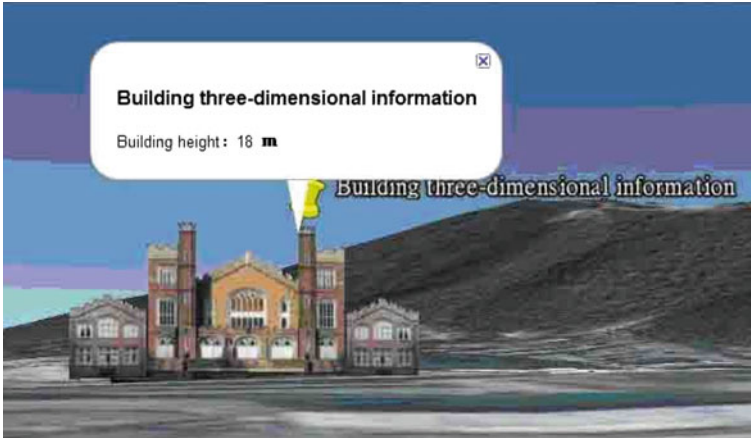


Fig. 180.2 Three-dimensional spatial query graph

kml's advantages to select objects, the three-dimensional objects have the identical IDs (identifier, unique number randomly generated by the database engine) with the corresponding spatial data stored in the geodatabase; Second, adopt the method to code kml simulate three-dimensional objects to realize the arbitrary choice of three-dimensional objects; Third, achieve to relate the selected objects and the corresponding attribute data by storing the corresponding ID in ArcSDE spatial database, program programming to obtain the corresponding attribute data, and return it through kml; What is more, display attribute data on the screen through the ball box. Fig. 180.2 is the result of point-select.

Three-dimensional measurement interactively: The function is realized by the calculation of arbitrary line's node coordinate in three-dimensional graphics. Transform any nodes' two-dimensional screen coordinates into Three-dimensional geographic coordinates on the line. The length of the segment calculated as formula (180.1), the distance between two adjacent points on the line is calculated as formula (180.2). Fig. 180.3

$$L = \sum_{i=1}^{n-1} Di \quad (180.1)$$

$$Di = \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2 + (z_{i+1} - z_i)^2} \quad (180.2)$$

180.3.4 Other Three-dimensional Spatial Analysis Function

Three-dimensional spatial analysis include slope and aspect, volume calculation and visible analysis.

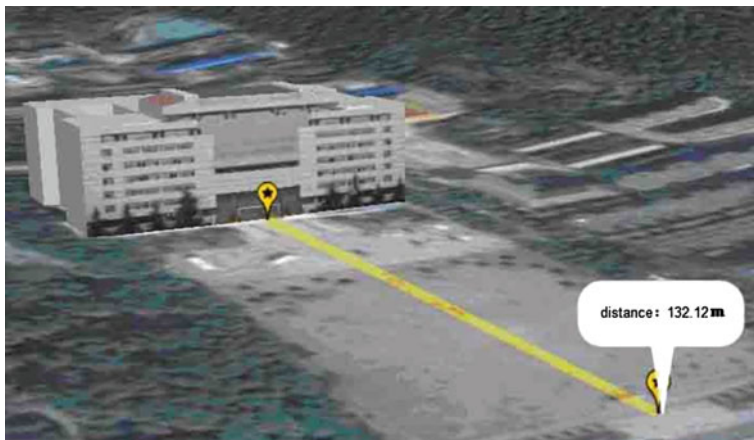


Fig. 180.3 Two-dimensional distance measurement

Slope and aspect at first choose an algorithm such as Kriging method, inverse distance method, Minimum curvature method, etc. Using selected algorithm, points with elevation are then spatially interpolated to directly obtain DEM of stratum surface. Then calculated by using the calculation model on slope provided by ArcGIS through spatial objects'ID selected by users and the calculation result generated by code behind, and publishing it to Google Earth Web inserts in kml form.

Volume calculation is done through the compilation of spatial analysis module. The corresponding length of the line is calculated through the node coordinates of three-dimensional spatial object obtained by three-dimensional measurements of the nodes, and volume calculation is done automatically by volume formula.

Implementation of the visible analysis deals with the visible area, the intersection with multi-point visible area, the calculation of the area blocked, and the calculation of the location and the high degree of the visible port etc. The ID of homogeneous spatial objects is adopted through the compilation of spatial analysis module to calculate the results of the visible analysis, and the results are published in Google Earth Web Plug-in by the way of kml.

180.4 Conclusion

- (1) Integration between the ArcGIS and SektchUp technology implements interactive three-dimensional modeling, and the Google Earth Web Plug-in completes scenario-driven and network establishment and release for three-dimensional model.

- (2) By ArcSDE + SQL, spatial database is established to achieve the CAD data and GIS data integration management. On the basis of the establishment of the shared spatial database three-dimensional objects are simulated by kml, and the Google Earth Web Plugin implements the three-dimensional interactive graphic-attribute mutual query.
- (3) Three-dimensional spatial analysis of WebGIS is implemented by using Google Earth Web Plug-in as the platform, ArcSDE spatial database as the basis, kml/kmz file as an intermediary, and ArcGIS as a background analysis or your own spatial analysis module.
- (4) With the Google Earth Web Plugin constantly updated, more API opened and three-dimensional modeling extended, the three-dimensional WebGIS will play an increasingly important role in the park planning through the integration between CAD systems and GIS systems.

Project support: National Natural Science Foundation of China, "Digital Mine" construction on ZhiJiaDi district, Shanxi Province.

Project Number: 40473029/2010002113008.

References

1. Zhu Q, Wu B, Zhong Z (2006) Integration of 3-D GIS and highway CAD [J]. *China J Highw Trans* 19(4):1–6
2. Xiao L-B, Zhong E-S, Liu J-Y (2001) A discussion on basic problems of three-dimensional GIS[J]. *J Image Gr* 9:842–845
3. Ai L-S (2004) Research on the application of the three-dimensional visualization GIS in city planning[D]. Tsinghua University, BeiJing
4. Shan N, Kuang M-S, Li Y-G (2009) Study on three-dimensional GIS development technology based on SketchUp and ArcGIS [J]. *Railw Comput Appl* 18(05):14–17
5. Zhu X, Wu H (2006) The realization of WebGIS based on Google Earth[C]. In: Proceedings of the fourth symposium on GIS development on cross-strait and the tenth annual meeting of China GIS association
6. Chen P, Meng LK, Song Y (2007) R-tree structure appended with spatial topology restrictions in three-dimensional GIS [J]. *J Wuhan Univ* 32(4):347–349
7. Zhou Q (2003) Key issues for data transformation from CAD to GIS [J]. *J Wuhan Univ* 36(3):64–67
8. <http://code.google.com/intl/zh-CN/apis/earth/documentation/index.html>
9. <http://code.google.com/intl/zh-CN/apis/kml/documentation/>

Chapter 181

Hierarchical Base-k Chord Based on Semantic Networks

Huayun Yan and Deqian Xue

Abstract Chord is a famous structured peer-to-peer networks. There are many variants which change the structure and the routing algorithm of Chord, and all these variants achieve higher routing efficiency than those of standard Chord, one variant of Chord is called Base- k Chord. To improve the routing efficiency of Base- k Chord, this chapter proposes a structured model for Base- k Chord by using hierarchical structure, which have two layers, one is called local layer (like the general layer of Base- k Chord),the other is called global layer which composed by supernodes (it is a semantic networks layer).Both the two layers of hierarchical Base- k Chord are composed by general structure of Base- k Chord. The routing algorithms of our model based on the base- k Chord's, that is, the routing method is the same as the base- k Chord's when the search key include in local area; otherwise, the node need transfer the search message to a related supernode of semantic layer, finally search the key in corresponding local area. The experimental results show that our model has better routing efficiency than its counterparts.

Keywords Peer-to-peer(P2P) · Distributed hash table(DHT) · Chord · Base- k Chord · Semantic networks

H. Yan (✉) · D. Xue
School of Information and Engineering, Huzhou Teachers College,
Huzhou 313000, China
e-mail: yanhy@hutc.zj.cn

D. Xue
e-mail: dqxue@hutc.zj.cn

181.1 Introduction

Peer-to-peer(P2P) systems are distributed systems without any centralized control, where the software running at each node is equivalent in functionality, that is all the nodes playing the server and the client in a peer-to-peer networks, it includes some characters in P2P networks, such as load balance, decentralization, efficient location, etc. There are two classes of P2P systems: The first class called unstructured P2P systems, the most popular are the file sharing systems. In general, sharing other resources such as Napster [1], scalability has been recognized as the primary challenge in designing unstructured P2P systems. The second class called structured P2P systems, to obtain a scalable system, structured P2P systems are based on distributed hash table (DHT) schemes. In DHT schemes, the basic idea is to associate objects with unique identifiers as their keys. Each participating node also has a unique identifier. There is a globally known function that maps object keys to node identifiers so that objects are inserted into the network by placing them (or their indexes) in the nodes responsible for their keys. So, locating an object becomes a routing problem from the requesting node to a destination identifier. To route messages, each node maintains some information about other nodes as guide to forward queries. Such DHT P2P system include Chord [2], CAN[3] etc.

Chord is the best known structured P2P protocol. It provides support for just one operation: given a key, it maps the key onto a node. Data location can be easily implemented on top of Chord by associating a key with each data item, and storing the key/data item pair at the node to which the key maps.

In Chord, a node needs “routing” information about only a few other nodes, because the routing table is distributed, a node resolves the hash function by communicating with a few other nodes, each node maintains information only about $O(\log_2^N)$ other nodes, and resolves all lookups via $O(\log_2^N)$ messages to other nodes. Therefore, Chord adapts efficiently as nodes join and leave the system, and can answer queries even if the system is continuously changing. It addressing the scalability problems of unstructured P2P systems. Chord includes the same features of structured P2P systems, such as load balance, decentralization, scalability, availability, flexible naming etc.

There are many variants on Chord, one kinds of flat variants aiming to the structure of the Chord,such as Base- k Chord [4, 5], F-chord [5],we presented asymmetrically bidirectional Base- k Chord [6] and symmetrical bidirectional Base- k Chord [7], all these kinds of variants on Chord are flat structures. All these flat variants of Chord can improve the routing efficiency. There are another variants on Chord called hierarchical Chord, such as Chord² [8] etc. All these variants cannot provide semantic attribute.

A semantic network is a network which represents semantic relations among concepts. This is often used as a form of knowledge representation. It is a directed or undirected graph consisting of vertices, which represent concepts, and edges. Literature [9] present statistical analyses of the large-scale structure of semantic

networks with graph idea. From this chapter, we know that all the keywords can combine a semantic networks, but we cannot efficient use this semantic work in P2P networks. We present a modified base- k Chord called hierarchical Base- k Chord to solve the semantic and stabilization problem of Base- k Chord. The core idea of this chapter: the flat structure is changed to hierarchical structure of Base- k Chord. Concretely, we present a hierarchical Base- k Chord model based on the Base- k Chord, there are two virtual layer of our model, the top layer is decided by semantic content (combined by Base- k Chord), the bottom layer is a general Base- k Chord (more details see Sect. 181.3) and we choose a steady node in some semantic area of bottom layer as a supernode on top layer and we presented a new routing algorithm, which always chooses the local finger table firstly, else transmit the routing message to the supernode of semantic layer.

The rest of the chapter is organized as follows: Sect. 181.2 presents the related work of Base- k Chord. Sect. 181.3 presents hierarchical Base- k Chord and related algorithms. Sect. 181.4 is the simulations. Sect. 181.5 is the conclusions.

181.2 Related Work

181.2.1 Base-k Chord

Standard Chord can be seen as a virtual ring of N nodes labelled from 0 to $N-1$. The edges, representing the overlay network, the finger table is from identifier x to identifier $x + 2^i$, for each $x \in \{0, 1, \dots, N-1\}$ and $i < \log_2 N$. The degree (finger items value) and the diameter (the largest routing jump) are $\log_2 N$, the average path length is $(\log_2 N)/2$ in Chord.

Because the degree of Chord is $\log_2 N$, when N is one million, the degree of Chord only reach to 20, that is too small. Thus we can expand the finger table (increasing the degree of standard Chord) to get better routing efficiency. Recently, many variants is appeared on Chord.

Yan [6] presents a variant called F-Chord based on the standard Chord, it introduces Fibonacci sequence numbers as the finger table to substitute the original sequence numbers of Chord. F-Chord can be seen as a virtual ring of N identifiers labelled from 0 to $N-1$. The edges, representing the overlay network, go from identifier x to identifier $\text{Fib}(i)$, where $\text{Fib}(i)$ denote the i th Fibonacci number. The degree and the diameter are below $\log_2 N$, the average path length is below $(\log_2 N)/2$. So the routing efficiency of F-Chord is more efficiency than the original Chord's.

Chiola [5] introduces the base- k sequence numbers to substitute the sequence numbers of standard Chord, details see the structure of Base- k Chord (where the k is 3) in Fig. 181.1. Base- k Chord can be seen as a virtual ring of N identifiers labelled from 0 to $N-1$.

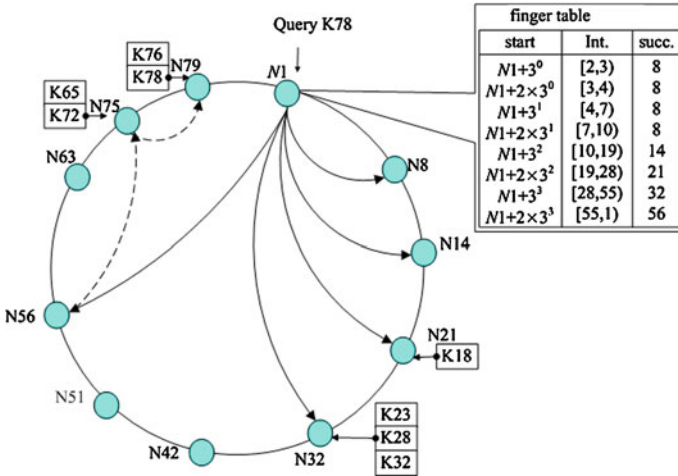


Fig. 181.1 Structure of a Base- k (where $k = 3$) Chord (including a finger table of node 1 and a routing query of node 1)

For any $k > 2$ for all the $l = 0, 1, \dots, \log_k N - 1$, and all the $i' = 0, 1, \dots, k - 2$, the hops of finger sequence are as follows:

$$J(i) = (i' + 1)k^l \quad (i = (K - 1)l + i') \tag{181.1}$$

Where $J(i)$ denotes the hop distance of the $(i + 1)$ th finger.

Figure 181.1 denotes a Base- k Chord structure, the finger table of node 1 and the query process of node 1, where the k is 3, the node space is 81 (that is $N = 81$), we can see it need 3 jumps when node 1 query key 78.

The degree is $(k-1) \log_k N$, the average path length is $\frac{k-1}{k} \log_k N$ on the Base- k Chord. Otherhand, we know the degree is $\log_2 N$, the average path length is $\log_2 N / 2$ on the standard Chord. So the routing efficiency of Base- k Chord is more efficient than the standard Chord's. We can see Chord as a special example of Base-2 Chord, where $k = 2$. In Base- k Chord, the start sequence (the first column of the Base-3 Chord in Fig. 181.1) is different from the sequence of the Chord's. The other aspect of the Base- k Chord is same as the Chord's. The procedure of involving finger table must be changed correspondingly, such as the procedure of find-successor, find-predecessor and routing algorithms etc.

Although the Base- k Chord can improve some routing efficiency of standard Chord, there are some problems to resolve. Firstly, there is no semantic element in flat Chord search. In Fig. 181.1, we can see that all the identifiers of nodes are created by a hash function, and hash function is a random function, thus all the identifiers are random values, thus there is no semantic element in flat Chord search. But we know that the search includes semantic element in P2P networks. Secondly, cause Chord is a dynamic networks, there is a churn problem in Base- k Chord.

181.3 Hierarchical Base-k Chord

Aim to the above two problems, we presented a hierarchical Base- k Chord based on semantic.

181.3.1 Structure of Hierarchical Base-k Chord

Figure 181.2 is the structure of hierarchical Base- k Chord. The nodes are classed two types in hierarchical Base- k Chord. One called general nodes, such as the nodes $N8$, $N14$, $N32$, etc. the general nodes just like the standard Chord and Base- k Chord's nodes. The others called supernodes, each supernode is a agent of the area itself, thus the supernodes must be steady nodes, every area represent a semantic types, such as the nodes $S1(N1)$, $S51(N51)$, $S79(N79)$. Supernodes are organised through a Base- k Chord ring. Supernodes and general nodes have two finger tables, one is local finger table, the other is global finger table, the global finger table of general nodes are copied through the supernodes, that is when the global finger table of a supernode is variety, the supernodes must broadcast in his local area.

In hierarchical Base- k Chord, nodes identifier is composed by two parts, the first part is the area identifier, the second part is the local identifier, all the identifier are created by hash functions, such as the MD5 or SHA function. For example, in Fig. 181.2, the identifier of $N1$ is $\text{hash}(\text{news})\|\text{hash}(\text{node's ID})$, where the node's ID is nodes' MAC addresss or IP address, this can ensure the node's ID is a exclusive value. Thus $\text{hash}(\text{news})$ is the area identifier, that is $A1$; and $\text{hash}(\text{node's ID})$ is node's local identifier.

When a content's owner want to publish the content to a supervisory node, it must confirm the identifier of the supervisory node. Firstly, we choose a keyword of the content, and decide the keyword belong to some semantic word, then the identifier is confirmed by $\text{hash}(\text{semantic word})\|\text{hash}(\text{keyword})$. The $\text{hash}(\text{semantic word})$ decide the semantic area (we call it semantic area ID), and $\text{hash}(\text{keyword})$ decide some node of this area (we call it as local ID). Cause there exist a $\text{hash}(\text{semantic word})$ part, we can search by semantic. There is a problem that the semantic networks is a graph, not a tree, that means some keyword belong to multi semantic words. In this case, we'll publish such content to multi semantic area through the value of $\text{hash}(\text{semantic word})\|\text{hash}(\text{keyword})$ with each semantic word.

When a node want to search a content through a keyword, it must determine the keyword's semantic word firstly (that is the keyword belong to some semantic word, we will choose one as the semantic word if there exist multi semantic words), then we'll compute the value of $\text{hash}(\text{semantic word})\|\text{hash}(\text{keyword})$, then the node decide wether forward the search message to corresponding supernode, if the $\text{hash}(\text{semantic word})$ is the local area's name, then search in local finger table; else forward the search message to corresponding supernode.

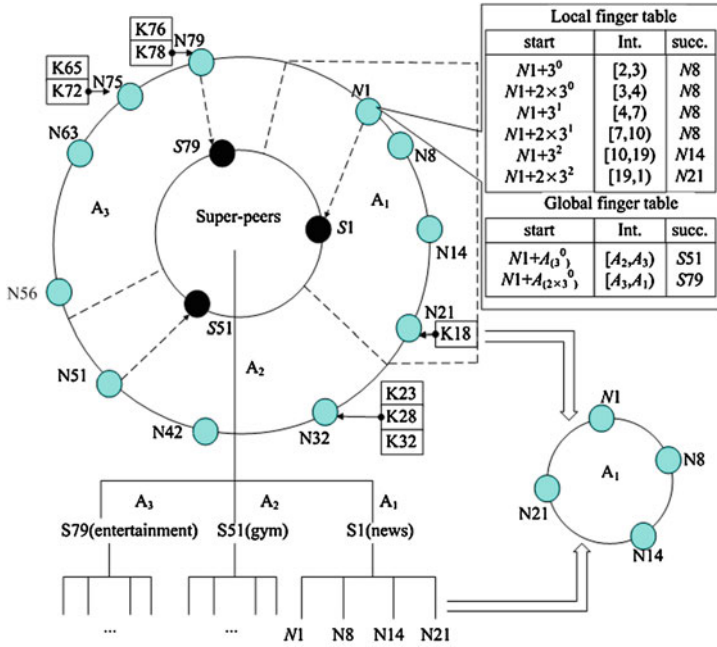


Fig. 181.2 Structure of hierarchical Base- k Chord (where $k = 3$, including two layers, the inner ring is a semantic layer, the outer ring is a general Base- k Chord ring, and the bottom hierarchical structure is a map of the two layer rings)

We modified the join and delete procedure based on the structure changes, cause the join and delete procedure are similar to the Base- k Chord's, we omitted here. We will discuss the finger tables and the routing algorithm of our model following.

181.3.2 Finger Table

In hierarchical Base- k Chord, suppose that the threshold total number of nodes is N , suppose that the semantic areas number is M . We called all the semantic area as A_1, A_2, \dots, A_M , the range of nodes identifier on $A_i (1 \leq i \leq M)$ is from $(i-1)N/M$ to iN/M .

There are two types finger table in hierarchical Base- k Chord. General nodes only have local finger table, supernodes have local and global finger table.

Definition 1 The jump sequences in hierarchical Base- k Chord's local finger table as following:

For any chosen base $k \geq 2$, for all $l = 0, \dots, \log_k N/M - 1$, for all $i' = 0, \dots, k-2$

$$J_{local}(i) = (i' + 1)k^l \quad (i = (K - 1)l + i') \tag{181.2}$$

Where the $J_{local}(i)$ denotes the hop (also called jump) of the $(i + 1)$ th *item* in local finger table from the considered node. □

In Fig. 181.2, according to the Definition 1, where the $k = 3$, suppose the hierarchical Base- k Chord containing nodes number N is 81, and semantic areas number M is 3, thus the semantic areas A1, A2, A3 are [0,26], [27,53], [54,80] separately. The local finger table items of j th area in hierarchical Base- k Chord can describe as following:

$$finger_local_N[i] = N_{A_j} + J_{local}(i) \bmod \left(\frac{(j - 1)N}{M} \right) \tag{181.3}$$

Where $finger_local_N[i]$ is the i th finger item of local finger table on node N_{A_j} , the N_{A_j} is a node of j th area A_j in hierarchical Base- k Chord. Expression (181.3) denotes the finger items are organized a local ring based on the module operation.

Definition 2 The jump sequences in hierarchical Base- k Chord’s global finger table as following: For any chosen base $k \geq 2$, for all $l = 0, \dots, M - 1$, for all $i' = 0, \dots, k - 2$

$$J_{global}(i) = (i' + 1)k^l N / M \quad (i = (K - 1)l + i') \tag{181.4}$$

Where the $J_{global}(i)$ denotes the hop (also called jump) of the $(i + 1)$ th *item* in global finger table from the considered node.

In Fig. 181.2, according to the Definition 2, there are three semantic areas A1, A2, A3, and S1, S51, S79 are the supernodes of each areas.

The global finger table items in hierarchical Base- k Chord can describe as following:

$$finger_global_N[i] = S + J(i) \bmod (N) \tag{181.5}$$

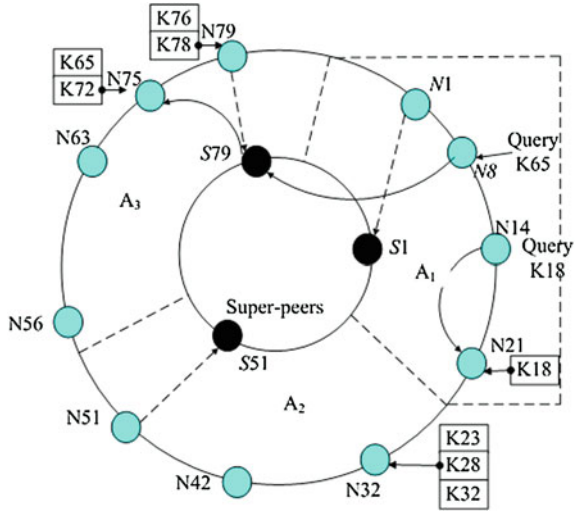
Where $finger_global_N[i]$ is the i th finger item of global finger table on node S, the S is a supernodes in hierarchical Base- k Chord. Expression (181.5) denotes the finger items are organized a global ring based on the module operation.

181.4 Routing

Figure 181.3 denotes the routing process.

In the routing figure, consider the two layers structure character of hierarchical Base- k Chord, we modify the routing algorithm of the Chord. The routing algorithm include two parts, one is local routing, the other is global routing. Firstly, node search a *key* in lcoal area when the *key* include in the local area’s semantic. The local search just like the N14 search the *key* 18 in Fig. 181.3. Secondly, if the query key do not include in local area, the node need transfer the

Fig. 181.3 Routing on hierarchical Base-*k* Chord (where the *k* is 3, there are two types, the local routing, just like the query K18 on N21, and the global query, just like the query K65 on N8)



search message to corresponding supernode S_j , then S_j search the key in its local area. The global search just like the N8 search key 65, N8 found that key 65 don't include in area A1 first, then N8 transfer the search message to S79, then S79 search the key 65 in area A3.

181.5 Simulations

In this section, we evaluate the relationship between the average number of hops and the nodes size, and the value of *k* how to influence the average number of hops with a fixed number of nodes. We consider the interests of nodes query, that is, a node will search more queries in the area itself.

In this section, we set ten semantic areas in our model.

In Fig. 181.4, for hierarchical Base-*k* Chord(where the *k* is 3),we consider the query percent in its own area, there are four cases: HB3(10 percent) denotes the query percent in its own area is 10%, and others are likes. In Fig. 181.4, B3 denotes the result of Base-*k* Chord.

From Fig. 181.4, we can see that all the average hops value of hierarchical Base-*k* Chord are less than the Base-*k* Chord's, and the average hops value of hierarchical Base-*k* Chord is decreasing with the query percent of its own area is increasing. Thus we can say that our hierarchical Base-*k* Chord get more efficiently than Base-*k* Chord, and the effect is more distinctness with the query percent in its own area is increasing.

In Fig. 181.5, we expand the *k* value from 3 to 4 and 5, and we compare hierarchical base-*k* Chord and base-*k* Chord. There are three cases in hierarchical base-*k* Chord and base-*k* Chord each, B3 denotes the result of base-3 Chord, and

Fig. 181.4 Average number of hops as a function of nodes sizes (HB3 denotes the case of hierarchical Base-k Chord, B3 denotes the case of Base-k Chord, where the k is 3, the percent denotes the local query ratio on hierarchical Base-k Chord)

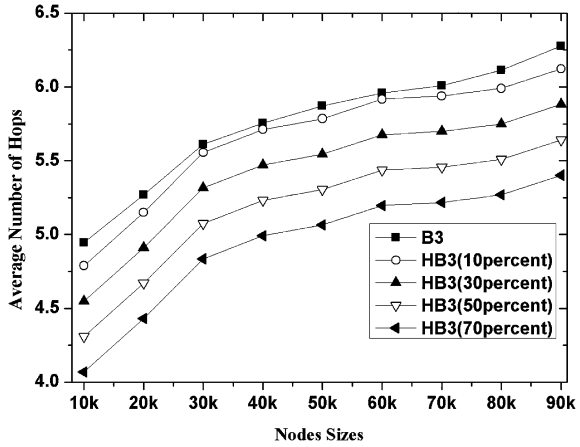
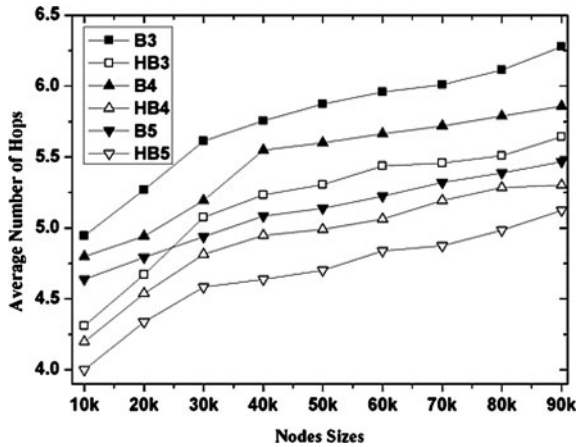


Fig. 181.5 Average number of hops as a function of nodes size (HB3, HB4 and HB5 denotes the case of hierarchical Base-k Chord, B3, B4 and B5 denotes the case of Base-k Chord, where the k is 3, 4 and 5 respectively)



HB3 denotes the result of hierarchical base- k Chord, where k is 3, and the query it is own area percent is 50 percent, the others symbols are likes.

In Fig. 181.5, we can see that all the results of ours are less than the base- k Chord's, that means our model get more efficiently than base- k Chord, and the effect is more distinctness with the k value is increasing.

We will analysis the simulations results through theoretic analysis following.

We suppose that the nodes number is N in hierarchical base- k Chord, and there are M semantic area in the hierarchical base- k Chord, thus the average nodes number is N/M in every local area. We suppose that the local routing ratio is R . We know that the average jumps of base- k Chord is $k - 1/k \log_k N$ in a base- k Chord with N nodes. Thus we can estimate the average jumps is $k - 1/k \log_k N/M$ in local

area, and the average jumps is $k - 1/k \log_k M$ in semantic area (the supernodes ring), thus we can get the average jumps of our model as the following expression:

$$\begin{aligned} R \frac{k-1}{k} \log_k N/M + (1-R) \left(\frac{k-1}{k} \log_k M + \frac{k-1}{k} \log_k N/M \right) \\ = \frac{k-1}{k} \log_k N/M + (1-R) \left(\frac{k-1}{k} \log_k M \right) \end{aligned} \quad (181.6)$$

In the left part of expression (181.6), the first part denotes the local routing case, the second part denotes the global routing case, we can see that the value of expression (181.6) will decrease with the ratio R increasing, that mean the routing efficiency will be improved with the ratio R increasing. In expression (181.6), we can easily prove that the expression is a monotonicity decreasing function when the k is not less than 3 through derivative function, that mean the routing efficiency will be improved with the parameter k increasing. And all these have validated in our simulations.

181.6 Conclusions

In this chapter, we present a hierarchical base- k Chord, we changed the structure of Base- k Chord and we introduce semantic networks in our system. The semantic networks is a graph, thus there exist some keywords belong to multi semantic words. We will publish all these cases through hash(semantic word)||hash(keyword), on the other hand, we will choose only one semantic word when a node search such keyword. We changed the routing table and algorithm correspondingly. Our model get more efficiently than base- k Chord through evaluations.

Our model only has two layer, we need study the effect if we expand our model to three or more layers in future.

Acknowledgments This research was partially supported by National Natural Science Foundation of China under grant 60872057 and 60803053, and by Zhejiang Provincial Natural Science Foundation of China under grant R1090244, Y107293 and Y1080212, and by Zhejiang Provincial Scientific and Technological Project of China under grant 2008C21083 and Huzhou Municipal Scientific and Technological Project of China under grant 2008GG11.

References

1. Napster. www.napster.com
2. Stoica I, Morris R, Karger D, et al (2001) Chord: a scalable peer-to-peer lookup service for internet applications. In: Proceedings of SIGCOMM San Diego, California, USA
3. Ratnasamy S, Francis S (2001) etc A scalable content-addressable network. In: Proceedings of ACM SIGCOMM San Diego, California, USA, pp 161–172

4. Chiola G, Cordasco G, Gargano L et al (2008) Optimizing the finger table in chord-like DHTs[J]. In concurrency and computation: practice and experience. Wiley, pp. 643–657
5. Cordasco G, Gargano L, Negro A et al (2008) F-Chord: improved uniform routing on Chord[J]. In networks Wiley, pp 323–330
6. Yan HY, Guan JH and Zhan WH, Jiang YL (2010) An asymmetrically bidirectional base- k Chord[J]. Telecomm Sci 26(10):71–79
7. Yan HY, Guan JH, Jiang YL (2009) Symmetrical bidirectional Base- k chord and its interesting character. The 5th international conference on semantics, knowledge and grid, Zhuhai, pp 371–375
8. Joung YJ, Wang JC Chord² (2007) A two-layer Chord for reducing maintenance overhead via heterogeneity. Comput Netw 51(3):712–731
9. Steyvers M, Tenenbaum JB (2005) The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. Cognit Sci 29:41–78

Chapter 182

Enterprises Application Integration Framework Based on Web Services and Its Interfaces

Zhangbing Li, Wujiang Che and Jianxun Liu

Abstract Traditional technologies for enterprise applications integration (EAI), such as CORBA and EJB/DCOM component or middleware, can not be handily extended in the systems for dynamic business requirements. To solve the problem, using the principle of service-oriented architecture (SOA), an EAI framework based on web services and its interfaces are designed. The framework is extensible to dynamic business requirement, and achieves well heterogeneous datum sharing and integration of application systems in enterprises.

Keywords WEB service · Enterprise applications integration · Heterogeneous system · SOAP · Interface

182.1 Introduction

With going deep into informatization, now a days most enterprises have designed much more application systems, such as enterprise resource planning (ERP), product data management (PDM), supply chain management (SCM), customer relationship management (CRM), and so on[1, 2, 3–6]. These application systems just have been made based on their functions, such called “independent solved project”. Because of the different development languages, platforms, communication protocols, they have clear boundaries and weak interaction of each other, So

Z. Li (✉) · W. Che · J. Liu
School of Computer Science and Engineering, Hunan University of Science
and Technology, Xiangtan, China
e-mail: LZB_XT@126.COM

Z. Li · J. Liu
Key Lab of Knowledge Processing and Networked Manufacture, Hunan
University of Science and Technology, Xiangtan, China

they have formed heterogeneous data resource of enterprises and caused “Isolated Information Island”, the information exchange of the inner or outer enterprise has been blocked in a certain extent [3, 4, 6–9]. With the more business requirement in enterprises, these application systems need to visit the data resource each other by network. So enterprise applications integration (EAI) is urgently needed for enterprises to reconstruct their business flows and offer faster, more convenient and better services to user. EAI is an integration of hardware, software, protocol and business process, which can realize integration to meet business requirement, realize information exchange and share among application systems in enterprises [3, 6–9].

This chapter analyses the characteristics of traditional EAI and the principle of web services, and gives a design of EAI framework and its interfaces for enterprise heterogeneous application system by using of web services and SOA component.

182.2 Analysis of the Traditional EAI

It is a great challenge to implement the sharing and integration for information and software resource under heterogeneous environment. There are two main problems for EAI of heterogeneous data resource [3, 7]. The one is the universal problem of heterogeneous data integration such as the conflict of data’s heterogeneity, integrity, performance, and semantic. The other is specific problem such as permission bottlenecks, additional restrictions and integrating content (business logic). Therefore, the global model offering by heterogeneous data integration system must satisfy as following: (1) It can describe variety of data format no matter whether it is structured or semi-structured, no matter whether it can support for all inquire language or simple text inquire; (2) it must be easy for delivering and exchanging variety of data.

182.2.1 Classification and Characteristics of EAI

EAI has different classification from different points of view. EAI has been classified to oriented-data integration and oriented-procedure integration according to the object of application integration. In terms of using the tool and technology of application integration, EAI has also been classified to the following types [1, 4, 6]:

- (1) User interface integration, which is a user-oriented integration by replacing the original terminal window and PC graphical interface with a standard interface such as a browser.
- (2) Data integration, which occurs between the database and data sources in enterprise and has been realized through the data migration from a data resource to another one. Data integration is a common form in current EAI solutions.

- (3) Business flow integration, which generates in business flows spanning multiple applications, and presents the integrated features by use of some high-level middlewares.
- (4) Function or method integration, which is the cross-platform applications in the network to an application integration (A2A), includes direct integration and strict integration. Generally it is interactive request-response mechanism based on the client (request programs) and server (response programs).

182.2.2 Traditional Methods of EAI

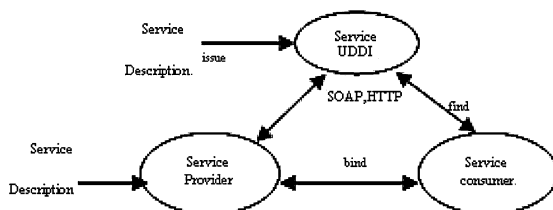
The traditional EAI mainly uses distributed technology such as point-to-point integration, message middleware, CORBA, DCOM, RMI and so on. But these kinds of integration model have its disadvantages: high degree of coupling, bad security and weak flexibility [1, 5, 8].

At present there are two methods of realizing heterogeneous database integration. The first method is data migration, that is, original data is migrated into the new database management system after it has been changed into new data type. Many relational database systems have offered such similar function. But the disadvantage of this method is that the original application software is abandoned or is redesigned for adapting to the new database management system with the data updating or renewing. Therefore, data migration method is not a practical project. The second method is integration by component or middleware. This method need not to change storing pattern and management of original data. The component or middleware locates at between heterogeneous database system (data layer) and integration application program (application layer), it corresponds database systems downwards meanwhile offers upwards general interface of uniform data model and data visiting for the applications visiting integration data [1, 2, 7–9]. The database applications system can independently complete their task, and the component or middleware mainly offers high-level services for searching heterogeneous data resources. But the component development is difficult for program design according to different business requirements. Although middleware, as an integration interface which must offer a global data model, can be purchased instead of programming, it is hard to adapt to dynamic business requirement of enterprise.

182.3 Design a Framework and its Interfaces of EAI Based on WEB Service

Web services can realize distributed service. First, client produces SOAP request message by WSDL, then transmits it to service offer by HTTP (FTP or STMP). At last service offer makes responses to service require and return result of SOAP to

Fig. 182.1 The architecture of WEB services



client [4, 5, 10]. So the technology of web service can be used for enterprise applications integration.

182.3.1 Architecture of WEB Service

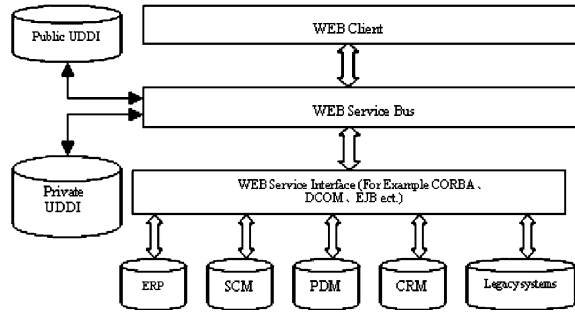
The Web service is a service-oriented architecture (SOA). WEB service architecture has three roles and three operations as shown in Fig. 182.1. Three roles are as following: (1) service offer, which offers and delivers services and makes response to requires of requestor; (2) service register center, which is a database of storing and describing service information. It is a medium between service offers and requestors. Service offers deliver their information and service for requestor searching and get binding information of service by this medium; (3) Service requestor, which is the client application of searching and using service. They look for services by service agent (UDDI) and then uses the service by binding it. Three operations be used among these roles as following: (1) Issue, which is that the service offers issue service descriptions into register servers, and then register their functions and interfaces towards service agent. (2) Find, which is that the service requestors query and find specific services with relation information by service broker. (3) Bind, which is that the service information, include access path, call parameter, return result, transport protocol, safety demand, ectal, is bound to the service to make the service requestors using those services truly.

Web services work in backstage of web server such as IIS. The SOAP requirement which is generated by web services caller has been embedded in a requirement of HTTP POST. Then it is sent to web server. These requirements are transmit to web service requirement processor by Web server. The processor analyzes the SOAP requirement and call Web services. Then the correlative SOAP response is made. After getting SOAP response, web server returns it to client by HTTP response [5, 10–12].

182.3.2 Design of a EAI Framework Based on WEB Service

The EAI framework based on web service is designed as shown in Fig. 182.2. It encapsules and integrates all kinds of heterogeneous application program and

Fig. 182.2 The framework of EAI based on WEB service



information system above-mentioned in manner of WEB services interface as a service offer. During integration, all kinds of logic components, application programs and message are packed, assembled, and integrated. The integration is publicized by web services interface.

The framework model is divided into four layers: customer layer, web service layer, business logic layer and data layer. Each layer has its functions as follows:

- (1) Customer layer: It is similar to traditional architecture of Browser/Server and Client/Server, the main function of customer layer is showing result to customer. The web browser is adopted as client main interface because it is widely used, and the other enterprise server or application program can exchange data by use of this model as client.
- (2) Web service layer: It is web service bus. It is a main layer of framework and made mainly of SOAP processor and WEB server. Web server's main function is to provide interactive Web pages and is responsible of transmitting the requirement from upper layer to SOAP processor and communicating with SOAP processor and client by HTTP.
- (3) Web service interface layer: This layer is composed of Web service interfaces and all kinds of deployed business logic components include existed and new developed. Its main function is to offer business logic and process by component. So this may not only improve the system reusability, maintenance and computing burden of distributed system but also be in favor of integrating with other application systems. The business logic or process of components deployed in this layer is generally crucial or core one in the application system. The business logic and data encapsulating mainly concern the business or process dealt with while some component method is called. Calling components is generated from upper layer. Web service tasks are visited by platform environment of operation system (Windows, UNIX, C#, JAVA etc.) and HTTP of internet, and more advanced components rather than distributed COM are allowed to use, which can expand data structure but need destroy client. Even if WEB services run in different platform or have different bottom implement mechanism, it can also be interactive and shared. This function can not be fulfilled by traditional technology such as CORBA, EJB or DCOM and so on. The more prominent thing is that the definition of Web service is

separated from its realization, namely loose-coupling. That means it becomes more convenient to modify web service rather than destroy current systems. It can promote the flexibility of EAI framework.

- (4) Data layer: This layer mainly concludes the enterprise application systems and various database systems. Enterprise application system comprises legacy system, customize system and new application system, etc. These application systems contain a large amount of historical datum which has very important value. The main functions of this layer are storing, updating, retrieval, modification, maintaining data security, integrity, consistency, etc. Due to application integration project based on WEB Service, each active customer visits data in database by sharing with other rather than the alone linking.

182.3.3 Interface Design Principle of WEB Service in the EAI Framework

WEB services are connected and interact with each other by their interface. The interface should be defined by means of neutral manner, which is independent of hardware, operation system and program language so that communication can be realized in the uniform and standard manner during constructing system service.

During the design of web service interface based on XML, the scalability and universality of the interface for application should be fully considered. The design of integration interface should abide by the following principles: (1) realizing the essential functions in order to exert the effectiveness of the interface; (2) hiding the implementation details and exposing the essential information; (3) unifying the style of interface and making it to be a certain expansibility; (4) creating and releasing the resources at the same level; (5) finding the errors in the lower levels and dealing with it in the higher levels.

Specific requirements are as follows: (1) General message interface of service defines or classifies the general message of each WEB service branches relation to the system; (2) Message mode can be assorted into several types such as requests, request/response, and notice. The dependent relationship between message should be reduced as far as possible; (3) Data granularity of interface should be reasonably segmented to improve service efficiency; (4) Giant data should be not contained into the service request or response message for the efficiency of serializing or desterilizing the SOAP message. It should be transmitted by other ways such as File Transfer Protocol; (5) The interface grammar adopts WSDL format. The same function should use the standard or same tag name, the same attribute name and the same namespace, and avoid repetitious definition and data confusion. Simple or short data is expressed with properties and the binary data or string more than 100 bytes characters is put into content of tags with XML format; (6) The safety measures are taken as follows: (a) encrypting the transfer protocol such as SSL/TSL and the core elements by XML encryption; (b) authenticating the user's

access permission for server/client with the keys and certificates; (c) using the digital signature to put an end to tampering the transmitting messages.

182.3.4 Implementation of Interfaces in the EAI Framework Based on WEB Service

A Web service can define one or multiple interfaces. Through these interfaces, Web services can exchange message with its business partners. But one interface can only bind a partnership service. Therefore Web service interaction occurs in two web service interfaces. If the system contains multiple Web services, any of two interactive Web services is connected by a pair of particular interfaces.

The descriptions of web service interface define control message interaction rules. In the SCM services, For example, a seller requirement logic is that an invoice news and transportation news are delivered only after received in the payment news. Therefore, the description of interface includes two parts: message description and process description. Message description defines all news sent or received by Web services. Process description defines the constraint relations between the messages. It regulates a legitimate sequence in time when WEB service is sending or receiving messages. The service interface call can be implemented by the following steps:

- (1) Generating two WSDL documents by the Wsdngen tools, one is Provider_A_Order_Service.wsdl which defines the reusable part of the service and includes types of message and operations for services, another is Provider_A_Order_Service.wsdl which is special part for services implementing and binding.
- (2) After accurately developing and deploying WSDL documents, the Java agent is generated by Proxygen in WSTK. It generates Provider_A_Order_ServiceBinding class, which contains the methods: Provider_A_Appointment_ServiceBinding (Java.net,URL endPointURL) .

182.4 Compared with Traditional EAI Method

In this framework, Web services are issued to private UDDI used for enterprise internal application system and public UDDI called for external enterprises or commerce fellows. WEB service can uniformly encapsule message, action, data and business flow. So the application programs can become reused components. The components encapsuled enterprise core business functions by web services can be well shared among enterprise departments. It is arranged in UDDI once all service consumers in network can call and integrate those WEB services at any time.

Table 182.1 The characteristics of the framework compared with traditional EAI method

EAI method	Data description	Coupling degree	Business adaptability	Expansibility
P2P data migration	Language related	Tight	Bad	Bad
Component/middleware	Language related	General	General	General
This framework	XML format	Loose	Good	Good

This framework has completely changed the point-to-point integration mode in traditional EAI by defining public functions which can interact each other among the application programs based on different operation system or development language, it can be rapidly and low-costly developed, disposed, discovered and dynamically binds the applications in an aggregate form. The characteristics of the framework compared with traditional EAI method are shown in the Table 182.1.

182.5 Conclusion

EAI is a method and technology which integrates all kinds of application program based on different platforms and different development language. EAI combines all the enterprise heterogeneous systems, the applications and the data sources, and is an effective way to share and transmit data each other in the enterprise internal and external. Integration interface is an indispensable part of EAI system, which can form a plug-play relationship between different application or platform for realizing accurate data transmission.

By the typical characteristic of web services, such as really platform independence and language independence, EAI integration project based on web services has good expansibility, reconstruction and flexibility. Therefore, the designed framework can commendably realize integration of various enterprise application systems, and also has characteristics of loose-coupling, location transparency, agreement independent and supports dynamic business requirement. So the “information island” problem can be solved by this framework. It utilizes the reusability of SOA service component to integrate the information of heterogeneous system. With information integration, enterprise can be serviced by instant remote network rather than on-site guidance. The EAI framework and its interfaces based on web service offers an important reference solution to improve the service quality, shorten the response time and promote the competition for enterprises.

Acknowledgments This paper is supported by NSFC, under grant No. 90818004, and by Scheduled Scientific Research Fund of Hunan Provincial Science. & Technology. Department, under grant No. 2007FJ3091.

References

1. Santos RLT, Pablo AR (2008) A web services-based framework for building componentized digital libraries. *J Syst Softw* 81:809–822
2. Davies NJ, Fensel D, Richardson M (2004) The future of web services. *BT Technol J* 22(1):118–30
3. Zhang HJ (2008) Research and implementation enterprise application integration framework based on SOA[J]. *Comput Eng Des* 29(8):2085–2092 (in Chinese)
4. Bin J, Yan GG, Zhu XX (2007) SOA-based business process integration system for small and medium-sized manufacturing enterprise[J]. *J Comput-Aided Des Comput Graph* 19(1):125–129 (in Chinese)
5. Deng S-G, Ying L (2007) Determination and computation of behavioral compatibility for web services[J]. *J Softw* 18(12):3001–3014 (In Chinese)
6. Song L, Fu XJ (2005) On the architecture of enterprise application integration for oriented-services[J]. *J Jilin Univ* 6:657–663 (In Chinese)
7. Singh PM, Huhns MN (2005) *Service oriented computing, semantics, processes, agents*. Wiley Press, New York
8. Huang Y, Chung J (2003) A web services-based framework for business integration solutions[J]. *Electron Comm Res Appl* 2:15–26
9. Xianpeng HP (2010) Research on web services discovering algorithm based on information description framework[J]. *Comput Sci* 37(2):134–138 (In Chinese)
10. CCW. EAI. From Interface to Services [EB/OL]. (23 January 2007) (01 October 2007) <http://www.amteam.org/k/Theory/2007-1/0/539573.html> 1.2010.10
11. Lin BP (2009) Research on compatibility and mediation of web service interfaces[D]. A dissertation for doctor's degree, University of Science and Technology of China 5:17–29 (in Chinese)
12. Han L-Q, Zhang C-S (2003) XML-based interface design of WEB services[J]. *Comput Eng Des* 24(9):47–49

Chapter 183

The Automatic Classification 3D Point Clouds Based Associative Markov Network Using Context Information

Gang Wang, Ming Li, TingTing Zhou and Longgang Chen

Abstract Many applications of mobile mapping want to automatic classification point clouds into different classes for further processing. In this chapter we present a new approach for labeling 3D point clouds with using a novel feature descriptor—the four directions scan line gradient, and context classification models—associative Markov network (AMN). To build informative and robust 3D feature point representations, our descriptors encode the underlying surface geometry around a point using multi-scanlines gradients. It is more stable and reliable than normal vectors in urban environments with wide variety of natural and manmade objects. By defining objects models of 3D geometric surfaces and making use of contextual information of AMN, our system is able to successfully segment and label 3D point clouds. We use FC09 datasets to evaluate the proposed algorithm.

Keywords 3D point clouds · Associative Markov network · Four directions scan line gradient

G. Wang (✉) · L. Chen
Henan Electric Power Survey and Design Institute, Zhengzhou 450007,
Henan, China
e-mail: hepsdi@163.com

M. Li
School of Computer Science, Wuhan University, Wuhan
430072, Hubei, China
e-mail: liming751218@gmail.com

T. Zhou
Power and Mechanical College, Wuhan University, Wuhan
430072, Hubei, China
e-mail: zhoutingwhu@163.com

183.1 Introduction

Recently, LiDAR had proven efficient for environment perception as their resolution exceeds radar and ultrasonic sensors, and provides more direct distance measurements than camera [1]. With the development of positioning systems, the accurate and dense point clouds of large scale environments are easier than ever to be collected [2]. The challenge for these applications is extracting features robustly and correctly, because there are different kinds of objects in urban environments, such as bushes, trees, or curvy objects whose shapes are hard to define.

In order to solve those problems, we propose an approach that is capable of automatic classification 3D point clouds in urban environments. The major contributions are:

First, different to only using tangent, normal and principal and least principal eigenvectors features, we introduce four directions scan line gradient features (4DSG). Second, due to the SVM usually fails to enforce local consistency of the classification predictions [3]. For example arches on buildings and other less planar regions are consistently confused for trees, even though they are surrounded entirely by buildings. We used associative Markov network (AMN) [4] as context model to consider the interactions with neighbors during point's classification.

This chapter is organized as follows. In the next [Sect. 183.2](#) we first give an outline of relevant works, followed by the detailed description of our approach. In [Sect. 183.4](#) we provide experimental results. [Section 183.5](#) concludes this chapter and draws a conclusion.

183.2 Previous Work

A common approach for point cloud classification is to learn a classifier that assigns a label to each point, independently of its neighbors' assignments. The widely used features are surface's estimated curvature, normal [5] or local descriptors such as moment invariants, and spherical harmonic invariants [6]. They have been successfully used registration, segmentation and classification, which computationally fast and produce good results when the extracted features are discriminative. However, when the extracted features are noisy, this approach can produce noisy classifications where points' labels are not locally consistent. In general, descriptors that characterize points with a single value are not expressive enough to make the necessary distinctions for point-surface classification. As most scenes will contain many points with the same or very similar feature values, thus reducing their informative characteristics.

In order to model context, it is necessary to consider the interactions with many neighbors for each point during classification. The label of the points are then determined jointly, not independently. The Markov random field framework is a popular choice because it models the spatial interactions present in the scene [7].

They classify objects by incorporating statistical relationships between nearby object parts. Huber [8] presents an approach for parts-based object recognition. This method provides a better classification because nearby parts that is easier to identify than others help to guide the classification. A similar idea of detecting object components has been presented by Limketkai et al. [9]. The idea here is to exploit the spatial relationship between nearby objects, which in this case consist of 2D line segments. The work that is mostly related to the approach described in this chapter is presented by Anguelov et al. [10], in which an AMN approach is used in a supervised learning setting to classify 3D range data.

The classification point cloud with Markov random field can be formulated in terms of minimizing an energy function. And the unary and pairwise clique potentials unable to capture the rich statistics of natural scenes. Higher order clique potentials have the capability to model complex interactions of random variables and thus could overcome this problem. However, the lack of efficient algorithms for performing inference in these models has limited their popularity. The runtime complexity of commonly used inference algorithms such as belief propagation (BP) [11] or tree reweighted message passing (TRW) [12] grows exponentially with the clique size, which makes them inapplicable to functions defined on even moderate sized cliques. Recent work on message passing algorithms has been partly successful in improving their performance for certain classes of higher order potential functions. Lan et al. [13] proposed approximation methods for BP to make efficient inference possible in higher order MRFs. However, as these methods are based on BP, they are quite slow and take minutes or even hours to converge. Kohli et al. [14] recently showed how certain higher order clique potentials can be minimized using the expansion and swap [15] move making algorithms for approximate energy minimization.

Our classification approach is used discriminative models AMN as a natural way to model correlations between classification labels. The contributions of this chapter are using a novel feature descriptor—the four directions scan line gradient, and context classification models—AMN. Our descriptors encode the underlying surface geometry around a point using multi-scanlines gradients. It is more stable and reliable than normal vectors in urban environments with wide variety of natural and manmade objects.

183.3 Associative Markov Random Fields

As the features are extracted robustly, the structure and context information of points is used to eliminating false classifications in sparse sampled regions and keep consistency of classification result. This is modeled in a mathematical framework known as Markov random fields.

Suppose we are given a set of N data points p_1, \dots, p_N and K object classes C_1, \dots, C_K . For each data point p_i , we are also given a feature vector x_i . The classification task is to find a label $y_i \in \{1, \dots, K\}$ for each p_i so that all labels

y_1, \dots, y_N are optimal. A likelihood function of the labels y given the features x is defined as $P_w(y|x)$. Assuming labels \hat{y} of each point of training data have been assigned by us in advance. Then, the classification problem can be divided into two stages:

Learning step: Find $w^* = \arg \max_w P_w(\hat{y}|x)$, which we need to find good parameters w^* for the likelihood function $P_w(y|x)$.

Inference step: Find $y^* = \arg \max_y P_w(\hat{y}|x)$, which we seek for good labels y^* that maximize this likelihood.

183.3.1 Definitions

An associative Markov network is an undirected graph in which the nodes are represented by N random variables y_1, \dots, y_N . These random variables are corresponding to the labels of each of the data points p_1, \dots, p_N . Each node y_i and each edge (y_i, y_j) in the graph has an associated non-negative value $\varphi(x_i, y_j)$ and $\psi(x_{ij}, y_i, y_j)$ respectively.

These are also known as the potentials of the nodes and edges. The node potentials reflect the fact that for a given feature vector x_i some labels are more likely to be assigned to p_i than others, whereas the edge potentials encode the interactions of the labels of neighboring nodes given the edge features x_{ij} . The conditional probability that is represented by the network can be expressed as

$$P(y|x) = \frac{1}{Z} \prod_{i=1}^N \varphi(x_i, y_i) \prod_{(i,j) \in E} \psi(x_{ij}, y_i, y_j) \tag{183.1}$$

Here, Z denotes the partition function which by definition is given as $Z = \sum_{y'} \prod_{i=1}^N \varphi(x_i, y'_i) \prod_{(i,j) \in E} \psi(x_{ij}, y'_i, y'_j)$.

In Taskar et al., the potentials are defined using the log-linear model. In this model, a weight vector w^k is introduced for each class label $k = 1, \dots, K$. The node potential φ is then defined so that $\log \varphi(x_i, y_i) = w_n^k \cdot x_i$ where $k = y_i$. Accordingly, the edge potentials are defined as $\log \psi(x_{ij}, y_i, y_j) = w_e^{k,l} \cdot x_i$ where $k = y_i$ and $l = y_j$. Note that there are different weight vectors $w_n^k \in R^{d_n}$ and $w_e^{k,l} \in R^{d_e}$ for the nodes and edges.

183.3.2 Learning

First, the Equation 183.1 is used as the conditional probability $P_w(y|x)$ where the parameters w are expressed by the weight vectors $w = (w_n, w_e)$, we obtain that $\log P_w(y|x)$ equals

$$\sum_{i=1}^N \sum_{k=1}^K (w_n^k \cdot x_i) y_i^k + \sum_{(ij) \in E} \sum_{k=1}^K (w_e^{k,k} \cdot x_{ij}) y_i^k y_j^k - \log Z_w(x). \quad (183.2)$$

Note that the partition function Z only depends on w and x , but not on the labels y .

As mentioned above, in a standard supervised learning task the goal is to maximize $P_w(y|x)$, however, if we instead maximize the margin between the optimal labeling \hat{y} and any other labeling y defined by

$$\log P_w(\hat{y}|x) - \log P_w(y|x), \quad (183.3)$$

The term $Z_w(x)$ cancels out and the maximization can be done efficiently. This method is referred to as maximum margin optimization. The problem is reduced to a quadratic program (QP) of the form:

$$\begin{aligned} & \min \frac{1}{w} \|w\|^2 + C\xi & (183.4) \\ \text{s.t.} \quad & wX\hat{y} + \xi - \sum_{i=1}^N \alpha_i \geq N; \quad w_e \geq 0; \\ & \alpha_i - \sum_{ij \in E} \alpha_{ij}^k - w_n^k \cdot x_i \geq -\hat{y}_i^k, \quad \forall i, k; \\ & \alpha_{ij}^k + \alpha_{ii}^k - w_e^k \cdot x_{ij} \geq 0, \quad \alpha_{ij}^k, \alpha_{ii}^k \geq 0, \quad \forall ij \in E, k \end{aligned}$$

Here, the variables that are solved for in the QP are the weights $w = (w_n, w_e)$, a slack variable ξ and additional variables α_i, α_{ij} and α_{ij} .

183.3.3 Inference

Once the optimal weights w have been calculated, it can be inference on an unlabeled test data set. This is done by finding the labels y that maximize $\log P_w(y|x)$. As mentioned above, Z does not depend on y so that the maximization in Eq. 183.4 can be carried out without considering the last term. With the constraints imposed on the variables y_i^k this leads to a linear program of the form

$$\begin{aligned} & \max \sum_{i=1}^N \sum_{k=1}^K (w_n^k \cdot x_i) y_i^k + \sum_{ij \in E} \sum_{k=1}^K (w_e^k \cdot x_{ij}) y_{ij}^k & (183.5) \\ \text{s.t.} \quad & y_i^k \geq 0, \quad \forall i, k; \quad \sum_{k=1}^K y_i^k = 1, \quad \forall i \end{aligned}$$

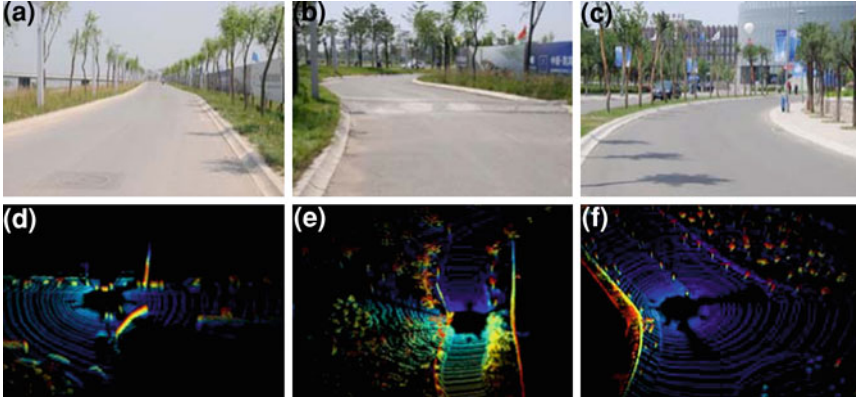


Fig. 183.1 T Classification results, *first column* are test results of straight course. *Second column* are test results of curve course, *Third column* are test results of turn round in intersection. (Color legend: tree/green, building/red, road/blue, pedestrian/yellow, vehicle/pink, ground/gray blue). **a** The view of straight course. **b** The view of curve course. **c** The view of intersection. **d** The point cloud viewed in color. **e** The ground truth of straight course. **f** Bayes classifier. **g** The AMN classifier

$$y_{ij}^k \leq y_i^k, \quad y_{ij}^k \leq y_j^k, \quad \forall ij \in E, k$$

Here, we introduced variables y_{ij}^k representing the labels of two points connected by an edge.

183.4 Experiments

To validate our framework, we have performed experiments of point cloud classification for real world datasets. The 3D point clouds were captured by SmartV-II. In order to evaluation classification performance, the samples of point cloud have marked as road, buildings, vehicles, pedestrians, trees, and ground. Then, the recall rate is used to evaluating performance.

The baseline and AMN classification results are compared in Fig. 183.1, (green indicates tree, red indicates building, blue indicates road, yellow indicates pedestrian, pink indicates vehicle, and gray blue indicates ground). The pictures views of environments are list in first row, the second row are pint clouds viewed in color. The segment results using normal vector is show in center picture, and the segment results using gradient in four directions is also show in right picture. The walls in the right picture are well segment contrast to center picture.

First, the straight course is a flat ground in urban area with large trees and building, thus the keyword flat will be used to refer to this frame. The bayes

classifier discriminated the road and building correctly. In this example, there are three different causes of misclassifications: edge effect (at the border of surface areas such as curbs), presence of several surfaces (junction of vehicle with ground surface) and scatter effect (at the bottom of the building and vehicle, due to the classifier cannot distinguished which points are tree or other objects). The results of AMN had markedly improved, the “scatter” and “linear” class are fused together and the ground surface class points are separated from other surface class points using a our method presented, One can note that the tree trunk, even though surrounded by vegetation, is correctly classified. But there still have some error such the curbs of road are misclassified.

The curve course is more complex scenes. The surface of road has undulating, separated from other grounds by curbs, ditches etc. In order to distinguish road from other ground, its structure must be considered. The AMN had will deal with it compared with the bayes for the pavement segmented well.

The intersection is more complex scenes, the pavements are irregular shape. Obviously, curbs are borders of road. It can be used as dividing line between pavement and other ground. The patterns of curbs can be divided into two groups: one is the direction of curb perpendicular to scan line; another is the direction of curb parallel to scan line. The 4DSG can reflect the changes of point cloud on curb in the four directions, which used to detect road edge very well. The third column shows the classification result and the largest connected components for a scene containing many curbs, sidewalk, ditches short trees and a chain-link fence covered by vegetation. Some of the ground points are misclassified as linear because of the scanning pattern of the sensor. The main causes of misclassification are complex structures of vegetation such as junction of tree trunks with ground surface. The AMN had will deal with it compared with the bayes for the tree segmented well.

183.5 Conclusion

We present a new approach for labeling 3D point clouds with using a novel feature descriptor—the four directions scan line gradient, and context classification models—AMN.

The maximum-margin learning is used to find the optimal tradeoff between the node and edge features. Then our system is able to successfully label 3D point clouds according to point’s geometry and context information, which can guarantee spatial contiguity of classification results.

Acknowledgments National Natural Science Foundation of China under Grant No. 41050110437,41001306. “Remote Collaboration Office Management Information System” Science and Technology Project of Henan Electric Power Company (2010 Batch 1 Class 3 Item 14).

References

1. Montemerlo M, Becker J, Bhat S, Dahlkamp H, Dolg D, Ettinger S, Haehnel D (2008) Junior: the stanford entry in the urban challenge. *J Field Robot* 25(9):569–597 C
2. Lalonde J-F, Vandapel N, Hebert M (2007) Data structures for efficient dynamic processing in 3-d. *Int J Robot Res* 26(8):777–796
3. Taskar B, Chatalbashev V, Koller D (2004) Learning associative Markov networks. In: Twenty first international conference on machine learning
4. Alexa M, Adamson A (2004) On normals and projection operators for surfaces defined by point sets. In: Proceedings of symposium on point-based graphics, pp 149–155
5. Mitra NJ, Nguyen A (2003) Estimating surface normals in noisy point cloud data. In: SCG'03: Proceedings of the nineteenth annual symposium on computational geometry, pp 322–328
6. Burel G, H'encocq H (1995) Three-dimensional invariants and their application to object recognition. *Signal Process* 45(1):1–22
7. Gelfand N, Mitra NJ, Guibas LJ, Pottmann H (2005) Robust global egistration. In: Proceedings of symposium on geometric processing
8. Anguelov D, Taskar B, Chatalbashev V, Koller D, Gupta D, Heitz G, Ng A (2005) Discriminative learning of Markov random fields for segmentation of 3d scan data. In: Proceedings of the conference on computer vision and pattern recognition (CVPR) pp 169–176
9. Frome A, Huber D, Kolluri R, Bulow T, Malik J (2004) Recognizing objects in range data using regional point descriptors. In: Proceedings of ECCV
10. Limketkai B, Liao L, Fox D (2005) Relational object maps for mobile robots. In: Proceedings of the international joint conference on artificial intelligence (IJCAI), pp 1471–1476
11. Anguelov D, Taskar B, Chatalbashev V, Koller D, Gupta D, Heitz G, Ng A (2005) Discriminative learning of markov random fields for segmentation of 3d scan data. In: Proceedings of the conference on computer vision and pattern recognition (CVPR), pp 169–176
12. Meltzer T, Yanover C, Weiss Y (2005) Globally optimal solutions for energy minimization in stereo vision using reweighted belief propagation. In: ICCV pp 428–435
13. Kolmogorov V (2006) Convergent tree-reweighted message passing for energy minimization. *PAMI* 28(10):1568–1583
14. Kohli P, Kumar M, Torr P (2007) P3 & beyond: solving energies with higher order cliques. In: CVPR
15. Boykov Y, Veksler O, Zabih R, Fast approximate energy minimization via graph cut, *IEEE Transactions on* 23:1222–1239

Author Index

B

Bao, Fuguang, [1181](#)
Bao, Mengjie, [945](#)
Bao, Zhigang, [1575](#)

C

Cai, Yangang, [1347](#)
Cao, Weiwei, [1479](#)
Cao, Xiaopeng, [1171](#)
Cao, Zhengjun, [925](#), [1539](#)
Catterall, Noel, [1609](#)
Cen, Qin, [1357](#)
Che, Wujiang, [1685](#)
Chen, Feng, [1017](#), [1467](#), [1471](#)
Chen, Haixiong, [1385](#)
Chen, Li, [991](#), [1007](#)
Chen, Longgang, [1695](#)
Chen, Meng, [1453](#)
Chen, Peiling, [1405](#)
Chen, Ting-gui, [915](#)
Chen, Wei, [1171](#)
Chen, Weigang, [1549](#)
Chen, Yangfei, [1559](#)
Chen, Ying, [1479](#)
Chen, Yongxian, [1575](#)
Cheng, Chuanzhou, [1665](#)
Cheng, Niansheng, [1129](#)
Chu, Jianfeng, [881](#), [889](#)

D

Dai, Shaowu, [1501](#)
Deng, Hepu, [1599](#)
Ding, Shuhui, [1375](#)

Ding, Xinyi, [1347](#)
Ding, Yingchun, [904](#)
Dun, Bin, [1017](#), [1467](#), [1471](#)

F

Fan, Decheng, [1521](#)
Fan, Feng, [935](#)
Fang, Jie, [1639](#)
Fei, Yulian, [1479](#)
Feng, Qiming, [1297](#)
Feng, Xiaofei, [1119](#), [1633](#)
Feng, Yi, [1617](#)
Fu, Jie, [1249](#)
Fu, Pei-Hua, [1259](#)
Fu, Peihua, [1559](#)

G

Ge, Yujia, [1119](#), [1549](#), [1583](#), [1633](#)
Geng, Qingtian, [1053](#)
Gong, Liquan, [1639](#)
Gorbenko, Anna, [971](#), [1289](#)
Guo, Jianwei, [1327](#)
Guo, Li, [1221](#)
Guo, WenYing, [1647](#)
Guo, ZhouWei, [1153](#)

H

Han, Na, [1591](#)
Han, Peiyong, [1583](#)
He, Guozhu, [1569](#)
He, Tao, [1655](#)
He, Yuanliu, [1079](#)

H (*cont.*)

Hu, Juan, 1241
 Hu, Lemin, 1665
 Hu, Liang, 881, 889
 Hu, Zepeng, 1583
 Hua, Ertian, 1181
 Huang, Jun, 897
 Huang, Yihao, 961

J

Jiang, Bo, 1655
 Jiang, Gangyi, 1089, 1099
 Jiang, Nan, 1053
 Jiang, Nana, 1033
 Jiang, Ping, 1267
 Jiang, Xiaoning, 961
 Ju, Chun-hua, 1231
 Ju, ChunHua, 1367

K

Kavs, Aleksander J, 1453

L

Lang, Jihai, 897
 LI, Bin, 1281, 1489
 Li, Fucui, 1099
 Li, Guangyuan, 1327
 Li, Hongtu, 881, 889
 Li, Jian, 1591
 Li, Jianghua, 1241
 Li, Juefeng, 1453
 Li, Jungang, 1209
 Li, Lei, 1249
 Li, Lu, 1201
 Li, Ming, 1695
 Li, Renwang, 1575
 Li, Shi-Jin, 1317
 Li, Shuangming, 1501
 Li, Xiaodong, 863
 Li, XiaoLong, 1191
 Li, Xueyi, 1375
 Li, Yulan, 981
 Li, Zhangbing, 1685
 Liang, Xiao-yi, 1633
 Liang, Yongquan, 1337
 Lin, Jing, 904
 Ling, Yun, 1307
 Liu, Changchun, 1079
 Liu, Cheng, 1569
 Liu, Chunxiao, 1249, 1347
 Liu, Cun, 1079, 1453

Liu, Dong, 1191
 Liu, Dongsheng, 1443
 Liu, Jiankang, 1569
 Liu, Jianxun, 1685
 Liu, Junling, 1053
 Liu, Kan, 1521
 Liu, Lihua, 1539
 Liu, Ling, 1427
 Liu, Mei, 1375
 Liu, Meifeng, 1145
 Liu, Tong, 1337
 LIU, Xiaoming, 991, 1007
 Liu, Yafei, 999
 Liu, Yujing, 1221
 Liu, Zhen, 981
 Lu, Haijun, 945
 Lu, Ling, 1145
 LU, Mei, 1395
 LU, Ming, 1017, 1467, 1471
 LU, Ying, 1017, 1467, 1471
 Lu, Zailin, 1575
 Luo, Jiabin, 1347

M

Ma, Nan, 1327
 Meng, Jianjun, 999
 Meng, Xingbao, 904

N

Ni, Weijian, 1337

O

Ouyang, Yi, 871

P

Peng, Hao-Yu, 1109
 Peng, Yang, 1259
 Peng, Zongju, 1089, 1099
 Popov, Vladimir, 971, 1289

Q

Qi, Wenzhe, 999
 Qian, Rong, 863
 Qing, Ji, 945

S

Sha, Qian, 925
 Shang, Weilie, 1171

Shao, Feng, 1089, 1099
 Sheka, Andrey, 971
 Shen, Sunyuan, 953
 Shi, Lingping, 999
 Shu, Bo, 1591
 Shuai, Xing-xia, 1231
 Si, Jiandong, 1181
 Song, Jiandong, 1375
 Song, Yijun, 1137
 Su, Juan, 1079
 Sun, Fengfei, 1099
 Sun, Jie, 1109

T

Tan, Runhua, 1267
 Tao, Xiaobo, 1513

U

Ullah, Waheed, 935

W

Wan, Ding-Sheng, 1317
 Wan, Wenlong, 981
 Wan, Youhong, 1209
 Wang, Gang, 1695
 Wang, Ji-Min, 1317
 Wang, Jiale, 1655
 WANG, Jian-dong, 1281, 1489
 Wang, Jiaqi, 1267
 Wang, Jie, 1347
 WANG, Liang-liang, 1161
 Wang, Shoujue, 953
 Wang, Xianbao, 953
 Wang, Xingyu, 1201
 Wang, Xiping, 961
 Wang, Xun, 1109
 WANG, Zhixue, 991, 1007
 Wibowo, Santoso, 1599
 Wong, Szemin, 1443
 Wu, Jingru, 1427
 Wu, Pin, 1171
 Wu, Qianhong, 1665
 Wu, Zhijun, 897

X

Xiao, Junjun, 1479
 Xiao, Wenxian, 981
 Xiao Xiao, 953
 Xie, Qi, 945
 Xing, Jianguo, 871

Xu, Bin, 1453, 1549, 1583
 Xu, Guangjian, 1307
 Xu, Hongxia, 1079
 Xu, Jie, 1221
 Xu, Jingxun, 1417
 Xu, Yingying, 1249
 Xue, Deqian, 1673

Y

Yan, Huayun, 1673
 Yang, Bingru, 1327
 Yang, Gelan, 1137
 Yang, Hai-han, 1231
 Yang, Li, 1665
 Yang, Yang, 1267
 Yang, Yanmei, 999
 Yang, Zhiyong, 1501
 Yang, Jiangtao, 935
 YE, Zhaoqing, 1531
 Yi, Tian, 1405
 Yin, Qian, 1575
 Yu, Mei, 1089, 1099
 Yuan, Wei, 881, 889

Z

Zhan, Penghui, 1575
 Zhang, Aiju, 1633
 Zhang, Faming, 1531
 Zhang, Guofu, 1043
 Zhang, Hongwei, 1417
 Zhang, Jing, 1267
 Zhang, Kejun, 863
 Zhang, Lei, 1181
 Zhang, Lili, 1079
 Zhang, Rong-fu, 1161
 Zhang, Tie-zhu, 1231
 Zhang, Xia, 1171
 Zhang, Xiaokun, 863
 Zhang, Xun, 1435
 Zhang, Yan, 1435
 Zhang, Ying, 991, 1007
 Zhang, YiShun, 1367
 Zhang, Zhenchuan, 1033
 Zhang, Zihou, 1201
 Zhao, Geng, 863
 Zhao, Hongwei, 1053
 Zhao, Jinfang, 1025
 Zhao, Xiang-Jun, 1395
 Zhao, Xuying, 863
 Zheng, Dequn, 1171
 Zheng, Huali, 1153
 Zheng, Qiaoyan, 1089

Z (*cont.*)Zheng, Yanling, [1079](#)Zhong, Peisi, [1375](#)Zhou, Houkui, [1071](#)Zhou, Huanyu, [1549](#)Zhou, Jinbo, [1161](#)Zhou, TingTing, [1695](#)Zhou, Xiaohong, [1079](#)Zhou, Yan, [1071](#)Zhou, Yongqiang, [1249](#)Zhu, Qiwei, [1453](#)Zhu, Xiao-ling, [915](#)Zhu, Xin-feng, [1281](#), [1489](#)Zhu, Yue-Long, [1317](#)Zou, Shurong, [1417](#)