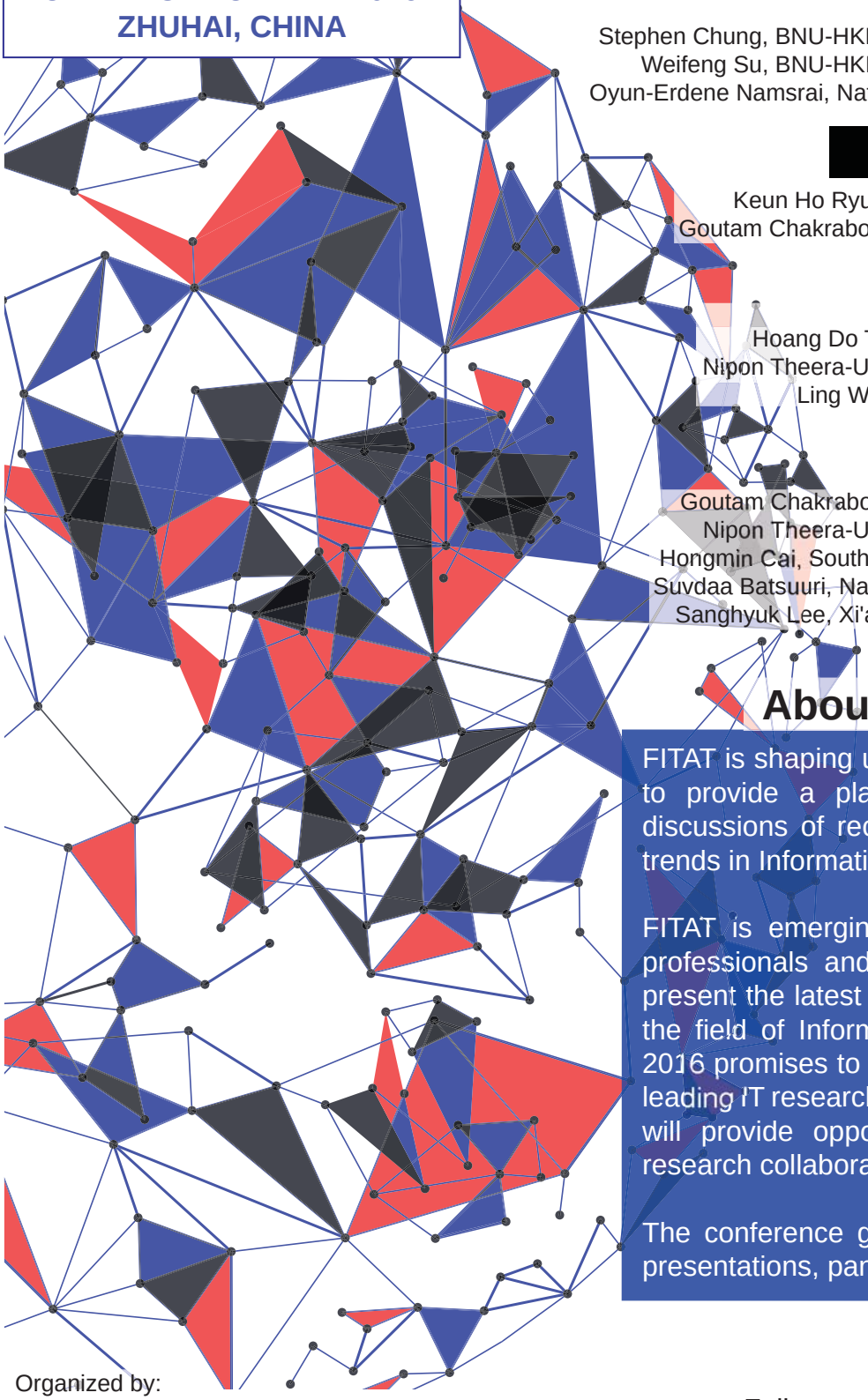
FITAT 2016

ISSN : 2288-9973
www.fitat.org

The 9th International Conference on Frontiers of Information Technology, Applications and Tools

31 MARCH - 3 APRIL 2016
ZHUHAI, CHINA



General Chairs

Stephen Chung, BNU-HKBU United International College, China
Weifeng Su, BNU-HKBU United International College, China
Oyun-Erdene Namsrai, National University of Mongolia, Mongolia

International Advisory

Keun Ho Ryu, Chungbuk National University, Korea
Goutam Chakraborty, Iwate Prefectural University, Japan

PC Chairs

Hoang Do Thanh Tung, (IOIT) of (VAST), Vietnam
Nipon Theera-Umpon, Chiang Mai University, Thailand
Ling Wang, Northeast Dianli University, China

Keynote Speakers

Goutam Chakraborty, Iwate Prefectural University, Japan
Nipon Theera-Umpon, Chiang Mai University, Thailand
Hongmin Cai, South China University of Technology, China
Suvdaa Batsuuri, National University of Mongolia, Mongolia
Sanghyuk Lee, Xi'an Jiaotong-Liverpool University, China

About Conference

FITAT is shaping up to be an annual conference to provide a platform for presentations and discussions of recent developments and future trends in Information Technology.

FITAT is emerging as a leading forum for IT professionals and researchers to discuss and present the latest research trends and results in the field of Information Technology (IT). FITAT 2016 promises to be exciting event that will host leading IT researchers from across the globe and will provide opportunity to build international research collaborations.

The conference generally consists of technical presentations, panel discussions and posters.

Organized by:

Important deadline

Full paper submission due : 9 March 2016
Poster session abstract submission due : 9 March 2016
Notification of acceptance : 11 March 2016
Camera ready due : 15 March 2016
Registration due : 21 March 2016

Proceeding of the Ninth International Conference on Frontiers of Information Technology, Applications and Tools

FITAT 2016

Organized by
United International College, China
Chungbuk National University, Republic of Korea
National University of Mongolia, Mongolia

March 31 – April 3, 2016
Zhuhai, China

Dear Distinguished Delegates and Guests,

On behalf of the FITAT 2016 Conference Committees, it is our great pleasure to welcome you to BNU-HKBU United International College (UIC), Zhuhai, China for the 9th International Conference on Frontiers of Information Technology, Applications and Tools. We are grateful to have this year's conference held in UIC, Zhuhai, China. The International Conference on FITAT 2016 is emerging as a leading forum for IT professionals and researchers to discuss and present the latest research trends and results in the field of Information Technology (IT). FITAT 2016 promises to be exciting event that will host leading IT researchers from across the globe and will provide opportunity to build international research collaborations. We hope that you will benefit both scientifically and personally from the forum and enjoy the FITAT 2016.

This volume contains the Proceedings of the FITAT 2016, held in UIC, Zhuhai, China, and Mar 31 – Apr 3, 2016. The FITAT 2016 conference includes 5 keynote speeches, 26 speeches in oral presentation where each speech is assigned into one of the 6 conference sessions, and 22 posters in interactive presentation session.

The papers were submitted from many countries across the globe and many program committee members from 7 countries worked hard to prepare the FITAT 2016. We, on behalf of the Program Committees of FITAT 2016, would like to thank the many people who volunteered their time to help make the conference a success. We also thank all of the authors who submitted their best work to the conference. The research presented here represents a tremendous effort on the part of all of these people. We hope that the proceedings will serve as a valuable resource for the community.

With our warmest regards,

The Organizing Committees
Mar 31 – Apr 3, 2016

Dear Distinguished Delegates and Guests,

First of all, on behalf of the Organizing Committee, I would like to express my warmest welcome to all of you attending FITAT2016, the ninth International Conference of the Frontiers of Information Technology, Application and Tools (FITAT), at the Beijing Normal University Hong Kong Baptist University United International College (UIC).

As you express yourselves at the IT frontiers, UIC is a small college but at the educational frontier in China. We are the first university from Hong Kong to set up a higher education institution in Mainland China, we are the first university in China to adopt the liberal arts education concept, we are one of the first two universities in China that use only English as the medium of instruction. Since its first enrollment in 2005, UIC has grown into an international institution having partnership with many universities from the US, Europe, Australia and Korea. We adopt innovative educational approaches such as Outcome-Based Learning, Criteria-Referenced Assessment, E-Learning etc. 2015 marks the 10th Anniversary of UIC. In this short period of time, we have progressed from a Band 3 university to a Band 1 university. In recognition of our accomplishments, the Zhuhai Government has given us a plot of land to build our new campus. We are glad that at this important juncture of UIC, we have the opportunity to host FITAT2016.

Although UIC is not considered as an engineering school, concepts of information and communication technology (ICT) for engineers are applicable to liberal arts students. Recently, many leading universities are imposing ICT subjects as compulsory to every student. Therefore, we not only play host for this Conference, we will also be eager learners in the Conference.

We are thankful to our keynote speakers, Prof. Goutam Chakraborty, Prof. Nipon Theera-Umpon, Prof. Hong-Min Cai, Prof. Suvdaa Batsuuri and Prof. Sang-Hyuk Lee. We also specially thank to the leader of FITAT, Prof. Keun-Ho Ryu.

We hope that you enjoy your stay in Zhuhai, the most livable city in China.

President of BNU-HKBU United International College
Professor Ching Fai Ng

Distinguished Guests, Speakers, FITAT members, Friends and Families,

On behalf of the Committee of FITAT2016, I am pleased to welcome you to Zhuhai and the FITAT2016. FITAT2016 is the ninth International Conference on the Frontiers of Information Technology, Application and Tools. FITAT is emerging as a leading forum for IT professionals and researchers to discuss and present the latest research trends and results in the field of Information Technology (IT). FITAT 2016 promises to be exciting event that will host leading IT researchers from across the globe and will provide opportunity to build international research collaborations. During this conference, there will be many vigorous discussions so that we can sharpen our knowledge and ideas. Through this, we encourage ourselves to keep our endless passion for academic achievement.

I am grateful to the Committee members for organizing and hosting this year's conference at UIC. I extend a special word of gratitude to our Co General Chairs, Prof. Oyun-Erdene Namsrai, and Dr. Weifeng Su who have labored tirelessly to ensure an excellent event. I also offer my sincere appreciation to the Technical Program Committee Co-Chairs who, working with their respective committees, have assembled an outstanding program. Finally, I close by saying thank you to all attendees. Your participation and enthusiasm is essential to the ongoing success of FITAT. Enjoy the conference and welcome to UIC and Zhuhai

General Chair of FITAT2016
Prof. Stephen Chung
Prof. Weifeng Su
Prof. Oyun-Erdene Namsrai

FITAT 2016 COMMITTEES

General Chairs

Stephen Chung	BNU-HKBU United International College	China
Weifeng Su	BNU-HKBU United International College	China
Oyun-Erdene Namsrai	National University of Mongolia	Mongolia

International Advisory Committee

Keun Ho Ryu	Chungbuk National University	Republic of Korea
Goutam Chakraborty	Iwate Prefectural University	Japan

PC Chairs

Hoang Do Thanh Tung	Vietnam Institute of Information Technology (IOIT) of Vietnamese Academy of Science and Technology (VAST)	Vietnam
Nipon Theera-Umpon	Chiang Mai University	Thailand
Ling Wang	Northeast Dianli University	China

Local Coordinators

Ms. Chunyan Gigi	BNU-HKBU United International College	China
Aziz Nasridinov	Chungbuk National University	Republic of Korea

Publication Chairs

Seon-phil Jeong	BNU-HKBU United International College	China
Basabi Chakraborty	Iwate Prefectural University	Japan

Committee Secretary

Dingkun Li	Chungbuk National University	Republic of Korea

Program Committee

Altannar Chinchuluun	National University of Mongolia	Mongolia
Anwar F.A. Dafa-Alla	Garden City College	Sudan
Aziz Nasridinov	Chungbuk National University	Republic of Korea
Baasantseren Ganbat	National University of Mongolia	Mongolia
Basabi Chakraborty	Iwate Prefectural University	Japan
Batnyam Battulga	National University of Mongolia	Mongolia
Bayarpurev Mongolyn	National University of Mongolia	Mongolia
Bold Zagd	National University of Mongolia	Mongolia
Bu Hyun Hwang	Chonnam National University	Republic of Korea
Bum Ju Lee	Korea Institute of Oriental Medicine	Republic of Korea
Byungchul Kim	Baekseok University	Republic of Korea
Dong Gyu Lee	University of Tokyo	Japan
Ella Roubtsova	Open University	Netherlands
Erwin Bonsma	Philips	Netherlands
Eun Jong Cha	Korean NRF	Republic of Korea
Garmaa Dangaasuren	National University of Mongolia	Mongolia
Goce Naumoski	Bizzsphere	Netherland
Heon Gyu Lee	Electronics and Telecommunications Research Institute	Republic of Korea
Herman Hartmann	University of Groningen	Netherlands

Ho Sun Shon	Chungbuk National University	Republic of Korea
Hoang Do Thanh Tung	Vietnam Institute of Information Technology (IOIT) of Vietnamese Academy of Science and Technology (VAST)	Vietnam
Incheon Paik	The University of Aizu, Japan	Japan
Jeong Hee Chi	Konkuk University	Republic of Korea
Jeong Hee Hwang	Namseoul University	Republic of Korea
Jeong Ji Mun	Namseoul University	Republic of Korea
Jeong Seok Park	Chungju National University	Republic of Korea
Ji Moon Chung	Namseoul University	Republic of Korea
Jong Yun Lee	Chungbuk National University	Republic of Korea
Jung Hoon Shin	Chonbuk National University	Republic of Korea
Jungpil Shin	The University of Aizu	Japan
Keun Hwan Jeon	Kunjang University	Republic of Korea
Kwang Su Jung	Chungbuk National University	Republic of Korea
Kwang Woo Nam	Kunsan National University	Republic of Korea
Kyeong Ja Jeong	ChungCheong University	Republic of Korea
Kyung-Ah Kim	Chungbuk National University	Republic of Korea
Mark van den Brand	TU/e	Netherlands
Menno Lindwer	Intel	Netherlands
Michel Chaudron	Leiden University	Netherlands
Mohamed Ezzeldin A. Bashir	Faculty of Computer Science, University of Medical Sciences and Technology	Sudan
Moon Sun Shin	Konkuk University	Republic of Korea
Ms. Chunyan Gigi	BNU-HKBU United International College	China
Nyamjav. J	National University of Mongolia	Mongolia
Oyun-Erdene Namsrai	National University of Mongolia	Mongolia
Razvan Dinu	Philips	Netherlands
Sanghyuk Lee	Xi'an Jiaotong-Liverpool University	China
Seon-phil Jeong	BNU-HKBU United International College	China
Sun Shin Kim	Cha University	Republic of Korea
Sung Bo Seo	Turbo Soft Company	Republic of Korea
Sung Hee Park	Soongsil University	Republic of Korea
Supatra Sahaphong	Ramkhamhaeng University	Thailand
Tom Arbuckle	University of Limerick	Ireland
Vu Thi Hong Nhan	Vietnam National University	Vietnam
Weifeng Su	BNU-HKBU United International College	China
Yang Koo Lee	ETRI	Republic of Korea
Yanja Dajsuren	TU/e	Netherlands
Ye Ho Shin	Far East University	Republic of Korea
Yonsik Lee	Kunsan National University	Republic of Korea
Yoon Ae Ahn	Health and Medical Information Engineering, College of Life	Republic of Korea
YoungSung Cho	CEO	Republic of Korea

Conference Venue

BNU-HKBU United International College (UIC)
United International College (UIC), situated in Zhuhai and jointly founded by Beijing Normal University and Hong Kong Baptist University (HKBU), is the first full-scale cooperation in higher education between the Mainland and Hong Kong. Its charter has been approved by the Ministry of Education with full support from local authorities.

Address : 28 Jinfeng Road, Tangjiawan, Zhuhai, Guangdong Prov. 519085, China
(广东省珠海市唐家**湾**金凤路 28 **号** 519085)

FITAT 2016

The Ninth International Conference on the Frontiers of Information Technology, Application and Tools; (FITAT) is shaping up to be an annual conference to provide a platform for presentations and discussions of recent developments and future trends in Information Technology.

FITAT 2016 is emerging as a leading forum for IT professionals and researchers to discuss and present the latest research trends and results in the field of Information Technology (IT). FITAT 2016 promises to be exciting event that will host leading IT researchers from across the globe and will provide opportunity to build international research collaborations.

The FITAT 2016 will discuss the challenges facing information technology professionals with core database technologies, bio-medical informatics, sensor network technology and current IT applications.

We are glad to let you know that the proceeding of the FITAT 2011 was one of the ACM International Conference Proceedings Series (ACM ICPS) so that the full-text of the proceedings papers have been put into ACM Digital Library, under the heading of the "ACM International Conference Proceedings", by ACM.

Topics

The topics of interest include, but are not limited to :

- Core Database Technologies including database administration, indexing, performance tuning, and query processing
- E-commerce, emerging object/web technologies, and information economics related topics such as web services, remote monitoring, financial market analysis, data warehousing and the semantic web
- Parallel and distributed data mining algorithms, Mining on data streams and Sensor Data, Spatial data mining, Text, video, multimedia data mining, Mining social network data, and Data mining support for designing information systems
- Smart city, Smart buildings and urban development, Smart grid and microgrids, Energy harvesting, storage, and recycling, Renewable energy models and prediction.
- IT integrated manufacturing, medical informatics, digital libraries, and mobile computing
- Bioinformatics, Genomics, Biometrics, and Image interpretations
- Base disciplines of computer science, telecommunications, operations research, economics and cognitive sciences
- Methods and tools related to the model-driven development, domain-specific languages, software architecture, and quality of architecture and design for all domains including automotive, healthcare, telecommunications, and finance.

PROGRAM OF FITAT 2016

Outline

Day 1 Pre-reception and Meeting: 18:00 ~ 21:00, Thursday, March 31, 2016

Day 2 FITAT Oral session: 09:00 ~ 18:30, Friday, April 1, 2016
Poster sessions: 13:00 ~ 18:30, Friday, April 1, 2016

Day 3 FITAT Oral: 09:00 ~ 15:30, Saturday, April 2, 2016
Panel discussion: 15:30 ~ 18:00, Saturday, April 2, 2016

Day 4 Business Committee Meeting and Tour of UIC: 09:00 ~ 12:30, Sunday, April 3, 2016

Presentation	China	Korea	Japan	Sudan	Mongolia	Vietnam	Thailand	Total
Oral	7	10	2	1	9	1	1	31
Poster	0	6	0	0	15	1	0	22
Total	7	16	2	1	24	2	1	53

Friday, April 1, 2016

FITAT Oral Presentation (09:00 ~ 18:00)

Time	Sessions
08:30 ~ 09:00	**Registration** **Location: Lobby 1**
09:00 ~ 10:20	**Session Name: FITAT Opening and Keynote Session 1** **Location: Room B101** Session Chair: Weifeng Su *BNU-HKBU United International College, China* **Keynote Speech** **Anomaly Detection in Time-series Data - A Case Study with Continuously Monitored Periodic Bio-signals (40min)** Goutam Chakraborty *Iwate Prefectural University, Japan* **Mapping the Knowledge Domain of FITAT for Better Research Collaboration and Dissemination (20 min)** Musa Ibrahim M. Ishag, SangHun Han, Keun Ho Ryu *Chungbuk National University, Korea*
10:20 ~ 10:40	**Coffee Break**
10:40 ~ 12:00	**Session Name: Oral Session 1** **Location: Room B101** Session Chair: Oyun-Erdene Namsrai *National University of Mongolia, Mongolia* **Digital Processing of Signal Channel Spectrometry System (20min)** Tsend-Ayush Oldokh, Jamiyan Sukhbaatar, Nyamjav Jambaljav, Bold Zagd *National University of Mongolia, Mongolia* **A Study on Skyline Query Processing Using Entropy Score Curve (20min)** Jong Hyeok Choi, Aziz Nasridinov, Jong Yun LEE *Chungbuk National University, Korea* **An approach to detect TCP based attack using Data mining algorithms (20min)** Ugtakhbayar.N, Usukhbayar.B and Nyamjav.J *National University of Mongolia, Mongolia*

	Simulation Studies of Switching Arc Behavior in High Voltage Puffer Type SF6 Circuit Breakers (20 min) Kai Shen Ee, Dingkun Li, Yu Fu, Keun Ho Ryu *Chungbuk National University, Korea*
12:00 ~ 13:30	**Lunch** **Location: Restaurant Green Wood**
13:30 ~ 15:40	**Session Name: Oral Session 2** **Location: Room B101** Session Chair: Basabi Chakraborty *Iwate Prefectural University, Japan* **Invited Speech** **Traditional Mongolian Script Segmentation (25min)** Suvdaa Batsuuri *National University of Mongolia, Mongolia* **Effect of Cognitive Distraction on Driving Behaviour (25min)** Basabi Chakraborty, Yusuke Manabe, Sho yoshida and Kotaro Nakano *Iwate Prefectural University, Japan* **Diurnal Variation of Surface Radio Refractivity over Mongolia (20min)** Jamiyan Sukhbaatar, Tsend-Ayush Oldokh, Bold Zagd, Nyamjav Jambaljav *National University of Mongolia, Mongolia* **Real-time Document Ranking using Term Weight Estimation in Information Retrieval** **(20 min)** Erdenebileg Batbaatar, Aziz Nasridinov, Oyun-Erdene Namsrai, Keun Ho Ryu *Chungbuk National University, Korea* **Mining Association Rules from Educational Data to Improve Teaching and Learning Outcomes (20min)** Chunyan Ji, Clement Leung, Junru Zhong *BNU-HKBU United International College, China* **An Image Retrieval Framework based on Knowledge Ontology (20min)** Clement Leung, Yuanxi Li *BNU-HKBU United International College, China*
15:40 ~ 16:00	**Coffee Break**

16:00 ~ 18:00	**Session Name: Oral Session 3** **Location: Room B101** Session Chair: Seon-phil Jeong *BNU-HKBU United International College, China* **Keynote Speech** **Breast Abnormality Detection in Mammograms Using Fuzzy Inference Systems (40min)** Nipon Theera-Umpon *Chiang Mai University, Thailand* **Traffic Flow Analysis on Public Transport Access Data (20min)** Amarsanaa Ganbold, Tsolmon Zundui, Purev Jaimai *National University of Mongolia, Mongolia* **Finding Prognostic Factors to MACE in Patients with Myocardial Infarction (20min)** Young Joong Kim, Ho Sun Shon, Man Geun Jeong, Kyung Ah Kim, Jong Yung Lee *Chungbuk National University, Korea* **Automated Detection of Outliers in Cardiovascular Database (20min)** Man Geun Jeong, Young Joong Kim, Jong Yun Lee, Ho Sun Shon *Chungbuk National University, Korea* **Online Motivation Analysis Model over Cloud Computing Environment (20min)** Hai Jing Jiang, Zhi Yuan Chen, Wei Ding, Tie Hua Zhou, Ling Wang *Northeast Dianli University, China*

Friday, April 1, 2016

FITAT Poster Presentation (13:30 ~ 18:10)

Time	Sessions
13:30 ~ 18:10	**Session Name: Interactive Session 1** **Location: Lobby 1** Session Chair: Ho Sun Shon *Chungbuk National University, Korea* **P1-01: Horse Stamp Detection in Real Nomadic Environment** Gantuya Perenleilkhundev, Bold Zagd, Suvdaa Batsuuri **P1-02: Survey on 3D model based pose estimation methods** E.Tsetsegjargal, R.Javkhlan, D.Usukhbaatar, B.Suvdaa **P1-03: Feature Selection in Intrusion Detection Datasets** Ugtakhbayar.N, Usukhbayar.B, Ganbayar.U, Nyamjav.J **P1-04: Design and Implementation of 32 bit MIPS Processor** Battogtokh.J, Batpurev M, Bold.Z **P1-05: The number of non-trivial solutions in Quadratic Sieve** Gantulga.G, Bayarpurev.M, Garmaa D. **P1-06: SDN design for Enterprise Network** Ganbayar Uuganbayar, Ugtakhbayar Naidansuren, Naranbaatar Bold-Erdene, Usukhbayar Baldangombo **P1-07: A Finite-state Morphological Transducer for Khalkha Mongolian Nominal** Zoljargal Munkhjargal, Altangerel Chagnaa **P1-08: Improving the Result of the Model for Predicting the Class Fault Proneness Using Data Mining Anomaly Detection Techniques** Batnyam Battulga, Lkhamrolom Tsoodol, Erdenetuya Namsrai, Purev Jaimai **P1-09: Modern Trend of Mongolian Horse Stamp** Gantuya Perenleilkhundev, Shaariibuu Setev, Suvdaa Batsuuri **P1-10: Differential Wheeled Mobile Robot Real Time Self-localization and Path Planning Method for Microcontroller** Batbayar Unursaikhan, Bold Zagd

P1-11: Self-tuning PID Controller for Dynamic Systems
Batbayar Unursaikhan, Battur Ganbat, Lodoiravsal Choimaa

P1-12: An Improved Medical Decision Support System for Predicting the Stages of Chronic Obstructive Pulmonary Disease
Solongo Khurts, Nasantuya Namsrai, Erdenetuya Namsrai, Otgonnaran Ochirbat

P1-13: Land Management System with Instant Area Estimator
Oktyabar Enkhtaivan, Nasanbat Namsrai, Oyun-Erdene Namsrai

P1-14: Virtual Lab Management Using Citrix
Ankhzaya Jamsrandorj, Sodbileg Shirmen

P1-15: Building OpenWRT Embedded Linux in Atheros
Ankhzaya Jamsrandorj, Sodbileg Shirmen

P1-16: An Augmented Reality Integrated Pseudo-3D Map and Optical Tracking Application
Phuong Tien Nguyen, Tung Duong Vu, Hue Thi Le

P1-17: Listener's Preference Based Bayesian Learning for Recommendation in Music Site
Young Sung Cho, Song Chul Moon, Seon-Phil Jeong, Keun Ho Ryu

P1-18: Competitiveness Enhancement of Home IoT Service by Smart Home Mirror
Yeong Real Kim, Tae Gu Kang, Kyung Mun Kang

P1-19: Screening of Allosteric Inhibitors for p21-activated Kinases
Duk-Joong Kim, Chang-Ki Choi, Chan-Soo Lee, Kyung-Ah Kim, Eun-Young Shin, Eung-Gook Kim

P1-20: Flow Generator System for Calibration and Comparison of Air Flow Modules
Eun-Jong Cha, Mi-Jung Park, Ji-Sun Lim, Eun-Young Shin, Yang-Mi Kim, Ho-Sun Shon, Kyoung-Ok Kim, Kyung-Ah Kim

P1-21: Risk Factor of Non ST-segment Elevation Myocardial Infarction (NSTEMI) Patients with Diabetes
Ho Sun Shon, Kyung Ah Kim

P1-22: The Electrophysiological Role of Epigallocatechin-3-gallate and Quercetin as TREK2 Antagonists
Kyung-Ah Kim, Yangmi Kim

Saturday, April 2, 2016

FITAT Oral Presentation (09:00 ~ 15:30)

Time	Sessions
08:30 ~ 09:00	**Registration** **Location: Lobby 1**
09:00 ~ 10:10	**Session Name: Keynote Session 2 and Oral Session 4** **Location: Room B101** Session Chair: Lin Wang *Northeast Dianli University, China* **Invited Speech** **Energy Balance of Smart Grid (30min)** Sang Hyuk Lee *Xi'an Jiaotong-Liverpool University, China* **Analysis of The Risk factor of Death in Stomach Adenocarcinoma Patients (20 min)** Jeong Ho Lee, Kwang Ho Park, Keun Ho Ryu *Chungbuk National University, Korea* **Design of a Security Framework for Big Data (20min)** Razan Abualgasim, Anwar F.A. Dafa-Alla *Independent researcher Khartoum, Sudan Sudan*
10:10 ~ 10:30	**Coffee Break**
10:30 ~ 12:20	**Session Name: Oral Session 5** **Location: Room B101** Session Chair: Musa Ibrahim M. Ishag *Chungbuk National University, Korea* **Invited Speech** **Using Jointly Constrained Optimization to Identify Both Recurrent and Individual Copy Number Variations (CNVs) from Multisample aCGH (30min)** Hongmin Cai *South China University of Technology, China*

	New Method to Determine Viewing Angle Analysis of Point Light Source Display (20min) Densmaa Batbayar, Enkhmunkh Tumurbaatar, Ganbat Baasantseren *National University of Mongolia, Mongolia* **Comparison of Classification Algorithms for the fruit yields (20 min)** Jong Seon Woo, Yongjun Piao, Hyunwoo Park, Keun Ho Ryu *Chungbuk National University, Korea* **Development of Robotics Teaching (20min)** Yanyan Ji, Hui Zhang, Chunyan Ji *BNU-HKBU United International College, China* **The system design based on the real-time electricity pricing (20min)** Zhi Yuan Chen, Hai Jing Jiang, Ding Wei, Tie Hua Zhou, Ling Wang *Northeast Dianli University, China*
12:20 ~ 13:30	**Lunch** **Location: Restaurant Green Wood**
13:30 ~ 15:30	**Session Name: Oral Session 6** **Location: Room B101** Session Chair: Kyung Ah Kim *Dept. of Nursing, Woosong College, Korea* **A Data Mining Approach for Bearing Failure Prediction Using Multiple Non-linear Features (20min)** Heon Gyu Lee, Hoon Jung *Electronic and Telecommunications Research Institute Republic of Korea, Korea* **Making Virtual Tour Suitable for Oculus Rift (20min)** Javkhlan Rentsendorj, Baatarbileg Altangerel, Oyun-Erdene Namsrai *National University of Mongolia, Mongolia* **Anomaly Detection Based Performance Improvement of Existing Business Intelligence System (20min)** Tsatsral Amarbayasgalan, Iderbaatar Munkhuu, Otgonnaran Ochirbat, Oyun-Erdene Namsrai *National University of Mongolia, Mongolia*

	Activity Recognition based on Clustering Methods for Senior Homecare Services (20min) Thi Hong Nhan Vu, Yang Koo Lee, Oyun-Erdene Namsrai *Vietnam National University, Veitnam* **Path Planning of Mobile Robot using Position System and Virtual Plane Approach in Dynamic Environment (20min)** Enkhtsogt.P, Zorig.B, Khurelbaatar.Ts *National University of Mongolia, Mongolia* **Spatial Keyword Queries using Spark for Big Social Data (20min)** Pyoung Woo Yang, Kwang Woo Nam *Kunsan National University, Korea*

FITAT 2016 Keynote Speaker

Goutam Chakraborty

Professor, Intelligent Informatics Lab.
Faculty of Software and Information Science
Iwate Prefectural University, Japan

Short Bio

Prof. Goutam Chakraborty received his Ph.D. in 1993 from Tohoku University, Japan. Before joining Tohoku University, he worked in Telecommunication Industry in India. Presently he is Professor and head of the Intelligent Informatics laboratory, Department of the Software and Information Science, Iwate Prefectural University, Japan. His research interests are Soft Computing algorithms and their applications to solve pattern recognition, prediction, scheduling and optimization problems including applications in wired and wireless Networks. Recently, he is interested in the analysis of various time-series signals, collected by sensors from Human body as well as machines.

FITAT 2016 Keynote Speaker

Nipon Theera-Umpon

Associate Professor
Department of Electrical Engineering
Faculty of Engineering Chiang Mai University
Director, Biomedical Engineering Center, Chiang Mai University

Short Bio

Nipon Theera-Umpon received his B.Eng. (Hons.) degree from Chiang Mai University, M.S. degree from the University of Southern California, and Ph.D. degree from the University of Missouri-Columbia, all in Electrical Engineering. His has been with the Department of Electrical Engineering, Chiang Mai University since 1993. He had been a Visiting Scholar at the University of Missouri Columbia (2000-20001) and Kagawa University, Japan (2004). He has served as Editor, Associate Editor, Peer Reviewer, General Chair, Technical Chair and Committee Member of several journals and conferences. He served as Associate Dean of Engineering at the Faculty of Engineering, Chiang Mai University from 2005-2009 and as the Chairman for Graduate Study in Electrical Engineering from 2004-2007. He is presently serving as the Director of Biomedical Engineering Center, and the Chairman for Graduate Study in Biomedical Engineering, Chiang Mai University. He is a member of Thai Robotics Society, Biomedical Engineering Society of Thailand, Council of Engineers in Thailand. He has served as Vice President of the Thai Engineering in Medicine and Biology Society. Dr. Theera-Umpon is a senior member of the IEEE. His research interests include Pattern Recognition, Digital Image Processing, Neural Networks, Fuzzy Logic, Medical Signal and Image Processing.

FITAT 2016 Invited Speaker

Hongmin Cai

Associate Professor
School of Computer Science & Technology
South China University of Technology

Short Bio

Dr. Hongmin Cai received bachelor and master's degree from Harbin Institute of Technology in 2001 and 2003, respectively. He got Ph.D from University of Hong Kong in 2007. In 2005, he was a research fellow at the Center of Bioinformatics at Harvard University. In 2006, he was visiting Section for Biomedical Analysis of Prof. Christos Davatzikos at University of Pennsylvania. From 2008 to 2012, he was assistant Professor at the School of Information and Technology, The Sun Yat-Sen University. He has joined School of Computer Science and Technology, South China University of Technology as associate Professor from 2012. In 2014, he was awarded Professor in Ad and Outstanding Youth Teacher in Guangdong Province. He has principled and co-principled more than 10 fundings from NSF, Guangdong Province and other sources. He published around 30 papers in top journals and conferences with Google citation more than 300 times. His research interests including biomedical image process and bioinformatics, in particular gene sequencing data analysis and survival data mining.

FITAT 2016 Invited Speaker

Suvdaa Batsuuri

Associate professor
School of Engineering and Applied Sciences
National University of Mongolia

Short Bio

Suvdaa Batsuuri received bachelor and master's degree from National University of Mongolia in 2002 and 2004, respectively. She got Ph.D in Computer Science at department of Computer and Software Engineering, Kumoh National Institute of Technology, South Korea, in 2011. From 2002 to 2007, she was working at National University of Mongolia. In 2010, she was a part-time lecturer at Kumoh National Institute of Technology. Since 2011, she is associate professor at department of Information and Computer Science, school of Engineering and Applied Sciences of NUM. Her research interests including image processing, computer vision, pattern recognition, distance metric learning and artificial Intelligence.

FITAT 2016 Invited Speaker

Sanghyuk Lee

Professor
Department of Electrical and Electronic Engineering
Xi'an Jiaotong-Liverpool University

Short Bio

Sanghyuk Lee is an Associate Professor in the Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, China. He received Doctorate degree at Seoul National University as majoring Electrical Engineering in 1998. His main research interests include data evaluation with similarity measure, human signal analysis, high dimensional data analysis, controller design for linear/nonlinear system, and observer design for linear/nonlinear system. Professor Lee had joined as vice president of Korean Convergence Society 2012, and he organized the international conference several times with KCS. And he awarded multiple honors such as outstanding scholar award from KCS and Korean fuzzy society and best paper award. He has published referred over 150 journal papers, and 80 conference papers.

Table of Contents

FITAT Papers

Anomaly Detection in Time-Series Data – A case study with Continuously Monitored Periodic Bio-signals

Goutam Chakraborty
Iwate Prefectural University, Japan

Abstract

Anomaly detection of time–series data is one of an important data mining task. In most cases, a times series has several kinds of anomalies. Moreover, defining anomaly needs domain knowledge of the data. In this work, we consider periodic bio–signals. We detect anomalies of a sub–sequence by its distance from normal one. Other physical features, as used by medical practitioners for concluding discords are not used. To establish the effectiveness of our idea, we use ECG data for experiments. Though the signal is periodic, a closer look will reveal that the periods differ at different times. We will be able to find some anomalies at a glance of the signal from the period of the each point. We define anomalies as those subsequences of a longer time series that are maximally different from all the rest of the subsequences of the whole sequence. Anomalies could be detected by comparing every pair of subsequences. But, in this method, if subsequence have a little different length of period than fundamental period, the distance registered is large even though there is no anomaly. We want to avoid such wrong decisions about anomaly. We used dynamic time warping (DTW), where the distance measure could take care of such elongation. However, it is computationally very heavy to use DTW for all comparisons. To cope with this computation complexity, to propose computationally light discord detection algorithm, we introduce a new concept we named mother signal. Mother signal is the average of subsequence which occurs most frequently. We cluster subsequences using Ward hierarchical clustering algorithm. Cluster with highest membership is the cluster of mother signals. An average of the members of that cluster is used as mother signal. If we use mother signal for comparison, it is efficient to detect anomaly and the result improves.

Mapping the Knowledge Domain of FITAT for Better Research Collaboration and Dissemination

Musa Ibrahim M. Ishag[1], SangHun Han[2], Keun Ho Ryu[*]
[1,2,*)]*Database and Bioinformatics Laboratory, School of Electrical and Computer Engineering, Chungbuk National University, South Korea*
{[1]ibrahim, [*]khryu}@dblab.chungbuk.ac.kr, [2]likelamb@gmail.com

Abstract

This paper applies the knowledge domain map analysis to the proceedings of the FITAT conference. The network analysis tool- CiteSpace is used to visualize the data. The visualization revealed the research fields, countries, institutions, and the collaboration patterns between authors. Finally, bright suggestions were given to improve the conference and the dissemination of its publications.

Keywords: Knowledge Domain Mapping; FITAT, Network Analysis, CiteSpace

1. Introduction

The International Conference on the Frontiers of Information Technology, Application and Tools (FITAT) is a nine years old annual venue for researchers in the field of Information Technology, and Computer Science [1]. It aims at gathering researchers and professionals in the field along with disseminating the results of their work in the form of proceedings. Mapping a research domain involves applying techniques such as Data Mining [2, 3, 4], Social Network Analysis [5], and Visualization [6] in order to know the experts, institutions, research topics, publications, and the collaboration patterns in the field [7].

Mapping knowledge domain analysis has been applied to various fields ranging from Terrorism Informatics [7], Intelligence and Security Informatics [8], and recently in Medical Informatics Education [9]. However, to the best of our knowledge, it has not been applied to analyze a conference.

In this paper, motivated by analytical capabilities offered my domain map analysis, and the sheer amount of publication and bibliographic data made available by the FITAT conference proceedings, CiteSpace network analysis and visualization tool was used to analyze the data. In essence, this paper puts forward the following major contributions:

- Applying Knowledge Domain Mapping Analysis to the FITAT conference proceedings using CiteSpace.
- Highlighting Issues related to future collaborations and better dissemination of the proceedings.

The reminder of this manuscript is organized as follows; section 2 describes the research method. Section 3 discusses the issues along with proper ways to compact them. Finally, section 5 provides a concluding summary of this manuscript.

2. Research Method

In order to map the domain of knowledge that is hidden in FITAT proceedings, a three step procedure was followed which involve data collection, preprocessing, and analysis using CiteSpace.

2.1. Data Collection and Preprocessing

The dataset used in this paper comprises the proceedings of FITAT. However, since the proceedings were not disseminated through Thomson Reuter's Web of Science [14], a preprocessing step was needed to annotate the bibliographic data. In essence, a typical citation record is made of 41 attributes as shown in Figure 2. In this paper, however, only a selected attributes were chosen due to unavailability. For example, no citation was reported. Although the total publications of FITAT from its establishment in 2008 until 2015 is considerably large as explained in Figure 1, only the first 16 papers published in the proceedings of FITAT 2015 were considered for analysis. The papers were then converted into a tagged format to make a dataset that is acceptable for analysis using CiteSpace[15, 16, 17, 18, 19].

2.2. Domain Map Analysis

2.2.1. From the view of Keywords.

Keywords and phrases that occur frequently together were visualized in a form of network in figure 4. The most salient keywords include; "Computer Science Research", "Research Output", and "Collaboration" among others. These phrases give a general idea of the topics discussed in the conference.

Table 1. The standard attributes of a typical citation record.

Field	Tag	Meaning	Ref
1	PT	Publication Type (J=Journal; B=Book; S=Series)	[10]
2	AU	Authors	[10]
3	AF	Author Full Name	[10]
4	TI	Document Title	[10]
5	SO	Publication Name	[10]
6	LA	Language	[10]
7	DT	Document Type	[10]
8	DE	Author Keywords	[10]
9	ID	Keywords Plus®	[10]
10	AB	Abstract	[10]
11	C1	Author Address	[10]
12	RP	Reprint Address	[10]
13	EM	E-mail Address	[10]
14	FU	Funding Agency and Grant Number	[10]
15	FX	Funding Text	[10]
16	CR	Cited References	[10]
17	NR	Cited Reference Count	[10]
18	TC	Times Cited	[10]
19	Z9	Total Times Cited (Web of Science Core, BIOSIS Citation Index, and Chinese Science Citation Database)	[11]
20	U1	Usage Count (Last 180 Days)	[11]
21	U2	Usage Count (Since 2013)	[11]
22	PU	Publisher	[10]
23	PI	Publisher City	[10]
24	PA	Publisher Address	[10]
25	SN	International Standard Serial Number (ISSN)	[10]
26	EI	Electronic International Standard Serial Number (eISSN)	[12]
27	J9	29-Character Source Abbreviation	[13]
28	JI	ISO Source Abbreviation	[13]
29	PD	Publication Date	[13]
30	PY	Year Published	[13]
31	VL	Volume	[13]
32	IS	Issue	[13]
33	BP	Beginning Page	[13]
34	EP	Ending Page	[13]
35	DI	Digital Object Identifier (DOI)	[13]
36	PG	Page Count	[13]
37	WC	Web of Science Categories	[13]
38	SC	Research Areas	[13]
39	GA	Document Delivery Number	[13]
40	UT	Accession Number	[13]
41	ER	End of Record	[13]

2.2.2. The co-citation Network.

In order to get a bird's eye view of the literature from which most of the authors base their work, a co-citation network was constructed using CiteSpace. The network is shown in figure 3. In the figure nodes represent cited articles and there will be a link between two nodes if they were co-cited in at least one paper which was presented in FITAT 2015. The nodes in the network were further organized into clusters of papers with similar topics. Each cluster has a numeric ID and a label. Table2 summarizes the first cluster whose ID is 0 . The cluster has 8 members and a silhouette value of 1. It is labeled as mobile "Peer-to-peer network" by LLR, "Study" by TFIDF, and "..." by MI. The most active citer to the cluster is 1 Song,, R (2015) a study on resource discovery for mobile peer-to-peer networks.

2.2.3. Institutional Collaboration.

For the purpose of understanding the collaboration patterns among the researchers and their affiliations, two types of networks were formed: a co-authorship network, and an institutional level network. The co-authorship network is illustrated by figure 2. In which authors who publish together form a cohesive cluster. It is clearly seen that most of the authors work in separate groups without nodes that can serve as hubs to link more than one research group. Figure 4 on the other hand, expresses the collaboration between the research institutions. In the network, institutions from three different countries can clearly be distinguished. Namely, the countries are China, Mongolia, and South Korea. Even at this level few collaborations exists. Particularly, between Chungbuk National University and Gangneung-Wonju National University both of which are located in South Korea. Another collaboration exist between National Northeast Dianli University and China University of Minig and Technology Beijing. The two institutions are based in Chain.

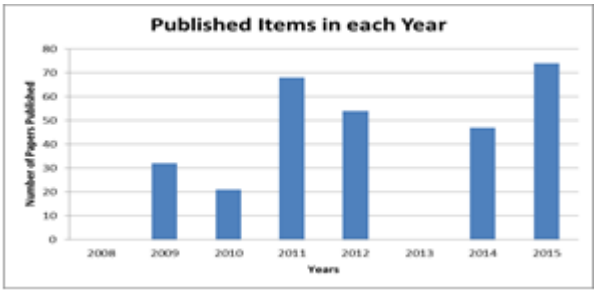


Figure 1. Publication Status of FITAT

3. Open Issues

From the preliminary results sought above, the following three directions can be presented for discussion during the PC, and Advisory board meetings of FITAT:

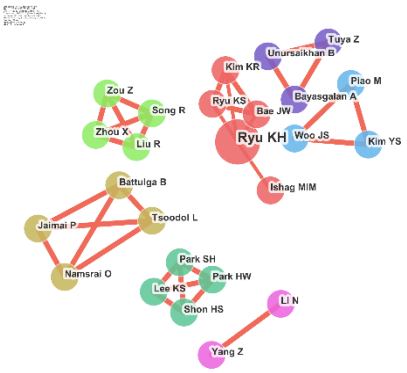


Figure 2. Co-author Network

Figure 3. Co-Citation Network

Table 2. Collaboration Network

ClusterID	Size	Silhouette	Label (TFIDF)	Label (LLR)	Label (MI)	Year Ave
0	8	1	(5.21) study	mobile peer-to-peer network (23.99, 1.0E-4)	…	2010

Figure 4. Collaboration Network

3.1. FITAT Journals.

FITAT can establish itself as a global network with its own journals focused on the topics discussed during the conference.

Figure 5. Key-words

3.2. Collaboration between the institutions.

Universities can carry out joint research.

3.3. Collaboration with ISI Publishers.

Collaboration with reputed journal is needed in order to disseminate selected papers which will then be extended and submitted for publications in possible volumes or special editions.

4. Conclusion

This paper exploited the knowledge domain map analysis to analyze the bibliographic data constructed from the proceedings of FITAT. Although the results are preliminary and not generalizable due to the sample of records selected for analysis, bright suggestions were presented in order to enhance the dissemination of future proceedings of FITAT for wider impact. Due to the laborious tasks involved in preparing the datasets, the authors are considering the proposal of automatic extraction of citation records from conference proceedings as a future research.

5. Acknowledgment

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2015-H8501-15-1013) supervised by the IITP(Institute for Information & communication Technology Promotion), and by Basic

Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No.2013R1A2A2A01068923).

6. References

[1] http://fitat.org/ [accessed on 3/18/2016: 12:44 PM]

[2] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. Introduction to data mining. Vol. 1, Pearson Addison Wesley, Boston, 2006.

[3] Han, Jiawei, Micheline Kamber, and Jian Pei, Data mining: concepts and techniques, Elsevier, 2011.

[4] Aggarwal, Charu C, Data mining: The textbook, Springer, 2015.

[5] Carrington, Peter J., John Scott, and Stanley Wasserman, eds. Models and methods in social network analysis. Vol. 28, Cambridge university press, 2005.

[6] Spence, Robert. Information visualization. Vol. 1, New York: Addison-Wesley, 2001.

[7] Hsinchun Chen, Edna Reid, Joshua Sinai, Andrew Silke, and Boaz Ganor, Terrorism Informatics: Knowledge Management and Data Mining for Homeland Security, Springer Science & Business Media , P. 3, 2008.

[8] Liu, Wenli, et al, "Collaboration Pattern and Topic Analysis on Intelligence and Security Informatics Research", IEEE Intelligent Systems 3 (2014), pp. 39-46

[9] Qian, Danmin, et al. "Mapping Knowledge Domain Analysis of Medical Informatics Education.", Frontier and Future Development of Information Technology in Medicine and Education, Springer, Netherlands, 2014, pp. 2209-2213.

[10] http://images.webofknowledge.com/WOKRS53B4/help/WOS/hs_wos_fieldtags.html [accessed on 3/27/2016: 9:22 PM]

[11] https://images.webofknowledge.com/WOKRS518B4/help/WOK/hs_alldb_fieldtags.html [accessed on 3/27/2016: 9:22 PM]

[12] https://images.webofknowledge.com/WOKRS515B5/help/WOK/hs_bp_fieldtags.html [accessed on 3/27/2016: 9:22 PM]

[13] https://images.webofknowledge.com/WOKRS57B4/help/WOS/hs_wos_fieldtags.html [accessed on 3/27/2016: 9:22 PM]

[14] Reuters, Thomson. "Web of Science.", (2012).

[15] Chen, Chaomei, and Loet Leydesdorff. "Patterns of connections and movements in dual map overlays: A new method of publication portfolio analysis." Journal of the association for information science and technology 65.2 (2014): 334-351.

[16] Chen, Chaomei. "Predictive effects of structural variation on citation counts." Journal of the American Society for Information Science and Technology 63.3 (2012): 431-449..

[17] Chen, Chaomei, Fidelia Ibekwe-SanJuan, and Jianhua Hou. "The structure and dynamics of cocitation clusters: A multiple- perspective cocitation analysis." Journal of the American Society for Information Science and Technology 61.7 (2010): 1386-1409.

[18] Chen, Chaomei. "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature." Journal of the American Society for information Science and Technology 57.3 (2006): 359-377.

[19] Chen, Chaomei. "Searching for intellectual turning points: Progressive knowledge domain visualization." Proceedings of the National Academy of Sciences 101.suppl 1 (2004): 5303-5310.]

Digital Processing of Single Channel Spectrometry System

Tsendayush Oldokh, Jamiyan Sukhbaatar, Nyamjav Jambaljav, Bold Zagd
Department of Electronics and Communication Engineering,
School of Engineering and Applied Sciences,
National University of Mongolia
{tsendayush, jamiyan, nyamjav, bold}@seas.num.edu.mn

Abstract

The purpose of this work was to evaluate the possibility of using digital processing of signals in the single channel system, registering signals from the output of preamplifier and amplifier, and comparing the results and see the advantages of digital processing of signals. The novelty of this study is first time signals from the single channel spectrometry system was digitally recorded and analyzed. The advantage of the digital system was evaluated and suggested for future applications. Measurements of signals were done using high precision ADC for Single Channel Nuclear Spectrometry system using Cs^{137} gamma sources, from two points of the system, output of preamplifier and the main amplifier. Results show that for the digital measurement, it can be done from the output of preamplifier, not using main amplifier.

Keywords:

1. Introduction

Nuclear Spectrometry Single Channel System is used widely in nuclear physics application for industry. The traditional single channel system is based on analog channel consisting of scintillation detector, preamplifier, amplifier, single channel analyzer and counter. This kind of system was used for determination of Flour Spar concentration into the ore on the truck and concentration in enriching reactor of the Bor-Undur Flour Spar Mining Company Mongolia [1.2].

Digital processing of single channel system is getting possible, thanks to rapid development of electronics, especially Analogy to Digital Converters (ADC) and development of big memory size of measuring system. The advantage of using ADC for Single Channel System is the availability of registering each signal from the detector in a digital way. This makes it possible to analyze each pulse by software that selects the right signals by disseminating the overlapped and noise influenced signals. Also it is possible to evaluate the system on the whole by determining the characteristics of the system to analyze noises.

The purpose of this work was to evaluate the possibility of using digital processing of signals in the single channel system, registering signals from the output of preamplifier and amplifier, and comparing the results and see the advantages of digital processing of signals.

The novelty of this study is first time signals from the single channel spectrometry system was digitally recorded and analyzed. The advantage of the digital system was evaluated and suggested for future applications

2. Results and Discussions

The nuclear spectrometry single channel system was described by the block diagram of the system and some basic blocks are shown in Figure 1. Conventional techniques to carry measurements by using single channel system and energy spectrum was also described. Modern ADCs and its characteristics are discussed. The PC oscilloscope (Picoscope 5000 series) system for registering detector signal was described. At the heart of the PicoScope 5000 is its ability to digitise signals accurately and with minimal distortion. The 250 MHz analog bandwidth is complemented by a real-time sample rate of 1 GS/s. For repetitive signals, an equivalent time sampling (ETS) mode increases the sampling to 20 GS/s.

An analog-to-digital converter (abbreviated ADC, A/D or A to D) is a device which converts a continuous quantity to a discrete time digital representation. An ADC may also provide an isolated measurement. Typically, an ADC is an electronic device that converts an input analog voltage or current to a digital number proportional to the magnitude of the voltage or current. However, some non-electronic or only partially

electronic devices, such as rotary encoders, can also be considered ADCs.

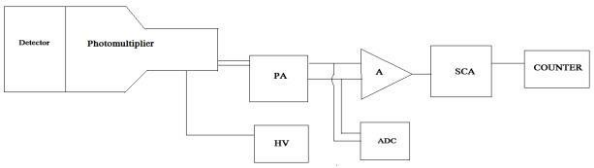


Figure 1

Figure 2. PicoScope 5000

The signals measured from the output of preamplifier and from the output of amplifier were shown in Figure 3 and 4. For the measurements, standard Cs^{137} gamma radiation source was used. For the experiment, single channel system of the Nuclear Research Center based on NIM crate was used. Blocks of the experiment is: NaJ scintillation detector with preamplifier, High Voltage block, Spectrometry Amplifier, single channel analyzer and counter, made in III Nuclear Instrumentation Factory, in Beijing, China were used.

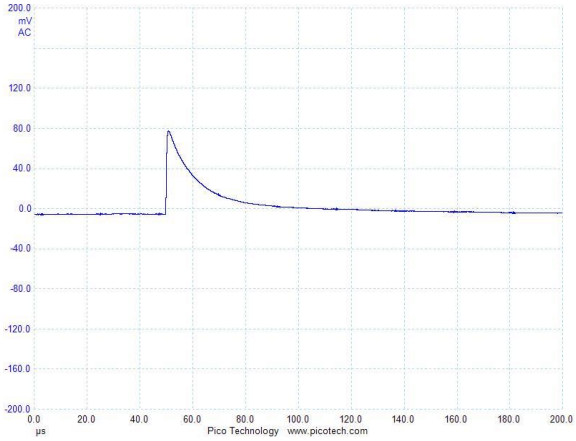


Figure 3. The signals were measured from the output of preamplifier.

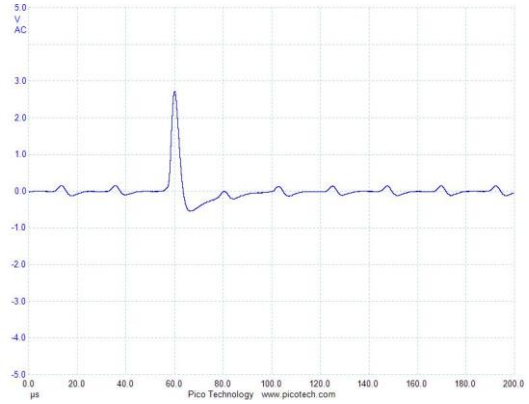


Figure 4. The signals were measured from the output of amplifier.

Spectrum Cs^{137} measured from the output of preamplifier is presented in Figure 5 and one from the output of amplifier not processed is presented in Figure 6.

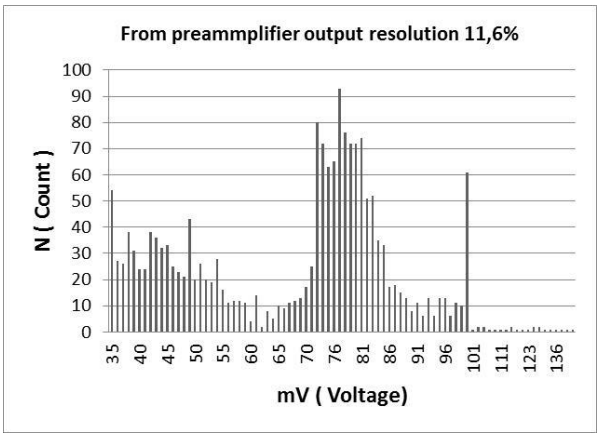


Figure 5. Spectrum Cs^{137} measured from the output of preamplifier.

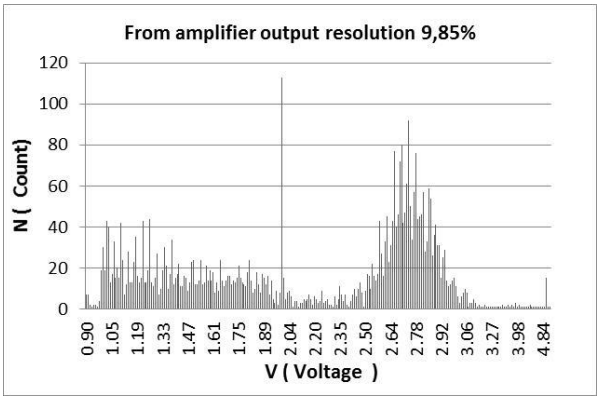


Figure 6. Spectrum Cs^{137} measured from the output of amplifier.

Signal forms from the output of preamplifier are presented in Figure 7. It was possible to measure the signals from the output of the preamplifier with a precision of 78mV, and we could sample the front edge of the pulses 20us times with a precision of 1ns.

Signal forms from the output of the spectrometric amplifier are presented in Figure 8. You can see that the amplitude is higher if compared with signals from the preamplifier we could sample each signal 10us times with a precision of 100ns. The amplitude was measured with a precision of 2.74 V.

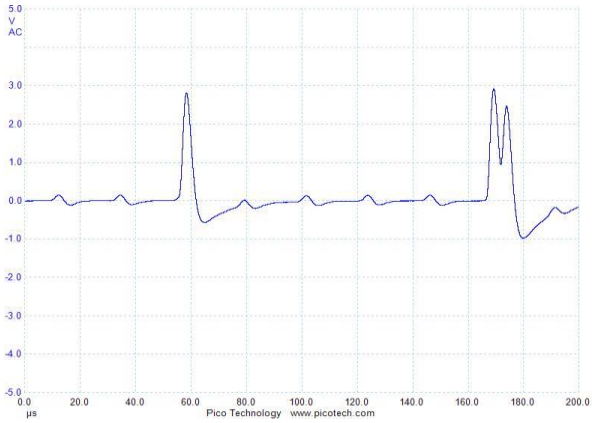


Figure 7. Signal forms from the output of preamplifier.

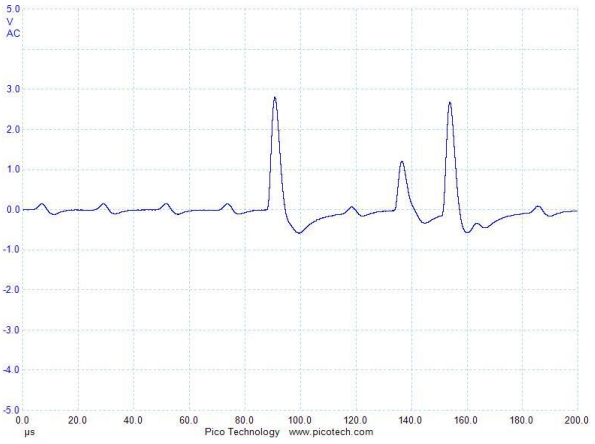


Figure 8. Signal forms from the output of the spectrometric amplifier.

Digital processing was done in the following way:

1. Signals which are overlapped were disseminated from the data (see Figure 7).
2. Signals overlapped on the considerable noises are disseminated from the data (see Figure 8).

The Cs^{137} spectrum after digital processing is presented in Figure 8. Several measurements made in the conventional way are presented in Figure 9.

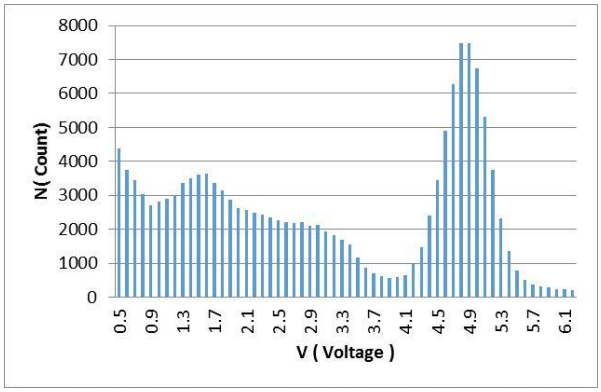


Figure 9. Several measurements made in conventional way.

After digital processing we can analyze the spectrums, we can see one high line because of the noise. Many experimenters who made measurements misunderstood that this is the peak of Cs^{137}, as this line had a little more amplitude than the Compton scattering peaks. But real Cs^{137} peak is a little higher voltage position (See Figure 9). This is the first advantage of digital processing to see spectrum and understand the nature of the spectrum.

In the above cases it was impossible to get energy spectrum if the digital processing was not switched on. Then we have analyzed real spectrum of the Cs^{137} and one can conclude that the resolution of the spectrometer is not much different from the spectrum measured in the output of the preamplifier and in the output of amplifier as they are 11.6% and 9.85%. From this we can conclude that it is not much gain of resolution if you get signals from the main amplifier. In spite of that theoretically it should be higher resolution from the output of spectrometry amplifier. This means that we can make measurements from the output of preamplifier not using the main amplifier for the experiments which is not requiring the higher resolution.

This way we can economize the blocks of amplifier, single channel analyzer and counter, and get spectrum in the result of using digital processing.

Main achievements of using of digital processing in the single channel spectrometry system are:

- By measuring each signal it can be registered, i.e. each signal can be measured digitally so that it can be described completely in each signal characteristic. This means each signal can be analyzed and processed;
- Detector and other block characteristics can be analyzed and troubleshot
- Dissemination of "bad" signals which overlapped with each other or overlapped with noise signals;
- Noises can be precisely analyzed, processed and depressed digitally by the software;
- Precision and accuracy of the measurement can be improved;
- Save some blocks namely, amplifier, single channel analyzer and counter;
- Make possible portable single channel instrument; Make multichannel analyzer using single channel system;

3. Conclusion

- Measurements of signals were done using high precision ADC for Single Channel Nuclear Spectrometry system using Cs^{137} gamma sources, from two points of the system, output of preamplifier and the main amplifier;
- Conventional measurements were done using the above system;
- Digital processing was done to improve the measurement results and the results were compared and discussed;
- Advantages of digital signal processing of nuclear single channel system were assessed;
- Digital signal processing can be applied for a Single Channel system;
- Advantages of digital signal processing for Single channel system is its ability to register each signal separately which are determining time, voltage and shape. Advantages described in chapter 5 are followed by this unique property;
- Results show that for digital measurement it can be done from the output of preamplifier, not using main amplifier;
- Use of digital processing of single channel system not only improve the quality of measurements, but make possible more portable system for use in the field;

4. References

[1] С.Лодойсамба, Н.Содном, Ш.Гэрбиш, Ж.Ганзориг, Б.Отгоолой, Г.Хүүхэнхүү, П.Улаанхүү, Д.Чүлтэм, Ж.Сэрээтэр, Б.Эрдэв, Д.Шагжжамба, Д.Баатархүү, Н.Норов, Х.Сиражет, "Разработка нейтронно-активационно и рентгено-флуресцентного методов анализа и применение их для определения вещественного состава различных образцов", *Эрдэм шинжилгээний ажлын тайлан 1981-1985*, МУИС,1985.

[2] Lodoysamba.S, Ganzorig.Zh, Gangaamaa.J, Otgooloi.B, Purev.Zh, Ulaankhuu.P, Shagjjamba.D, "*Electronic device for rapid fluorine ore analysis on Truck*", Abstracts for CAMAC`90 Nuclear Electronics and Interfaces Development and Application in Science and Process control, International Seminar, Warsaw, 1990.

A Study on Skyline Query Processing Using Entropy Score Curve

Jong Hyeok CHOI[1], Aziz Nasridinov[2], Jong Yun LEE[3]
[1]*Dept. of Computer Science, Chungbuk National University, Chungbuk, South Korea.*
[2,3)]*Dept. of Software Engineering, Chungbuk National University, Chungbuk, Korea.*
[1]*leopard@chungbuk.ac.kr,* {[2]*aziz,* [3]*jongyun*}*@chungbuk.ac.kr*

Abstract

Skyline is the set of tuples that are not dominated by any other tuple and have a better value at least one attribute than any other tuple. Especially skyline is very useful in the search areas because it suggests to the user the most representative results using the tuple is stored in the database. Therefore, various algorithms had been suggested that integrate SQL and Skyline operator to take advantage of stored data in the database. However, existing skyline algorithms had a problem because they show slow processing time that occurred due to a number of comparison process for searching the skyline. Therefore, we propose a method which can reduce the processing time by effectively reducing the number of comparisons in the search process for the skyline using entropy score curve.

Keywords: skyline, query processing, database management system

1. Introduction

Recently, due to rapid development of storage technologies, many companies have collected a massive amount of data. However, when this data is large, obtaining answer to a query may take a long time. On the other hand, we can use skyline queries that can help to quickly obtain top-k answers. Specifically, skyline is the set of tuples that are not dominated by any other tuples in the dataset, which means these tuples have a better value in at least one attribute [1]. Thus, skyline can be called as a representative set of tuples and used in searching applications.

The naïve method to get the skyline set is to compare a tuple with other tuples in the dataset. Block-Nester-Loops (BNL) [1] uses this naïve strategy for a comparison while maintaining a window of candidates. The drawback of BNL is that when a dataset is large it performs a large number of unnecessary comparisons, which result in enlarged processing time. Sort-Filter-Skyline (SFS) [2, 3] solves BNL's problem by pre-sorting the data according to the entropy score of each tuple. Once data is presorted, it is more likely that tuples with low entropy scores are not dominated by other tuples. Thus, SFS performs a small of amount comparisons comparing to BNL. However, when the data is large and has many attributes, performing SFS is not efficient as it takes a long processing time to calculate entropy score and sort the data.

In this paper, we propose a study on efficient skyline query processing using a novel entropy score curve. The proposed method is able to overcome the drawback of BNL and SFS by making fewer comparisons and without performing presorting of data.

2. Related Work

There have been many approaches proposed to efficiently compute the skyline. BNL and SFS are representative ones. In this section, we briefly describe these methods.

BNL [1] performs a pairwise comparison between every tuple in dataset. Here, if a tuple is neither dominated nor dominates the other tuples, then it is inserted into the candidate list that is maintained as a window in the main memory. If a tuple dominates other tuples in the candidate list, then it is inserted into the list and the tuples dominated by a current tuple are eliminated from the list. If a tuple is dominated by any other tuple, then it is eliminated. The main drawback of BNL is that the pairwise comparison in large databases may be too expensive.

In order to solve this problem, SFS [2, 3] is proposed. It first calculates the entropy score of all tuples and then presorts the data according to this entropy score. This strategy enables to eliminate most of the tuples in early comparison stage. The main drawback of SFS is that for large dataset with many attributes, calculating the entropy value of each tuple and sorting may take a long time. Several extensions [4] have been proposed to overcome the drawback of SFS. However, as these methods are based on sorting strategy, in large

and high-dimensional datasets, these methods have the same problem as SFS.

3. Entropy Curve Score

This section describes the proposed method to efficiently construct skyline for large and high-dimensional databases. The main idea of the proposed method is to calculate the entropy score similar to SFS algorithm. Instead of prior sorting of dataset, we propose to classify tuples according to the certain criteria, and using this, we propose a method to find the skyline quickly.

For classifying criteria, we use a topological feature that tuples with lower entropy scores has a high probability to dominate the other tuple. Based on this idea, we create and use a classification method of tuples based on entropy score, called entropy score curve. Each entropy score curve contains a set of tuples with the same tuples. And we propose a method that can quickly find the skyline with *entropy score curve*.

Our proposed method generates *entropy score* curve first, and classify tuples by comparison of the input tuple's *entropy score* and *entropy score curve*. After that, we create two windows for storing the classified tuples, one is the *comparison window* for tuples that have high probability to be skyline, another one is *storage window* for storing remaining tuples. Therefore, the input tuple is compared only with the *comparison window*. If an input tuple is not dominated by skyline candidates in the comparison window, it is stored into comparison window or storage window according to the *entropy score curve*. If the input tuple is dominated by skyline candidates, it is deleted immediately. Once the comparisons of all tuples are completed, the merger step performs comparison between comparison window and storage window for determining skyline points. Example 1 shows the scenario of finding skyline points of hotels.

Example 1: Figure 1 (a) shows the list of hotels with distance and price. Suppose a user wants to get the list of hotels with shortest distance to the beach and low price. Figure 2 (b) shows the data points in two-dimensional universe. Here, the entropy score curve was generated that divides given tuples into two windows, and based on this, tuples with the highest entropy score are stored in the comparison window and the remaining tuples are stored in the storage window. At first, p1 is stored in storage window without comparisons, because comparison window is initially empty. Also, p2 is stored without comparisons as well, because comparison window is empty, but this time p2 stored into comparison window by *entropy score curve*. p3 falls into the same curve with p2, however is not dominated by p2 and thus, stored into comparison window. p4 is dominated by p2 and eliminated in comparison processing. p5 is not dominated by skyline candidates in comparison window, so it is stored into storage window. p6 dominates the p3 in comparison window and then stored into comparison window. p7, p8, p9 is dominated by p2. p10 is not dominated by skyline candidates and stored into comparison window. Since then, through the merging step, <p1, p5> in storage window are compared with <p2, p6, 10> in comparison window. Since, <p1, p5> is dominated by <p2, p6>, skyline point is determined as <p2, p6, p10>.

ID	Distance	Price
p1	7	50
p2	1	50
p3	5	40
p4	3	85
p5	8	30
p6	2	30
p7	8	60
p8	4	70
p9	3	50
p10	3	20

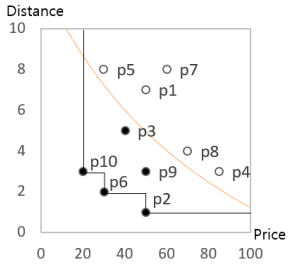


Figure 1. Example for suggestion method

4. Conclusions

In this paper, we have proposed a method which can reduce the processing time by effectively reducing the number of comparisons. Specifically, we have proposed a novel entropy score curve that does not require sorting the tuples. We have demonstrated the correctness of the proposed method with real-life example. In the future, we are planning to perform extensive experiment results that test the optimal number of curves needed to perform classification. Also, we are planning to compare the proposed method with other state-of-the-art methods to demonstrate the effectiveness of the proposed method.

5. Acknowledgement

This work was supported by the research grant of the Chungbuk National University in 2015.

References

[1] B. Stephan, K. Donald, and S. Konrad, “The skyline operator.”, *ICDE 2001*, 2001, pp.421-430.

[2] J. Chomicki, P. Godfrey, J. Gryz, and D. Liang, “Skyline with presorting.”, *ICDE 2003*, 2003, pp.717-719.

[3] J. Chomicki, P. Godfrey, J. Gryz, and D. Liang, “Skyline with presorting: Theory and optimizations.”, *Intelligent Information Processing and Web Mining*, Springer Berlin Heidelberg, 2005, pp.595-604.

[4] J. Chomicki, P. Ciaccia, and N. Meneghetti, “Skyline queries, front and back.”, *ACM SIGMOD Record*, Vol.42(3), 2013, pp.6-18.

An approach to detect TCP based attack using Data mining algorithms

Ugtakhbayar.N, Usukhbayar.B and Nyamjav.J
National University of Mongolia, Ulaanbaatar, Mongolia
44911.n@gmail.com

Abstract

Intrusion Detection Systems have become a necessary in computer networking security of largest networks. In the recent years, the system needs to identify new intrusion in largest datasets in a timely manner because internet to instantly access information at anytime from anywhere. That is a massive increasing of data traffic and internet nodes. Therefore, to refine IDS's performance and false alarm is a one of the important challenges in computer network security field. In this work we propose an approach to detect TCP connection based attacks using data mining algorithms. We gather raw network traffic and classify it into normal and abnormal traffic by Bro IDS. First we extract features in TCP headers of the packets such as sequence and acknowledge numbers, window size, control flags, and an event which is time between neighbor segments. Next, we evaluate the worth or merit of a features in novel attacks and select valuable subset of features. Finally, the selected features are given to learn the classifiers: J-48, Naïve Bayes. By adopting the concepts of machine learning and data-mining, we could detect 74% of novel attacks with 19 features.

Keywords: data mining, learning algorithms, IDS, intrusion detection

1. Introduction

Network security is still quickly developing in any information technology fields. In the last few years, due to the growing use of computer networks, network traffic is immediately increasing. There are several private as well as business sectors, government organizations that store valuable data over the computer network. Cause, new threats are showing up on quickly, while older often abide relevant. Therefore, more dynamic mechanisms such as Intrusion Detection Systems are should also be utilized. A Cisco report found the following: "Global IP traffic in 2012 stands at 43.6 exabytes per month and will grow threefold by 2017, to reach 120.6 exabytes per month, by 2019, there will be 24 billion networked devices and connections globally." [1].

Intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of possible attacks [2, 4]. Intrusion detection mechanism is divided in two; anomaly detection and misuse detection. Misuse detection is an approach where each suspected attack is compared to a set of known attack signatures [3]. It is in an exclusive manner the attacks in that database that can be detected, this method does not can for detection of unknown attacks. Unknown attack can be most zero day attacks. The role of anomaly detection is the identification of data points, substance, event and observations or attacks that do not conform to the expected pattern of a give collection [5].

Network traffic speeds and volume are increasing at an exponential rate. The conventional approach of tuning the hardware and software of the NIDS platform to maximize its performance can yield considerable improvements, but falls short in supporting next-generation networks operating at gigabits per second and faster.

This paper worked for data mining as a data processing technique, it can increasing detection ratio using data mining algorithms and decreasing processing time using feature selection method. We are using KDD 99 dataset and our university gateway traffic in this research. The KDD 99 dataset have been used for evaluating the most eminent available in the literature for feature selection and classification. KDD 99 dataset consists of nearly 5 million training connection records labeled as an intrusion or not an intrusion, and separate testing dataset consists of seen and unseen attacks [6]. In Methodology section, we presented Pearson correlation and J48, Naïve Bayes in our collected and KDD 99 datasets. In finally section we are summarize our paper and give final conclusion.

2. Related Work

Intrusion detection system, using artificial intelligence and data mining in intrusion detection

system. Intrusion detection and prevention systems used to detect and prevent the known and unknown attacks made by intruders. Moradi and Zulkernine, who are publisher of [8]. In this paper, there are presented an IDS that uses IDS for effective intrusion detection. One of the disadvantage of their approach is that it increases the time in training. In the literature [9] et al proposed a new method based on Continuous random function for selecting appropriate feature sets to perform network intrusion detection. Also, Liu and Gu have used Learning Vector Quantisation neural networks to detect attacks, that is supervised version of quantization, which can be used for pattern recognition, multi-class classification and data compression tasks [10]. In paper [11] has written a highly referenced article about intrusion detection using neural networks. In the article, he studies in detail the advantages and disadvantages of neural networks for this application. In the conclusion of this article that neural networks are very suitable for Intrusion detection system. The [12] have used a neural network to detect the number of zombies that have been involved in DDoS attacks. The objective of their work is to identify the relationship between the zombies and in sample entropy. Shon and Moon used genetic algorithm to extract optimized information from raw internet packets [13]. Ruchi Jain and Nasser S. Abouzakhar applied J48 decision tree algorithm to determine significant features from KDDCUP 1999 dataset for anomaly intrusion detection [14]. And experimental results demonstrate that the Hidden Markov model is able to classify network traffic with approximately 76% to 99%. Most proposed techniques utilize characteristics of network traffics to identify abnormalities absolutely. But, performing the real time network traffic detection with maintaining higher accuracy is restricted due to complex nature of networks.

3. Methodology

This research focuses on solving the issues in Intrusion Detection methods that can help the network and system administrators to make pre-processing, classification of network traffic. Most of attacks can be identified only after it happens. Data mining approaches have been implemented by many researchers to solve the abnormal detection problem. In this section, we are explain the proposed methodology for anomaly intrusion detection. We concentrated on data mining such as J48 algorithm, Naïve Bayes classifiers, because data mining approaches use strong statistical foundations to enhancing the dynamic and accurate learning that gives better accuracy, reduce false alarm rates, performance improvements, ability to detect novelty, protection against zero-day exploits.

The entire framework of proposed methodology shown in figure 1, we are collected our university's internet and intranet traffic using Bro IDS by sensor. Our method consists of two stage. In stage 1, collecting data with real time intrusion detection analyzer with Bro IDS system. In stage 2, in the figure with red frame, feature selecting in KDD 99 dataset and train to data mining algorithms.

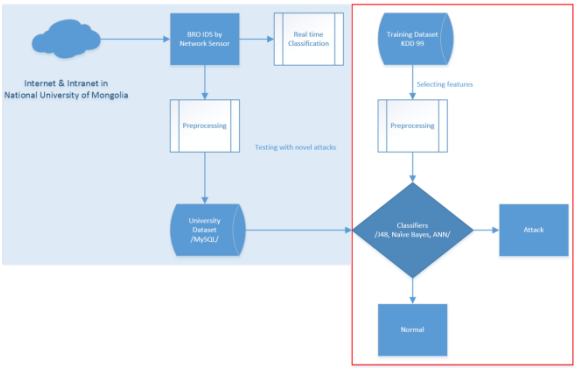


Figure 2 Anomaly detection Proposed method

In our approach, the dataset is divided into training and testing datasets. Data for the research paper originated from two sources. First, training data sets includes KDD 99's labeled datasets. The labeled datasets are applied to J-48, Naïve Bayes, classifier and the model is generated. The second set of data was from our collecting datasets from National University of Mongolia's gateway router. The router located inside of firewall. Our sensor system Bro IDS is running with the specifications of 2nd generation Intel Atom Dual core processor, 2GB DDR3 RAM disk, 128GB SSD hard disk. Also, we are collected testing dataset with novel attacks using Backtrack system collected by netflow, tcpdump. In the preprocessing section, our system is designed by applying feature extraction and feature selection. The dataset contains divers attack type that could be classified into four main categories. The dataset has 41 features for each connection record. The features divided into three categories that are host features, service features, traffic features.

4. Result of simulation

KDD 99 dataset has been used in this research work of which 100% is treated as training data. We are implemented the proposed method in Weka data mining tool. This tool contained the tools required for the analysis.

Our computing environment for this paper included Ubuntu operating system that ran on a Dell desktop. The system's hardware consists of an Intel I5 processor, 16GB of memory and 2TB hard disk space.

We selected important features using the Markov blanked model [15, 16]. And found that 16 features of the dataset form the Markov blanket. These 19 features are "duration, protocol-type, service, src_bytes, land, wrong_fragment, num_failed_logins, logged_in, root_shell, num_file_creations, num_outbound_cmds, is_guest_login, count, srv_count, serror_rate, srv_serror_rate, diff_srv_rate, dst_host_count, dst_host_srv_count". Moreover, a J-48, Naïve Bayes classifiers was constructed using the KDD 99 dataset and then the classifier was used on the our collected dataset to classify the data as an attack or normal data. The result shown in table 1 and table 2.

Attack class		*J-48*	*Naïve Bayes*
Normal		76.4	56.1
Probe		89.2	93.6
DOS		98.2	95.1
U2R		45.7	33.2
R2L		66.3	90.4
	Over all	75.16	73.6

Table 1 Detection ratio with full feature dataset

Attack class		*J-48*	*Naïve Bayes*
Normal		75.1	53.2
Probe		88.3	92.1
DOS		98.9	96.2
U2R		45.2	10.6
R2L		63.8	73.4
	Over all	74.26	65.1

Table 2 Detection ratio with selected 19 feature dataset

5. Summary and Future work

In this paper, we present an intrusion detection system using J-48 and Naïve Bayes for classification and data mining methods for feature selection. To implement and classify of our system we used KDD 99 dataset and our University's traffic. The principal challenge in intrusion detection is to obtain high detection rate and reduce false alarm rate with novel attacks. In our experimental result shown as single classifier is not sufficient to obtain the high result and feature selection is the most important to detection ratio. Therefore maybe more than one classifier can be improve the attack detection ratio. Further we will do feature selection differently to improve the results and choose the more than one classifier.

6. References

[1] Cisco, *"Cisco Visual Networking Index: Forecast and Methodology"*, 2012-2017, Cisco, 2013.

[2] Guide to Intrusion Detection and Prevention Systems (IDPS), *"Recommendations of the National Institute of Standards and Technology"*, Technology Administration U.S. Department of Commerce. NIST Special Publication 800-94.

[3] A. K. Pathan, "*The State of the Art in Intrusion Prevention and Detection",* CRC Press, 2014.

[4] Li Hanguang, Ni Yu, "*Intrusion Detection Technology Research Based on Apriori Algorithm",* 2012 International Conference on Applied Physics and Industrial Engineering, Physics Procedia 24, 2012, 1615 – 1620

[5] M.Naga Surya Lakshmi, Dr. Y. Radhika *"A complete study on intrusion detection using data mining techniques"* Volume IX, IJCEA Issue VI, June 2015

[6] Miroslav Stampar *"Artificial Intelligence in Network Intrusion Detection"*

[7] Senthilnayaki Balakrishnan, Venkatalakshmi K, Kannan A *"Intrusion detection system using Feature selection and Classification technique"* IJCSA Volume 3, Issue 4, 2014, pp. 144-151.

[8] Moradi M and Zulkernine M *"A Neural Network based System for Intrusion Detection and Classification of Attacks"*, Proceedings of IEEE International Conference on Advances in Intelligent Systems – Theory and Applications, Luxembourg, Vol. 148, 2004, pp. 1-6.

[9] Wang Jianping, Chen Min and Wu Xianwen, *"A Novel Network Attack Audit System based on Multi-Agent Technology",* Physics Procedia, Elsevier, Vol. 25, 2012, pp. 2152 – 2157.

[10] J.Li, Y.Liu and L.Gu *"DDoS attack detection based on neural network"*: Proceedings of the 2nd International Symposium on Aware Computing (ISAC), Tainan, 1–4 Nov.2010, pp.196–199.

[11] James Cannady. *"Artificial neural networks for misuse detection".* In Proceedings of the 1998 National Information Systems Security Conference (NISSC'98) October 5-8 1998. *Arlington, VA.*, 1998, pages 443–456.

[12] B.B. Gupta, C.Joshi and M.Misra *"ANN based scheme to predict number of zombies in a DDoS attack"*, International Journal of Network Security. 13(3)(2011)216–225.

[13] T. Shon and J. Moon, *"A hybrid machine learning approach to network anomaly detection,"* Inf. Sci., vol. 177, no. 18, 2007, pp. 3799–3821, Sep.

[14] R. Jain and N. Abouzakhar, *"A comparative study of hidden markov model and support vector machine in anomaly intrusion detection,"* Journal of Internet Technology and Secured Transactions (JITST), vol. 2, no. 1/2/3/4, 2013, pp. 176–184.

[15] Cho S-B, Park H-J, *"Efficient anomaly detection by modeling privilege flows with hidden Markov model."* Computers and Security, 22(1), 2003, pp. 45-55.

[16] Tsamardinos I, Aliferis CF, Statnikov A, *"Time and sample efficient discovery of Markov blankets and direct causal relations."* 9th ACM SIGKDD international conference on knowledge discovery and data mining, ACM press, 2003, pp. 673-678.

.

Simulations Studies of Switching Arc Behaviour in High Voltage Puffer Type SF6 Circuit Breakers

Kai Shen Ee[1], Dingkun Li[2], Yu Fu[3], Keun Ho Ryu[*]
[1,2,3,*)]*Database and Bioinformatics Laboratory, School of Electrical and Computer Engineering, Chungbuk National University, South Korea*
{[1]kaishen, [2]jerryli, [3]fuyu, [*]khryu}@dblab.chungbuk.ac.kr

Abstract

During the last two decades, major improvement in the technology and design of high voltage sulfur hexafluoride (SF6) gas circuit breakers (GCBs). The main reason of advancement of GCB in technology is due to the progress made in the numerical simulation and the availability of powerful personal computers. There are various type of SF6 breaking technology in the market, where the puffer type Circuit Breaker using SF6 gas as a current-interrupting medium are widely used in most of high-voltage circuit breaker. The reason being is SF6 gas has brilliant insulation and current-interruption properties. Besides, puffer structure is very simple and reliable in circuit breaker design. Computation simulation of the arc behaviour in high voltage puffer type SF6 circuit breaker has been performed by using a commercial computational fluid dynamics (CFD) package, PHOENICS[1].

Keywords: Arc behaviour, SF6, circuit breakers, CFD, Puffer type, PHOENICS

1. Introduction

High voltage circuit breakers are used in the transmission substations to connect, isolate and switch among different sets of circuits. Circuit breaker play an important role in power system. When circuit breaker is closed, it is an ideal conductor, when it is open, it serves as ideal insulator. When a circuit breaker is closed, it can respond to interrupt normal or fault current within an interval during which system stability is not lost and no serious damage to equipment while avoiding system over voltage. Meanwhile, when it is open, it should be able to close under normal and short-circuit conditions where faults may not be permanents. Because circuit breaker is so crucial, the requirements to consider when design and manufacture a circuit break is its lifespan, reliability, size and cost. Circuit breaker is a current interruption or switches in a power system. It helps to de-energize part of a power system for maintenance and repair, isolate the part of the system where a fault has developed and control of power flow in a power system.

[1] PHOENICS is provided by CHAM Ltd, London, UK

2. The Arc Model

High voltage circuit breakers are used in the transmission substations to connect, isolate and switch among different sets of circuits. Circuit breaker play an important role in power system. When circuit breaker is closed, it is an ideal conductor, when it is open, it serves as ideal insulator. When a circuit breaker is closed, it can respond to interrupt normal or fault current within an interval during which system stability is not lost and no serious damage to equipment while avoiding system over voltage. Meanwhile, when it is open, it should be able to close under normal and short-circuit conditions where faults may not be permanents. Because circuit breaker is so crucial, the requirements to consider when design and manufacture a circuit break is its lifespan, reliability, size and cost. Circuit breaker is a current interruption or switches in a power system. It helps to de-energize part of a power system for maintenance and repair, isolate the part of the system where a fault has developed and control of power flow in a power system.

2.1. Governing Equations

Low frequency electrical arcs at atmospheric pressure or above have a state close to local thermal equilibrium (LTE) because of the frequent collisions between heavy particles and electrons. The geometry of high voltage circuit breakers is axisymmetric except some of the minor details and assume that the arc together with its surrounding gas flow is axis-symmetric and in LTE state. The system can be mathematically described by the time-averaged Navier-

Stokes equations taking account of Ohmic heating, Lorenz force electromagnetic effect, nozzle ablation, radiation loss and turbulence enhanced mass, momentum and energy transport. The general conservation equations can be written in the follow form:

$$\frac{\partial(\rho\phi)}{\partial t} + \nabla \cdot \left(\rho\phi\vec{V}\right) - \nabla \cdot \left(\Gamma_\phi \nabla\phi\right) = S_\phi \qquad (1)$$

In a 2D axisymmetric system using cylindrical coordinates, it will be

$$\frac{\partial(\rho\phi)}{\partial t} + \frac{1}{r}\frac{\partial}{\partial r}\left[r\rho v\phi - r\Gamma_\phi \frac{\partial\phi}{\partial r}\right] + \frac{\partial}{\partial z}\left[\rho w\phi - \Gamma_\phi \frac{\partial\phi}{\partial z}\right] = S_\phi \qquad (2)$$

The source terms (S_ϕ) and the diffusion coefficients (Γ_ϕ) are listed in Table 2.1 for different conservation equations, in which h is the enthalpy, μ the dynamic viscosity, P the pressure, j the current density and B the magnetic flux density. k is the thermal conductivity, $C_p$the specific heat at constant pressure, σ the electrical conductivity, E the electric field, q the net radiation loss per unit volume and t the time. C_m is the Polytetrafluoroethylene (PTFE) mass concentration and D the diffusivity. The subscript l denotes the laminar part of the transport coefficient and t the turbulent part. Viscous stresses are taken into account in the diffusion terms in the two momentum equations in Table 2.1. The part of viscous stresses in the radial momentum equation which cannot be written as part of a diffusion term is included in the source term. It has been found that molecular viscous effects are negligible in momentum balance for arcs in a supersonic nozzle. Viscous heating due to molecular and turbulent stresses is given in the source term for the enthalpy equation. The effects of Lorentz force generated by the interaction of the arc current with its own magnetic field, which is included in the momentum source term, can be neglected for low current arcs, which is below 2 kA in nozzle.

Table 1. Terms of governing equations

Equation	ϕ	Γ_ϕ	S_ϕ
Continuity	1	0	0
Z – momentum	W	$\mu_l + \mu_t$	$-\frac{\partial\rho}{\partial z}$
R – momentum	v	$\mu_l + \mu_t$	$-\frac{\partial\rho}{\partial r} + J\,x\,B - (\mu_l + \mu_t)\frac{v}{r^2}$
Enthalpy	h	$\frac{k_l + k_t}{C_p}$	$\frac{\partial\rho}{\partial t} + \sigma E^2 - q + (\mu_l + \mu_t)\{2\left[\left(\frac{\partial v}{\partial r}\right)^2 + \frac{v^2}{r^2} + \left(\frac{\partial w}{\partial z}\right)^2\right] + \left(\frac{\partial v}{\partial z} + \frac{\partial w}{\partial r}\right)^2\}$

2.2. Governing Equations

Ohmic heating (σE2) is included in the source term of energy and momentum conservation equations in Table 3.1. The electric conductivity is a function of temperature and pressure, which could be calculated by linear interpolation from a data table. In the high current phase, the radial size of the arc section varies along the axial direction significantly and the radial current density component is comparable with the axial one, the electric field needs thus to be calculated by the so called "Non-slender arc model", meaning the radial component of the electric field cannot be ignored and the electric potential equation has to be solved. The electrical potential is calculated by solving the current continuity equation, which is expressed as

$$\nabla \cdot (\sigma \nabla \varphi) = 0 \qquad (3)$$

where σ is the electrical conductivity and φ the electrical potential. To solve such equation with only the diffusion term, the diffusion coefficient can never be zero in order that a solution exists in the whole domain. It has been shown that Equation (4) produces correct results of the electric field in and around the arc section where the Lorentz force and Ohmic heating is needed. So for low temperature gas or insulating material, an electrical conductivity of 10-3 Ω-1m-1 is used. The radial and axial current density can then be calculated from the electrical potential distribution by

$$j_r = -\sigma\,\partial = \frac{\partial\varphi}{\partial r}\,;\quad j_z = -\sigma\,\partial = \frac{\partial\varphi}{\partial z} \qquad (4)$$

On all boundary surfaces of the domain for electric potential calculation except that intersecting with the current conducting contacts, a condition of zero current density is imposed, which is also the default PHOENICS boundary [2] condition on walls meaning that no current flows across the boundary. Because the boundaries for electric potential calculation are sufficiently away from the arc region, the electric field solution inside the arc section and in its vicinity is not affected by the use of the zero current density condition. The potential on the boundary surface where the live contact intersects with the boundary is set to zero, as a reference value.

2.3. Modelling of turbulence

The switching arc is turbulent, especially at the vicinity of current zero. In the present investigation,

[2] Boundary condition in PHOENICS. is retrieved from http://www.cham.co.uk/phoenics/d_polis/d_lecs/general/bcond.htm.

turbulence is modelled by the simple Prandtl mixing length model. The eddy viscosity is given by

$$\mu_t = \rho(cr_\delta)^2 \sqrt{(\frac{\delta w}{\delta r})^2 + (\frac{\delta v}{\delta z})^2} \quad (5)$$

where ρ is the gas density, c the turbulence model parameter and rδ a characteristic dimension of the arc section which is defined in the present work as the radial distance from the axis to the point of 5,000 K for the high current phase. The choice of 5,000 K is based on the observation that a large portion of the arcing space can sometimes be filled with hot vapour from nozzle ablation during the high current phase and the use of a lower temperature could lead to erroneous values for the characteristic dimension rδ. For current zero and post arc current simulation, the use of 5,000 K will lead to an underestimate of the size of the high speed jet where turbulence mixing is strong. SF6 at 5,000 K is still conducting. So in the current zero phase and the subsequent post arc current calculation, rδ is defined as the radius of 3,000 K.

3. Implementation of Arc model in PHOENICS

The model is implemented in PHOENICS. PHOENICS has mainly three parts. The first part is the Satellite with an input file called Q1. Q1 contains definitions of all important model parameters, the grid system, the definition of governing equations, the boundary conditions and the relaxation control. Computation is performed by Earth which is compiled from PHOENICS library files and a Fortran file called Ground.for where the arc model is coded. The visualisation of the results is done by a postprocessor called PHOTON. For details of the structure of each part, please refer to the PHOENICS manual which is provided for each license.

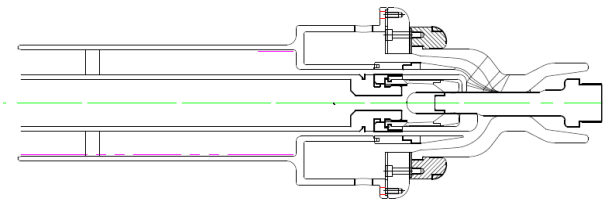


Figure 1. Geometry example of circuit breaker

4. Author name(s) and affiliation(s)

Computation has been performed for the arching chamber of puffer type circuit breaker as shown in Figure 1. Computationally, it is more convenient to move the solid contact rather than the hollow contact and nozzle assembly as in practice. This will not cause much error as the speed of the moving parts is much smaller than the sound speed of SF6 at room temperature. The circuit breaker is initially filled with SF6 at an absolute pressure of 0.6MPa at room temperature

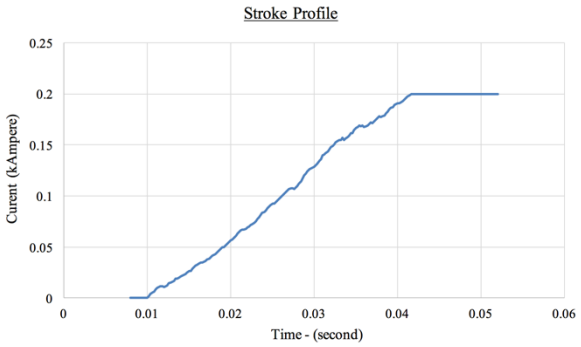


Figure 2. Model C stroke profile

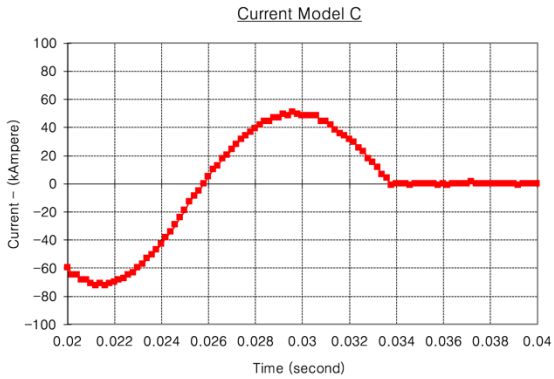


Figure 3. Model C fault current profile

The result and discussion of computational simulation can be break down into few stages include, before first current zero stage, current zero stage and after current zero stage. All the date can be obtain from Figure 2 and 3 which are generated from simulation. Figure 4. is an example of simulation of circuit breaker in PHOENICS.

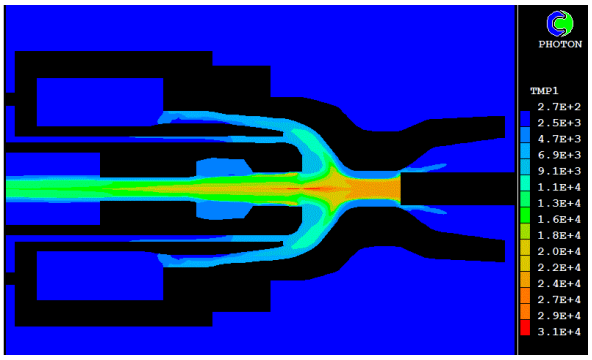


Figure 4. Temperature before first current zero.

5. Conclusion

The simulation of SF6 circuit breaker has been performed successfully with the commercial code PHOENICS. The results show that with the circuit breaker under investigation, pressurization of the thermal expansion chamber is a direct consequence of heating of SF6 gas from the main nozzle. The ablated vapor from the nozzle mainly leaks into the flow passage towards the piston chamber and does not contribute to the pressure rise in the thermal expansion chamber.

6. Acknowledgment

I wish to thanks all the members of Tenaga Nasional R&D Sdn. Bhd. (TNBR) which include Associate Professor Dr. C.K. Ng from Universiti Tenaga Nasional, C.N. Saniyyati, research assistant and Dr. H.M. Looe from TNBR for their valuable discussion and knowledge, which assisted me a lot in my thesis and understanding to the project. Besides, I would like to express my greatest appreciation to undergraduate supervisor, Assistant Professor Dr. M.M Al Iman and previous supervisor, Assistant Professor Dr. V.K. Liau for their support and continuous encouragement throughout this project.

7. References

[1] S. Yanabu, H. Mizoguchi, H. Ikeda, K. Suzuki, and M. Toyoda, "Development of Novel Hybrid Puffer Interrupting Chamber for SF6 Gas Circuit Breaker Utilizing Self-pressure-rise phenomena by Arc.", *IEEE Trans. On Power Delivery*, Vol.4(1), 1989, pp.355-361.

[2] J.Y. Trepanier, M. Reggio, and Y. Lauze, "Analysis of the dielectric strength of an SF6 circuit breaker.", *IEEE Trans. On Power Delivery*, Vol.6(2), 1991, pp.809-815

[3] M. Okamoto, M. Ishikawa, K. Suzuki, and H. Ikeda, "Computer Simulation of Phenomena associated with Hot gas in puffer-type gas circuit breaker.", *IEEE Trans. On Power Delivery*, Vol.6(2), 1991, pp. 833-839.

[4] K.Y. Park and M.T.C. Fang, "Mathematical Modeling of SF Puffer Circuit Breaker 1 High Current Region.", *IEEE Trans. On Plasma Science*, Vol.24(2), 1996, pp.490-502.

[5] P. Chevrier, M. Barrault, C. Fievet, J. Maftoul, and J.M. Fremillon, "Industrial applications of high, medium and low voltage arc modelling.", *J. Phys. D: Appl. Phys.* Vol.30, 1997, pp.1346-1355.

[6] K.Y. Park, X.J. Guo, R.E. Blundell, M.T.C. Fang, and Y.J Shin, "Mathematical Modeling of SF Puffer Circuit Breeaker 2 Current Zero Region", *IEEE Trans. On Plasma Science*, Vol.25(5), 1997, pp.967-973.

[7] X.B. Li, Q.P. Wang, Y.B. Li, and Y. Yang, "Numerical Analysis of Flow Field and the dynamic properties of arc in the interrupting chamber of an SF6 puffer circuit breaker.", *IEEE Trans. On Plasma Science*, Vol.25(5), 1997, pp.982-985.

[8] J.J. Lowke, R. Morrow, and J. Haidar, "A simplified unified theory of arcs and their electrodes.", *J. Phys. D: Appl. Phys*, Vol.30(14), 1997, pp.2033.

[9] J.D. Yan and M.T.C. Fang, "The development of PC based CAD tools for auto-expansion circuit breaker design.", *IEEE Trans. On Power Delivery*, Vol.14, 1999, pp.176-181.

[10] J.L. Zhang, J.D. Yan, A.B. Murphy, W. Hall, and M.T.C. Fang, "Computational Investigation of Arc Behavior in an Auto-Expansion Circuit Breaker Contaminated by Ablated Nozzle Vapor.", *IEEE Trans. On Plasma Science*, Vol.30, 2002, pp.706-19.

[11] C.M. Dixon, J.D. Yan and M.T.C. Fang, "A comparison of three radiation models for the calculation of nozzle arcs.", *J. Phys. D: Appl. Phys*, Vol.37, 2004, pp.3309-3318.

[12] T. Mori, H. Kawano, K. Iwamoto, Y. Tanaka, and E. Kaneko, "Gas-Flow Simulation With Contact Moving in GCB Considering High-Pressure and High-Temperature Transport Properties of SF6 Gas.", *IEEE Trans. On Power Delivery*, Vol.20(4), 2005, pp.2466-2472.

[13] S.H. Park, C.Y. Bae, H.K. Kim, and H.K. Jung, "Computer Simulation of Interaction of Arc-Gas Flow in SF6 Puffer Circuit Breaker Considering Effects of Ablated Nozzle Vapor.", *IEEE Trans. On Magnetic*, Vol.42(4), 2006, pp 1067-1070.

[14] M. Bartlova, N. Bogatyreva, V. Holcman, and J. Burstlova, "Approximate Description of Radiation Transfer in SF6 and PTFE Arc Plasmas.", *Proceedings in Research Conference in Technical Disciplines*, 2013, pp.16-20.

Traditional Mongolian Script Segmentation

Suvdaa Batsuuri
School of Engineering and Applied Sciences,
National University of Mongolia
suvdaa@num.edu.mn

Abstract

Mongolian many historical and cultural documents are stored as books, whch are written in traditional mongolian script. Therefore, traditional script recognition is one of the important topics in Mongolia.

In this paper, we have summarize results of segmentation of the traditional mongolian script. There are 3 methods are discussed; (1) handwritten or italic script segmentation by computing the slop, (2) segmentation by backbone width, and (3) segmentation by removing backbone. Specially, in the 'modon bar' format which is printing style without any separated parts called "dusal".

In experiments, we use "Ganjuur Danjuur sudar" cultural valuable book with 102 pages. In this time, we test 3 pages, 84 columns, 703 words, 2500 scripts. As a results, we achieved the column segmentation rate 100%, the word segmentation rate 98% and script segmentation rate 85%.

Effect of Cognitive Distraction on Driving Behaviour

Basabi Chakraborty[1], Yusuke Manabe[2], Sho Yoshida[3], Kotaro Nakano[3]
Faculty of Software and Information Science, Iwate Prefectural University, Japan[1],
Faculty of Information and Computer Science, Chiba Institute of Technology, Japan[2],
Graduate School of Software and Information Science, Iwate Prefectural University, Japan[3]
basabi@iwate-pu.ac.jp[1], ymanabe@net.it-chiba.ac.jp[2],
{g231n033, g231n024}@s.iwate-pu.ac.jp[3]

Abstract

In this work, the effect of driver's distraction due to cognitive load has been studied by analyzing data of driving behaviour from driving simulator. The simulation experiments are done with 4 drivers and three type of driving situation (normal driving and driving with various types of cognitive loads). From the analysis of time series data obtained from sensors in a driving simulator, it has been noticed that driving behavior changes with statistical significance for varying cognitive tasks. The feature of the driving simulator data that changes most with increasing cognitive load has been assessed. The classification accuracy for automatic detection of driving with and without cognitive load from sensor data by a simple classifier came out on the average as 66.3%.

Keywords: Driving behaviour, cognitive distraction, cognitive load, detection of distracted driving, driving simulator

1. Introduction

Driver distraction due to secondary activities while driving, is considered to be one of the main reason of road accidents [1]. Automatic detection of distraction and issuance of alert can help driver to adhere to safe driving. Many researches [2] are going on to model driving behavior and automatic detection of driver's distraction, yet to come up with a successful commercial application. It is known that driving behavior is affected by fatigue, visual and cognitive distraction and can be modelled from the data obtained by physical, physiological or environmental sensors. The effective method of automatic detection of driver's distraction or the effective combination of the sensors for detection are now the objectives of research.

Generally two major types of distractions are visual distraction and cognitive distraction. Visual distraction happens when the driver looks away from the road described as eyeoff- road, cognitive distraction occurs when the driver's mind is busy with something not directly related with driving known as mind-off-road. Visual distraction can be automatically detected by tracking the driver's eye movement. A general algorithm that considers driver's glance behaviour across a relatively short period, could detect visual distraction consistently across drivers. Some research works in this direction are presented in [3]. However, detecting cognitive distraction is much more complex as the signs of cognitive distraction are usually not straight forward and can vary across drivers. Moreover the driving behavior does not have a simple linear relationship with cognitive distraction. Some studies on cognitive distraction can be found in [4].

In this work, we restrict our study to the effect of cognitive distraction on driving behaviour of the driver and how it can be detected from the sensors available with the driving simulator and whether it is affected by the nature of the cognitive load. It is known from various studies [5] that the driving behaviour changes when the driver is exposed to cognitive demand of the secondary task. The change of driving behaviour may depend on individual or there might be some common trend. In this study, experiments are done with a driving simulator in different scenarios and drivers are asked to drive 1) with attention without any secondary task 2) with various secondary tasks. The sensors' data from driving simulator are collected and analyzed. Statistical tests are done to check whether there is any significant difference between the driving behaviour with and without secondary cognitive tasks and what feature or which sensor data indicates the most difference during driving with attention and driving with distraction.

The next section presents some related works followed by the section describing experimental studies and results. The final section presents summarization and conclusion.

2. Related Works and Background

The area of modelling driving behaviour is gaining an increasing attention in the research of safety driving. Some studies considered driving simulators [6] and some others used real cars equipped with various sensors [7]. The main objective is to automatically assess the distraction level of the driver that have considerable effect on driving performance. Frontal cameras can be useful to assess the visual distraction level of the driver. Relevant visual features include head pose, gaze and eyelid movements. In [8], an approach to monitor visual distractions using a low cost camera is presented. Measuring cognitive distraction from cameras or any nonintrusive sensors is difficult as many parameters need to be integrated for detecting cognitive distraction. In [9] [10] Support Vector Machines (SVM) and Dynamic Bayesian Networks (DBN) are used respectively for detecting cognitive distraction from driver's visual behaviour and driving performance. Car information also provide valuable information about the driver's behaviour. In [6] driver's behaviour is predicted from pedal position. In [11], driver's attention level is calculated from steering wheel, vehicle speed and lateral position of the car. There are some other research works [12] in which car component features like wheel angle, gas pedal or brake pressure are used for assessing driver's behaviour. Studies have also done considering physiological signals to assess cognitive load, attention and fatigue. EEG and bio signals are also used in some researches for detecting drowsiness and sloppy driving behaviour. But for measuring physiological signals one needs to use intrusive sensors which are not convenient for real life situation.

3. Experimental Study

In this study, we have used driving simulator D3Sim. The driving behaviour is assessed from the simulator output which contains time series data (steering angle, steering torque, accelerator stroke, brake stroke, car speed, car angle, engine speed etc). We have used various scenarios for driving and collected simulator output. The experimental study in detail is as follows.

1) 4 subjects have been used for this study. All of them are students in the age group 20-22 yrs.
2) For each subject, driving data for three situations have been collected: a) normal driving with attention b) driving while continuing conversation with co passenger c) driving while doing mental arithmetic at the elementary school level.
3) For each situations, different driving scenarios are used for example, simple route, route having curves and sharp bending and routes with multiple diversions.
4) All subjects are initially allowed to practice for a while in different routes. Each subject is then asked to drive in the designated routes (from simple to complex) consecutively and repeat driving for several times.
5) The time series output data from the driving simulator for steering wheel angle, steering torque, accelerator torque, brake stroke, car speed and engine speed are recorded.
6) Statistical analysis have been done for finding out the best parameter for distinguishing difference og driving behaviour for attentive driving and distracted driving.
7) Nearest neighbour classifier and Support Vector machine classifier are used to classify driving with cognitive load and without cognitive load.

4. Analysis and Results

In this study we selected 80 driving samples (for each person, normal driving 11 times, driving with conversation 4 times and driving with mental arithmetic 5 times) for analysis. The drivers were asked to drive following a car speeding 60 km per hour with a more or less constant separation. The route scenario used are as follows:

- Mountain roads
- Only curve roads without left/right turn
- Roads with ups and down

Each driver is asked to drive for 3 minutes in each route and repeat the course for 5 times. The parameters used from each time series data are maximum value σ_k , variance σ_k^2 and average value μ_k as in the following:

$$\sigma_k = max_1 \leq t \leq N_k(|y_k(t)|)$$
$$\sigma_k^2 = \frac{1}{N_k}\sum_{t=1}^{N_k}(|y_k(t) - \mu_k|)^2$$
$$\mu_k = \frac{1}{N_k}\sum_{t=1}^{N_k}|y_k(t)|$$

where $y_k(t)$; k(= 1, 2, ⋯, 6) is the time series data for kth series, k representing each of the 6 time series data collected from driving simulator. N_k , is the number of time intervals from beginning to end of the driving.

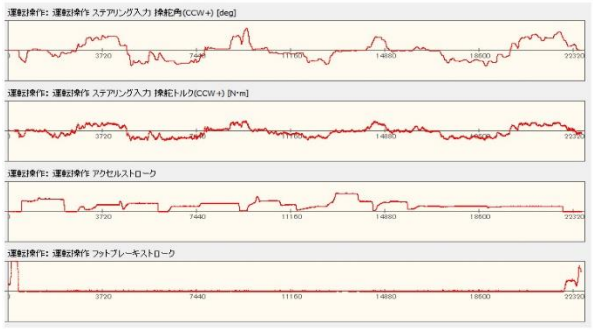


Figure 1. Data for Normal driving

Figure 2. Data for driving with conversation

Now for every feature and for every series, statistical significance is tested for confirming significant difference between normal and distracted driving. 1NN classifier and SVM with RBF kernel is used to classify the data of driving.

Figure 1, Figure 2 and Figure 3 represent the time series data from driving simulator for different time series associated with steering angle, steering torque, brake stroke, car speed etc. for normal driving and driving with various cognitive tasks. It can be found from visual inspection of the data that steering angle and steering torque show difference in case of driving with or without cognitive load. Moreover it is found that the difference is larger for driving with conversation than driving with simple mental arithmetic.

Figure 4 to Figure 8 represent the histograms of the features having significant difference for the two classes, normal and driving with conversation. It is also observed from statistical analysis that for driving with conversation, braking behaviour of driver changes, driver uses strong brakes and the number of use of brakes also increases. Similarly, steering wheel angle has also large changes in case of driving with conversation. Car speed is also slightly greater for the driving with conversation.

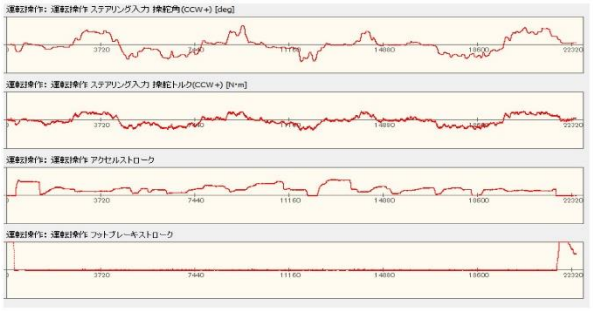


Figure 3. Data for driving with mental arithmetic

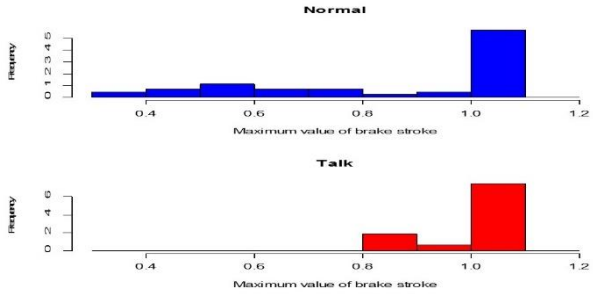


Figure 4. Maximum value of brake stroke (Normal vs Conversation)

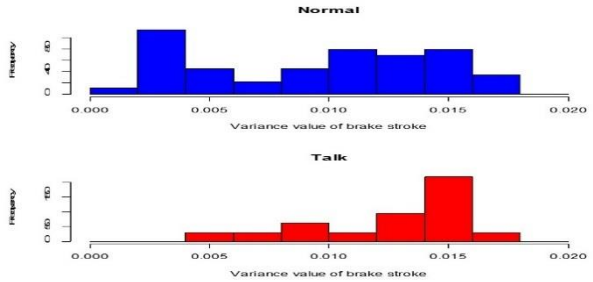


Figure 5. Variance of brake stroke (Normal vs Conversation)

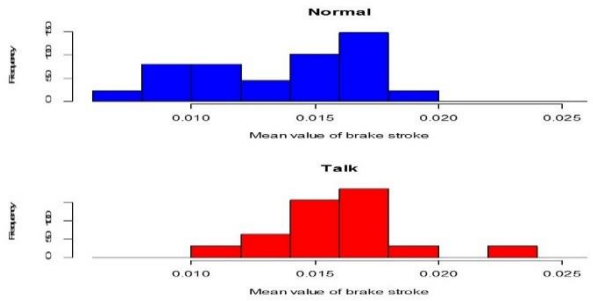


Figure 6. Average of brake stroke (Normal vs Conversation)

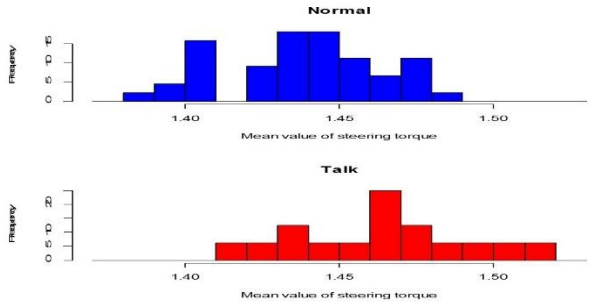


Figure 7. Average of steering torque (Normal vs Conversation)

Figure 9 to Figure 11 represent the histograms of features having significant difference for two classes normal and driving with mental arithmetic. It seems that brake stroke is the most important feature for distinguishing these two classes.

Using the best features from the statistical analysis, SVM is used to classify two classes of driving. Table 1 represents the results for the best values obtained. We have tried nearest neighbor classifier (1NN) also but we could achieve the average accuracy of classification as 59%.

5. Conclusion and Discussion

In this work a preliminary study has been done in order to assess the effect of cognitive load on driving behaviour and to detect distracted driving automatically from the sensor's data for recording driving and car behaviour. The study has been done using driving simulator and the driving behavior is assessed by analyzing the collected data from in-built nonintrusive sensors of the simulators. In the simulation study with young subjects, different levels of cognitiove loads were imposed on the driver to simulate cognitive distraction. The subjects were asked to engaze in conversation with the copassanger and to do mental arithmetic while driving.

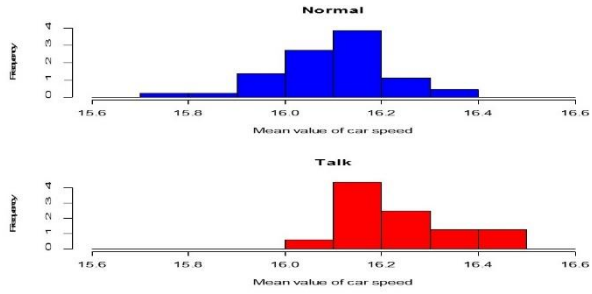


Figure 8. Average of car speed (Normal vs Conversation)

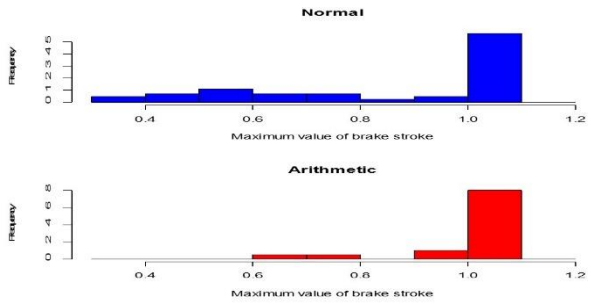


Figure 9. Maximum value of brake stroke (Normal vs Arithmetic)

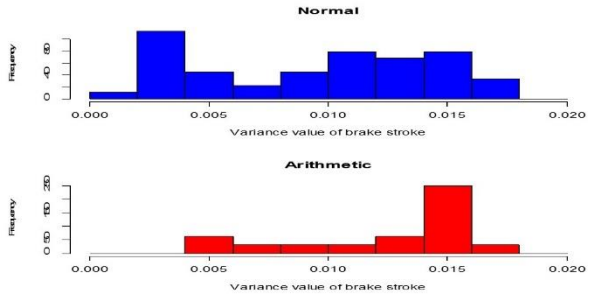


Figure 10. Variance of brake stroke (Normal vs Arithmetic)

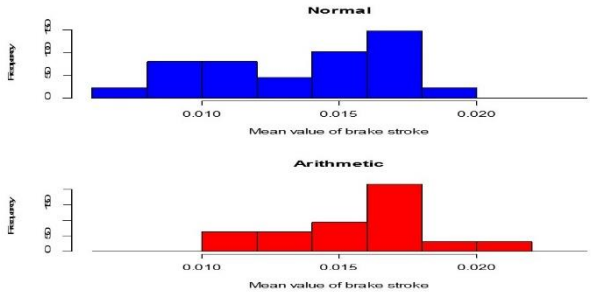


Figure 11. Average of brake stroke (Normal vs Arithmetic)

Table 1. Classification Results with SVM

		Predicted	
		Normal	With cognitive load
True	Normal	68.2%	31.8%
	With cognitive load	36.1%	63.9%

It has been observed that there is significant difference in terms of statistical test in driving behavior for attentive driving and distracted driving. From the analysis of the collected data it has been found that the braking behaviour of the driver is an important attribute of the driving behaviour which is affected the most. So it is used for classifying normal and driving with cognitive loads. The steering angle also is influenced by driving with cognitive load. The

classification accuracy is not very good but as a starting point this study is promising for further investigation. Other factors of driving behaviour can be integrated with braking stroke for improving the classifier accuracy.

For future study, we need to integrate other factors influencing cognitive distraction and also use some other sensors to detect cognitive distraction for more concrete results and increased classification accuracy for normal and distracted driving.

6. References

[1] D. Ascone, T. Lindsay and C. Varghese, "*Traffic Safety factor: An Examination of Driver Distraction as Recorded in NHTSA Databases*", National Highway Traffic Safety Administration's National Center for Statistical Analysis, 2009.

[2] L. Jin, Q. Niu, H. Hou, et.al., "*Driver Cognitive Distraction Detection Using Driving Performance Measures*", Discrete Dynamics in Nature and Society, Vol.2012, article ID 432634. 12 pages, 2012.

[3] Y. Liang, J. D. Lee and L. Yekhshatyan, "*How dangerous is looking away from the road? Algorithms predict crash risk from glance patterns in naturalistic driving*", Human Factors, Vol. 54, No. 6, 2012, pp. 1104-1116.

[4] Y. Liang and J. D. Lee, "*A hybrid Bayesian Network approach to detect driver cognitive distraction*", Transportation Research Part C, Vol. 38, 2014, pp. 146-155.

[5] N. Li, J. Jain and C. Busso, "*Modeling of Driver Behavior in Real World Scenarios using Multiple Noninvasive Sensors*", IEEE Trans. On Multimedia, Vol15, No.5, 2013, pp. 1223-1225.

[6] T. Ershal, H. Fuller et.al, "*Model based analysis and classification of driver distraction under secondary task*", IEEE Trans. on Intelligent Transport Systems, Vol.11, No.3, 2010, pp.692-701.

[7] M. Kutila, et.al, "*Driver distraction detection with a camera vision system*", IEEE International Conference on Image Processing ICIP 2007, Vol.6, 2007, pp. 201-204.

[8] M. Su, C. Hsuing and D. Huang, "*A simple approach to implementing a system for monitoring driver inattention*", IEEE International Conference on Systems, Man and Cybernetics

[9] Y. Liang et.al, "*Non-intrusive detection of driver cognitive distraction in real time using Bayesian networks*", Transportation Research Record: Journal of the Transportation Research Board (TRR) Vol. 2018, 2007, pp.1-8.

[10] Y. Liang et.al, "*Real-time detection of driver cognitive distraction using Support Vector Machines*", IEEE Transactions on Intelligent Transportation Systems, Vol 8, No.2, 2007, pp. 340-350.

[11] F. Tango and M. Botta,"*Evaluation of distraction in a driver-vehicle environment framework: An application of different data-mining techniques*", Advances in Data mining: Applications and Theoretical Aspects, Lecture Notes in Computer Science, Vol 5633, 2009, pp. 176-190.

[12] P. Angkititrakul et. al, "*Getting start with UTDrive: Driver behavior modelling and assessment of distraction for in-vehicle speech system*", Interspeech 2007, 2007, pp. 1334-1337.

Diurnal Variation of Surface Radio Refractivity over Mongolia

Jamiyan Sukhbaatar, Tsendayush Oldokh, Bold Zagd, Nyamjav Jambaljav
Department of Electronics and Communication Engineering,
School of Engineering and Applied Sciences,
National University of Mongolia
{jamiyan, tsendayush, bold, nyamjav}@seas.num.edu.mn

Abstract

The diurnal radio refractivity over Mongolia was studied. The values of radio refractivity have been determined 59 different locations. The diurnal refractivity was calculated first day of January, April, July and October. A total of more than twenty thousand refractivity measurements was considered in this analysis. The result showed that the refractivity values were lower in the morning and the night, and higher in the afternoon. This is the result of variations in meteorological parameters such as humidity, temperature and atmospheric pressure. The highest values observed in winter and the lowest values were in spring. The diurnal maximum radio refractifiy, 342 was in Khuvsgul aimag on January 1 at 12.00 and a minimum one, 292 was in Dundgovi aimag on April 1 at 21.00.

Keywords: radio refractivity, meteorological data

1. Introduction

The propagation of radio wave signal in the troposphere is affected by many processes which include the variations of meteorological parameters such as temperature, pressure and humidity. These are associated with the change in weather in different seasons of the year. These variations in meteorological parameters have resulted in refractivity changes. Multipath effects also occur as a result of large scale variations in atmospheric radio refractive index, such as different horizontal layers having different refractivity [1]. This effect occurs most often, when the same radio wave signals follow different paths thereby having different time of arrivals to its targeted point. This may result to interference of the radio wave signals with each other during propagation through the troposphere. The consequence of this large scale variation in the atmospheric refractive index is that radio waves propagating through the atmosphere become progressively curved towards the earth. Thus, the range of the radio waves is determined by the height dependence of the refractivity. Thus, the refractivity of the atmosphere will not only vary as the height changes but also affect radio signal.

The quality of radio wave signal reception and probability of the failure in radio wave propagations is largely governed by radio refractivity index gradient which is a function of meteorological parameters changing in lower atmosphere such as temperature, pressure and humidity.

Radio waves travel through vacuum with a speed equal to the speed of light. In material medium, the speed of the radio waves is approximately c/n where c is the speed of light in vacuum and n is the radio refractive index of the medium. The value of the radio refractive index (n) for dry air is almost the same for radio waves and the light waves. But the value of the radio refractive index (n) for water vapor, which is always present in some quantity in the lower troposphere, is different for the light waves and radio waves. This arises from the fact that water vapor molecule has a permanent dipole moment, which has different responses to the electric forces of different radio wave frequencies propagated within the atmosphere.

Radio–wave propagation is determined by changes in the refractive index of air in the troposphere. Changes in the value of the troposphere radio refractive index can curve the path of the propagating radio wave.

At standard atmosphere conditions near the Earth's surface, the radio refractive index is equal to approximately 1.0003 [2]. Since the value of refractive index is very close to unity, then the refractive index of

directly related with driving known as mind-off-road. Visual distraction can be automatically detected by air in the troposphere is often measured by a quantity

called the radio-refractivity N, which is related to refractive index, n as:

N= (n-1) 10^6 [3]

As the conditions of propagation in the atmosphere vary, the interference of radio-wave propagation is observed. Such interferences are incident with some meteorological parameters.

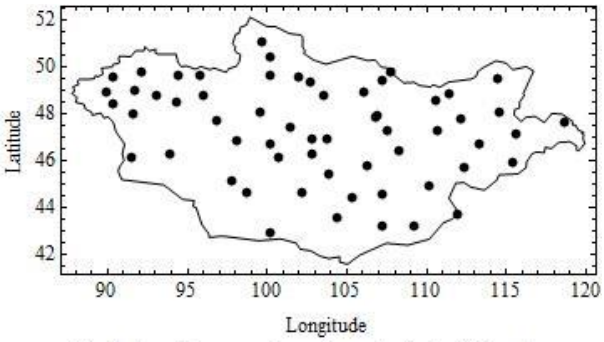
The atmospheric radio refractive index depends on air temperature, humidity, atmospheric pressure and water vapour pressure. Subsequently, meteorological parameters depend on the height at a point above the ground surface. Variation in any of these meteorological parameters can make a significant variation on radiowave propagation, because radio signals can be refracted over whole signal path [4]. In the atmosphere, pressure, temperature and humidity decrease exponentially as height h increases.

Atmosphere has an important feature [5]:- the vertical gradient of the refractive index, G. The vertical gradient of the refractive index is responsible for bending of propagation direction of the electromagnetic wave. If the value G is negative, the signal bends downward [6]. The characterization of the seasonal variation in fading and its dependence on meteorological parameters provides the way to improve transmission performance by better tailoring of performance equipment design and usage to the amount of fading expected at a given location and time of the year.

The main goal of this paper was to apply well known model [7] to find out the diurnal variation of the radio refractive index values using geographical and meteorological data of Mongolian localities in different days of different seasons.

2. Meteorological stations

In this study, we investigated the diurnal variation of refractivity over fifty nine localities. Meteorological parameters (pressure, temperature and relative humidity) used to calculate radio refractivity over Mongolia were obtained from weather stations which located fifty nine different locations. These stations give us reasonable geographic coverage. Since temperature, humidity, atmospheric pressure and water vapor pressure, which are important parameters for determination of radio refractivity are highly variable and change rapidly in time and from place to place, measurements of these parameters were considered fifty nine different locations. Meteorological data were taken in different period for different stations. Twenty seven stations cover 40 years from 1960 to 2015 and rest of stations covers up to 31 years from 1974 to 2015. The locations of the meteorological stations are shown in Figure 1.



"Fig 1. Location map for meteorological stations"

3. Calculation of Radio Refractivity

Radio refractive index, n, is equal to approximately 1.0003. Since n never exceeds unity by more than a few parts in 10^{-4}, it is convenient to consider scaled-up by 10^6 and measured by radio-refractivity N, which is related to the refractive index, *n* as:

$$N = (n-1)\times 10^6 \quad (1)$$

Radio refractivity [3] N is expressed by:

$$N = N_{dry} + N_{wet} = \frac{77.6}{T}\left(P + 4810\frac{e}{T}\right) \quad (2)$$

with the dry term, N_{dry}, of radio refractivity given by:

$$N_{dry} = 77.6\frac{P}{T} \quad (3)$$

and the wet term, N_{wet}, by:

$$N_{wet} = 3.732\times 10^5 \frac{e}{T^2} \quad (4)$$

where P is the atmospheric pressure (hPa), e is the water vapor pressure (hPa) and T is the absolute temperature (K).

The relationship between water vapor pressure e and relative humidity is given by [3]:

$$e = \frac{He_s}{100} \quad (5)$$

e_s is the saturation vapor pressure (hPa) at the temperature t (°C), and obtained from:

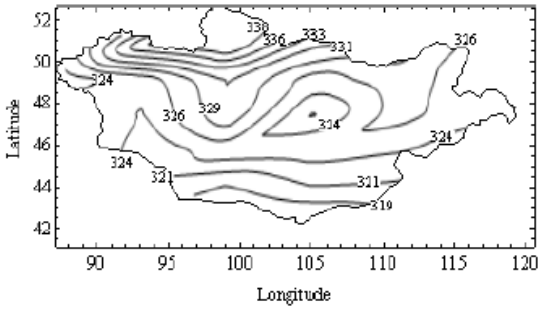
$$e_s = a\exp\left(\frac{bt}{t+c}\right) \quad (6)$$

where H is the relative humidity (%) and t is the Celsius temperature (°C). For water a=6.1121, b=17.502, c=240.97 (valid between -20° to + 50°, with an accuracy of ±20%) [3].

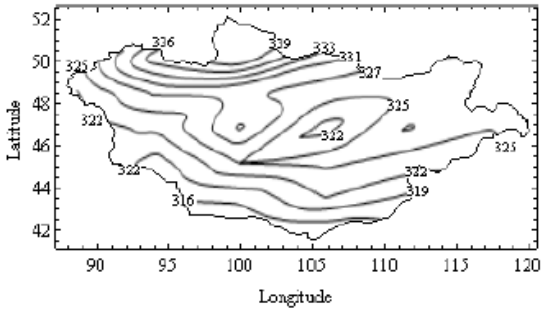
4. Results and Discussions

Mean values of the refractivity N have been determined by using (2) at the each 59 stations. The partial water vapor pressure e was determined by using (5) and (6). Up to forty four years (1960-2015) values of temperature, humidity and atmospheric pressure were taken from fifty nine meteorological stations. Each day, eight measurements of temperature, relative humidity and pressure were taken at 02.00, 05.00, 08.00, 11.00, 14.00, 17.00, 20.00 and 23.00 hours local time at all fifty nine stations. All calculations have been performed using Mathematica [8].

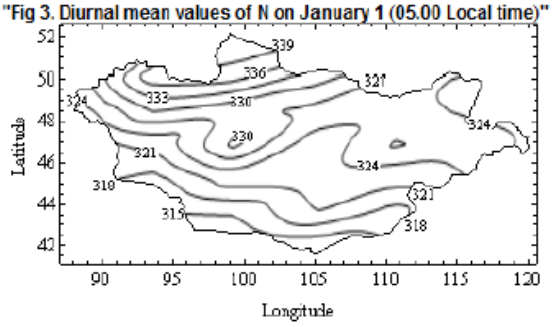
Figures from 2 to 9 present the contours of diurnal mean refractivity values on January first for different measurement hours. The maps were contoured by interpolation between the wide-spaced plotted data points using ListControurPlot command of the Mathematica. The options that used to plot contours were "Contours 9, MaxPlotPoints 7, InterpolationOrder 2". The result showed that the refractivity values were lower around 11.00, 14.00 and 17.00hr (fig 5-11). During this time, temperature and atmospheric pressure increase and humidity decreases. From the equation (2), it can be seen that when temperature increases, refractivity decreases. Sunrise around 08.40 and sunset around 17.00hr on January 1. It is clear that the temperature is higher at 11.00, 14.00 and 17.00hr than other times. It is noted that the southern areas display lower values of radio refractivity while the northern areas have higher values. The lowest value was 339 at 08.00hr in the Khuvsgul aimag. The highest value of radio refractivity was 314 at 20.00hr in the Umnuguvi aimag.



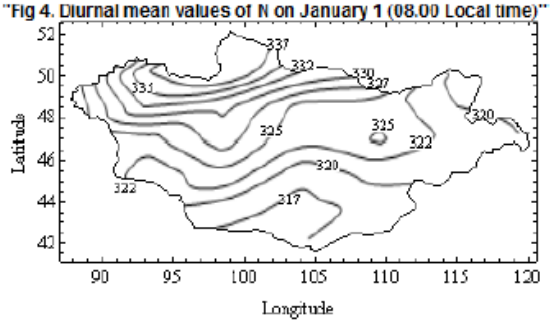
"Fig 2. Diurnal mean values of N on January 1 (02.00 Local time)"



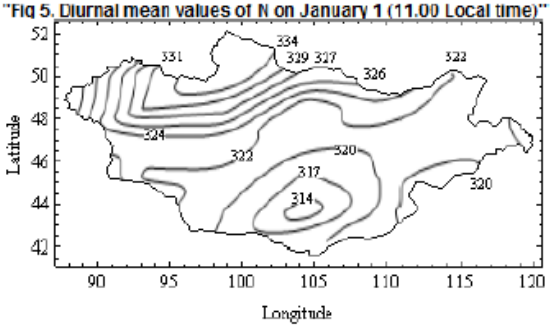
"Fig 3. Diurnal mean values of N on January 1 (05.00 Local time)"



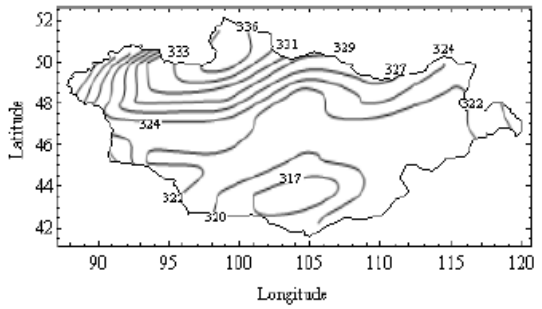
"Fig 4. Diurnal mean values of N on January 1 (08.00 Local time)"



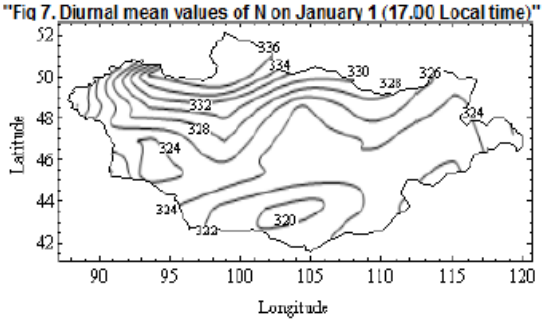
"Fig 5. Diurnal mean values of N on January 1 (11.00 Local time)"



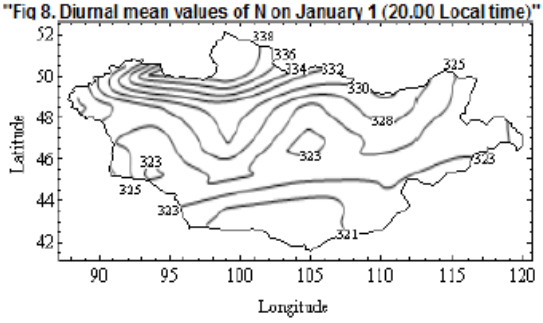
"Fig 6. Diurnal mean values of N on January 1 (14.00 Local time)"



"Fig 7. Diurnal mean values of N on January 1 (17.00 Local time)"

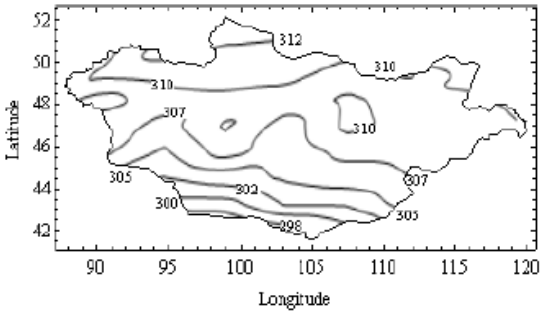


"Fig 8. Diurnal mean values of N on January 1 (20.00 Local time)"

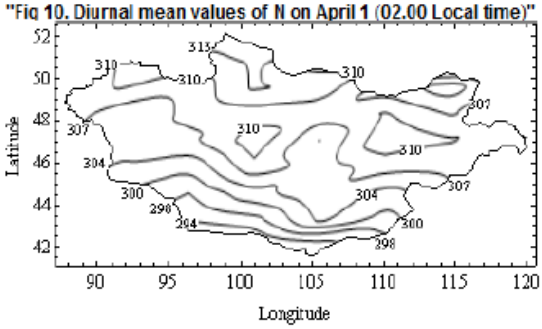


"Fig 9. Diurnal mean values of N on January 1 (23.00 Local time)"

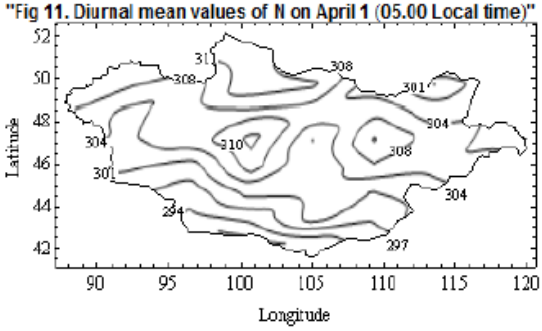
The hourly N values on April 1 presented in figures 10 through 17. The maps show that radio refractivity values were decreased all over the country from the late morning (11.00hrs, Fig 13) to evening (20.00hrs, Fig 16). It is due to that the temperature increase during these hours because sunrise around 06.30 in the morning. At about 20.00hr, radio refractivity was found to increase. It is because of sunset about at 19.20hr. Comparing with the January 1, refractivity values were lower for the whole country and since the day time is longer than first of January, N values were lower for longer periods from 11.00 to 20.00hr on April 1. Radio refractivity is found to have the peak values during night hours of the day (Fig 11 and Fig 17) and the maximum value was 314 at 23.00hr.



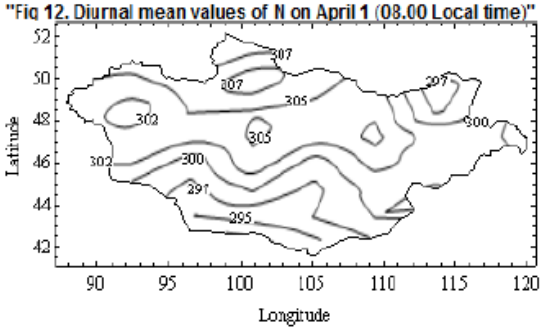
"Fig 10. Diurnal mean values of N on April 1 (02.00 Local time)"



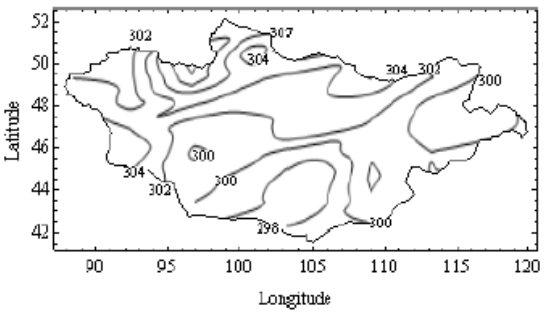
"Fig 11. Diurnal mean values of N on April 1 (05.00 Local time)"



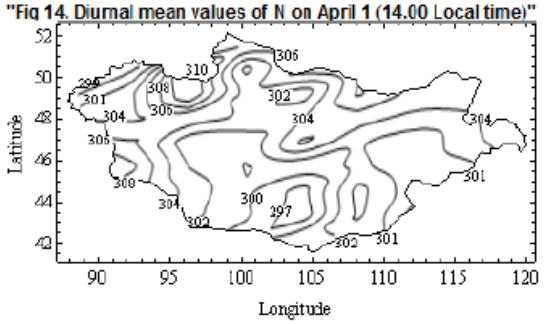
"Fig 12. Diurnal mean values of N on April 1 (08.00 Local time)"



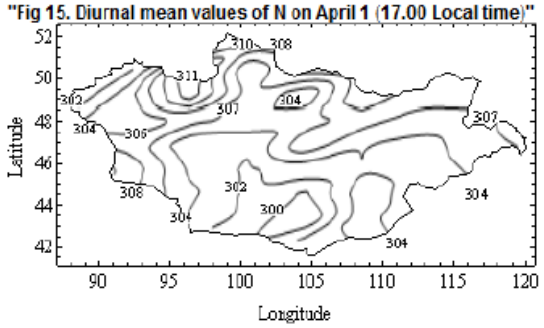
"Fig 13. Diurnal mean values of N on April 1 (11.00 Local time)"



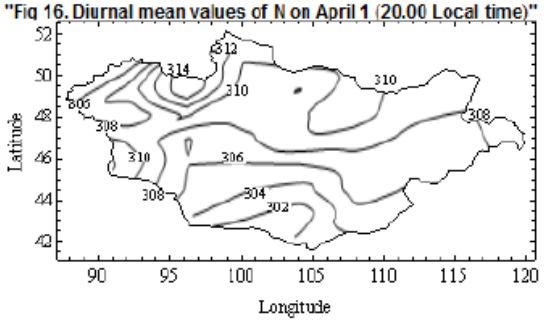
"Fig 14. Diurnal mean values of N on April 1 (14.00 Local time)"



"Fig 15. Diurnal mean values of N on April 1 (17.00 Local time)"

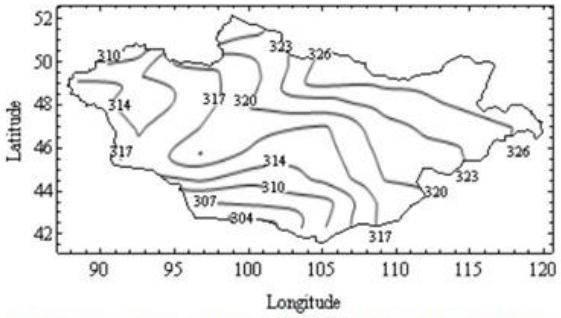


"Fig 16. Diurnal mean values of N on April 1 (20.00 Local time)"

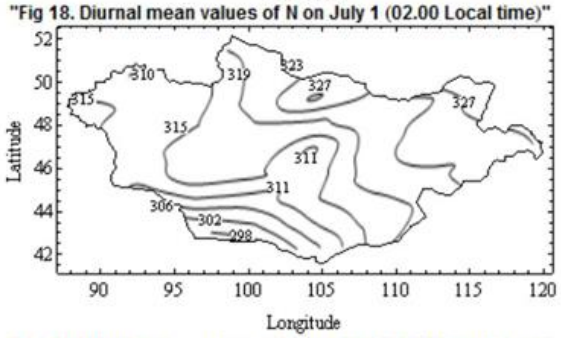


"Fig 17. Diurnal mean values of N on April 1 (23.00 Local time)"

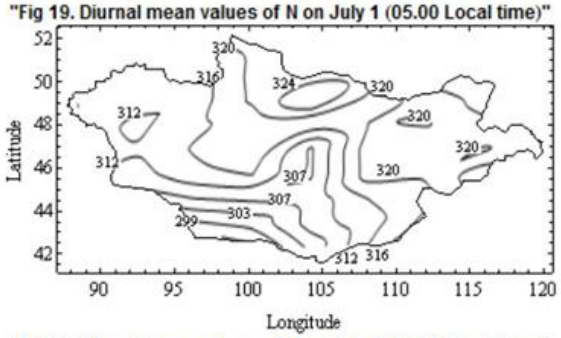
The diurnal variation of refractivity on July 1 is depicted in Figure 18 to Figure 25. The sun rises around 05:00hr and the sun sets around 21:00hr on the first of July. While Figure 18 and 19 showed a variation with high values during the night and early morning hours of the day, which drops gradually from late morning (Figure 21) until night (Figure 24) and increase again (figure 25). Mean refractivity values of contour maps were higher on July 1 than on April 1 for corresponding hours. The results also show that the variation of radio refractivity values was more stable in southern areas than northern regions for different measurement hours. The variation in the southern areas can be attributed to the influence of the wet term of refractivity which is mainly influenced by the humidity. The southern region of Mongolia is drier than northern region [9]. The highest value of N was 327 N-units around 23.00hr local time and the minimum value was 296 N-units appeared at 14.00hr in the southern region.



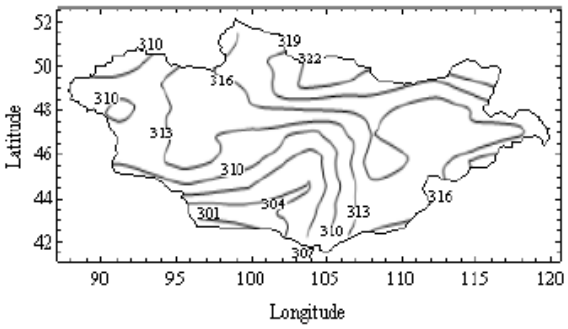
"Fig 18. Diurnal mean values of N on July 1 (02.00 Local time)"



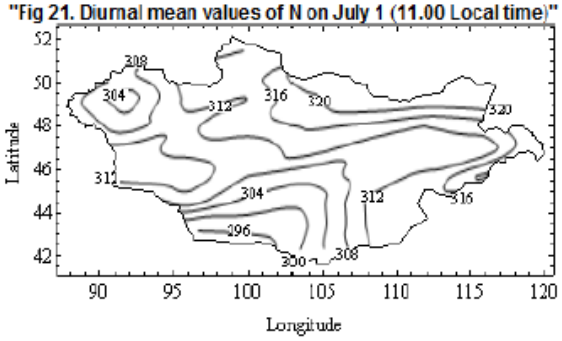
"Fig 19. Diurnal mean values of N on July 1 (05.00 Local time)"



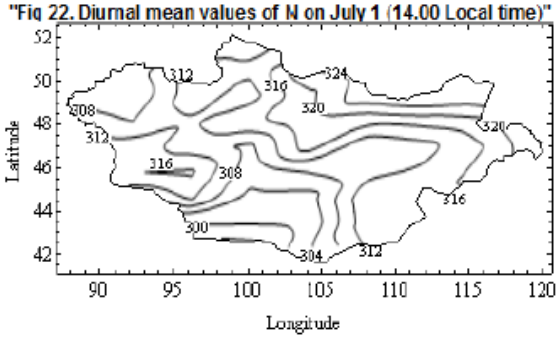
"Fig 20. Diurnal mean values of N on July 1 (08.00 Local time)"



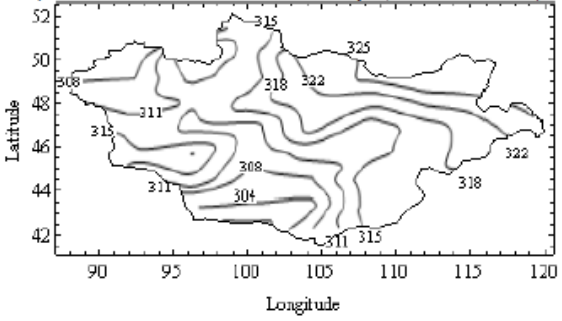
"Fig 21. Diurnal mean values of N on July 1 (11.00 Local time)"



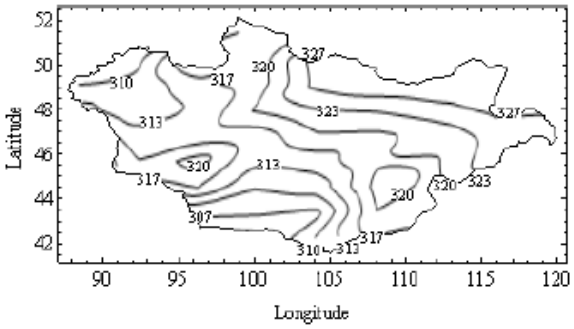
"Fig 22. Diurnal mean values of N on July 1 (14.00 Local time)"



"Fig 23. Diurnal mean values of N on July 1 (17.00 Local time)"

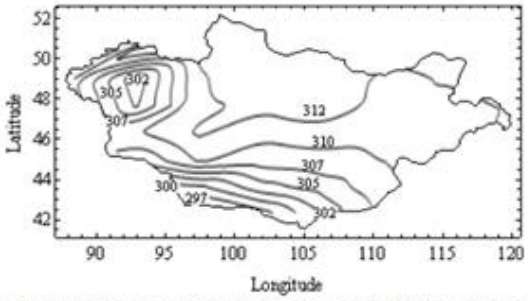


"Fig 24. Diurnal mean values of N on July 1 (20.00 Local time)"

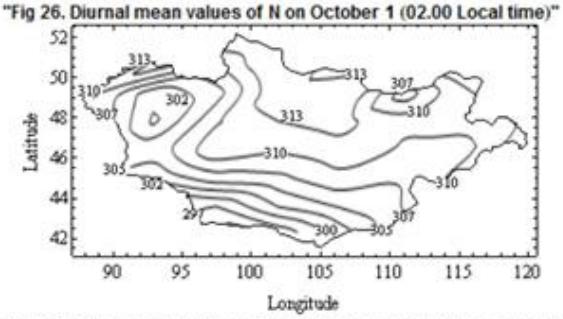


"Fig 25. Diurnal mean values of N on July 1 (23.00 Local time)"

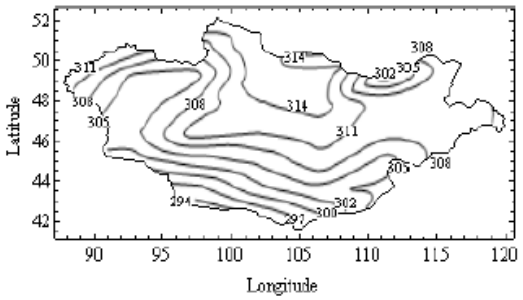
The contour maps in Figure 26 to Figure 33 present the mean refractivity values for Mongolia on October first. Changes in the refractivity values on October 1 were similar to April 1. The variation of N values of this day of the month was lower than January 1 and July 1. The refractivity peaked to about 314 N-units at 20.00hr and 23.00hr. The minimum value was 294 N-units at 08.00 hr. Sun up time on October 1 is 07.00hr and sunset around 18.30hr. The figures also show that the difference between the N values in the southern region and in the northern region is smaller on October 1 for all measurement hours.



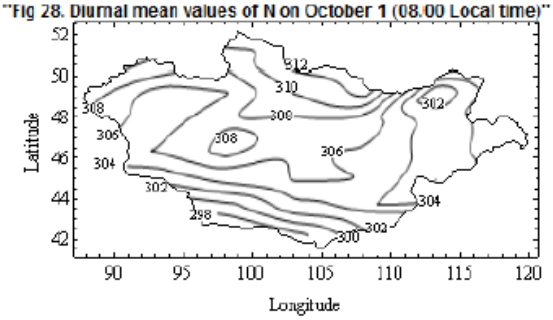
"Fig 26. Diurnal mean values of N on October 1 (02.00 Local time)"



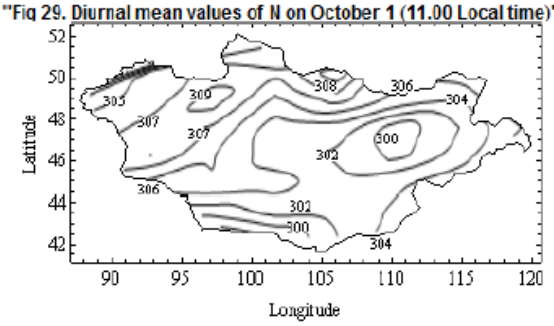
"Fig 27. Diurnal mean values of N on October 1 (05.00 Local time)"



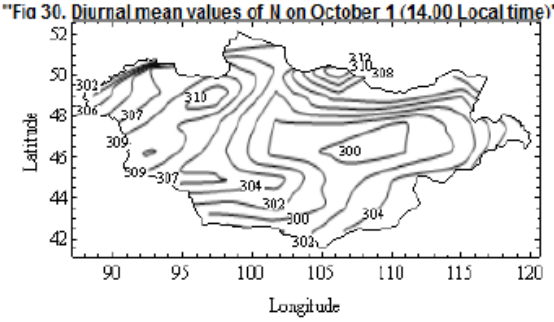
"Fig 28. Diurnal mean values of N on October 1 (08.00 Local time)"



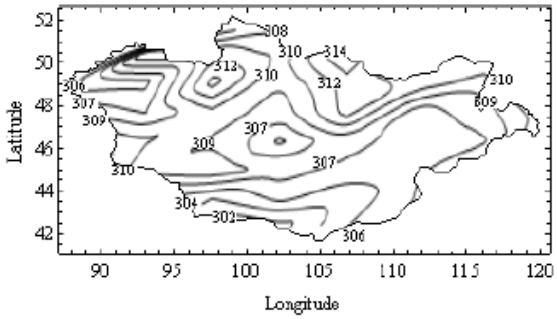
"Fig 29. Diurnal mean values of N on October 1 (11.00 Local time)"

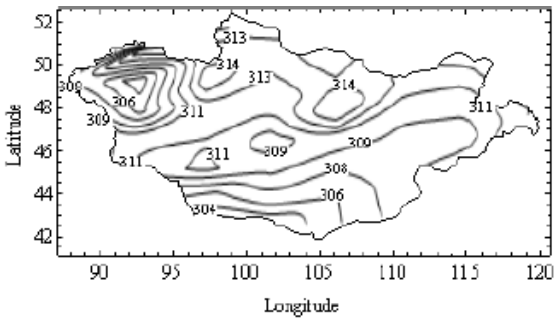


"Fig 30. Diurnal mean values of N on October 1 (14.00 Local time)"



"Fig 31. Diurnal mean values of N on October 1 (17.00 Local time)"





"Fig 33. Diurnal mean values of N on October 1 (23.00 Local time)"

5.Conclusion

a. Mean refractivity values were increased after sunrise and lower values were in the day time.
b. Higher mean refractivity values were on January 1 and lower values were on April 1.
c. Refractivity values were higher in the northern part than southern region for all maps in all days and hours.
d. Changes in the refractivity values on October were similar to April 1.
e. Maximum refractifiy, 342 was in Khuvsgul aimag on January 1 at 12.00 and a minimum one, 292 was in Dundgovi aimag on April 1 at

6.References

[1] Grabner, M., and Kvicera, V., "Radio Engineering 12", No.4, 50, 2008.

[2] Freeman R.L., "Radio System Design for Telecommunications", John Wiley&Sons Inc, Hoboken, New Jersey, 2007, pp. 880.

[3] ITU-R "The radio refractive index: Its

[4] Priestley J.T., and Hill R.J., "Measuring High-Frequency Refractive Index in the Surface Layer", Journal of Atmorpheric and Oceanic Technology, Vol.2, 1985, pp. 233-251

[5] Willoughby A.A., and Aro T.O., and Owalabi I.E., "*Seasonal variations of radio refractivity gradients in Nigeria*", Journal of Atmospheric and Solar Terrestrial Physics, Vol.64, 2002, pp. 417-425.

[6] Guo G., and Li, S., "*Study on the vertical profile of refractive index in the troposphere*", International Journal of Infrared and Millimeter Waves, Vol.21, No.7, 2000, pp. 1103-1112.

[7] Bean B.R., "*The Radio Refractive Index of Air*", Proc, I.R.E., 50, pp.260-73, March 1962.

[8] Wolfram S., "*The Mathematica Book*", Fifth Edition, Wolfram Editions, 2003.

[9] "*Mongolia's Country Studies Report on Climate Change*", Vol.1: Executive Summary. Ulaanbaatar: HMRI

Real-time Document Ranking using Term Weight Estimation in Information Retrieval

Erdenebileg Batbaatar[1], Aziz Nasridinov[2], Oyun-Erdene Namsrai[3], Keun Ho Ryu[*]
[1,2,*)]*Database and Bioinformatics Laboratory, School of Electrical and Computer Engineering, Chungbuk National University, South Korea*
[3]*School of Engineering and Applied Science, National University of Mongolia*
*{[1]eegii, *khryu}@dblab.chungbuk.ac.kr, [2]aziz@chungbuk.ac.kr, [3]oyunerdene@seas.num.edu.mn*

Abstract

In this paper, we propose a method term-weight based document ranking approach that exploits relevance information using search terms. A fundamental goal of search engine is to identify documents that have relevant text. But it is more useful to measure how important a word is to a document in a corpus. Our system can automatically give us effectiveness result that retrieve relevant information from already ranked documents by term weight. It enables the user to issue a flexible query and receive the results immediately even when the number of the documents that match the query is very large. Through an experiment on the corpus which contains sentence from the abstract and introduction of 30 scientific article that have been annotated according to a modified version of the Argumentative Zones annotation scheme.

Keywords: *Information retrieval; Document Ranking; Term weighting; Tokenization; Stemming; Removing Stopword;*

1. Introduction

The web creates new challenge for information retrieval. The amount of information on the web is growing rapidly, as well as the number of new users inexperienced in the art of web research. Online search engines use index database which map keywords to documents or, in a more general term, to information resources.

Most search engines for information retrieval is not consider ranked documents, it's more effective to evaluate their relative importance by considering their distribution in the full document collection. The intuition behind this approach is that more frequent a term is in a collection, the less discriminating it is. The most classical embodiment of this approach is the family of tf x idf scores where tf stands for the frequency of a term in a document, and idf for the "inverse document frequency". The user might wish to weight the importance of search terms. In information retrieval system has following kind of problems.

(1) To process large document collection quickly. The amount of online data has grown at least as quickly as the speed of computers

(2) To allow more flexible matching operations

(3) To allow ranked retrieval. In many cases the user wants the best answer to an information need among many documents that contain certain words

Our proposed method is to improve efficiency of information retrieval system by ranking document.

Assume that the user gives a standard query to our engine, and our engine gives back a ranked list of results automatically.

The article organized as follows: Section 2 describes some related works whereas Section 3 explain the methods used in our work and presents our key idea which is unique. Section 4 reports the experiments on the dataset we have selected. Finally Section 5 derives the conclusions.

2. Related works

In this Section we briefly summarize related works, there are a number of web-based text mining applications which can be used for information retrieval.

Text-based information retrieval is one of the oldest areas of research in Computer Science, and a number of approaches have been devised over the years [1, 2]. Methods based on 'bag-of-word' representations, where the frequency of terms in documents are used to

define a vector space remain dominant, with variations of the Term Frequency-Inverse Document Frequency model (TF-IDF) being the most popular (e.g. the Pivot TF-IDF technique[1]). Boolean and Bayesian approaches have also been well-studied and find use in practice.

An important data structure for supporting text search is the inverted index. In an inverted index, words or other tokens are mapped to documents that contain them. EBIMed (Rebholz-Schuhmann et al., 2007) receives a PubMed-style query from the user and analyzes the matched documents to recognize protein/gene names, GO annotations, drugs and species mentioned. Frequently occurring concepts are shown in a table, and the user can view the sentences corresponding to the associations. PolySearch (Cheng et al., 2008) can produce a list of concepts which are relevant to the user's query by analyzing multiple information sources including PubMed, OMIM, DrugBank and Swiss-Prot. It covers many types of biomedical concepts including diseases, genes/proteins, drugs, metabolites, SNPs, pathways and tissues. Systems that provide similar functionality include XplorMed (Perez-Iratxeta et al.,2003), MedlineR (Lin et ak., 2004). LitMiner (Maier et al., 2005) and Anii (Jelier et al ., 2008).

Although these applications are useful in exploring such information in the document, not many of them provide real-time responses. The users often have to wait for several minutes (or even hours) before they receive the results.

A widely used scalable full text inverted index library is the Lucene Java library, and fast, featureful full-text indexing and searching library implemented in pure Python. We adapt a TF-IDF-based scoring metric provided with Lucene and Whoosh.

3. Proposed method

Our engine receives a query from the user as the input which has two kind of input text. The one is a list of stopwords, the other is a list of documents. We can manually manage the lists and store any changes in database. For experiment, we used 900 articles which contains abstract, introduction article id and article domain.

Pseudo code of DocuRank

Input:
(1) *S* : a set of stopwords
(2) *D* : a set of documents
Output:
(1) *R* : the list of ranked documents
Algorithm:
1. Initialize the term list T = {}, ranked document R = {}
2. Tokenization
3. Removing stopwords
4. Stemming with Porter stemming algorithm
5. Weighting term
6. **FOR ALL** d ∈ D **DO**
7. **FOR ALL** s ∈ S **DO**
8. Weight for each term in each document, append to T
9. Search the documents by user given query in T, append to R
10. **RETURN** R

Figure 1. Pseudo code of DocuRank

Step 1: Tokenization. In most information retrieval system, basically important step is to select terms from unstructured sentence. Basically each token is a word, although the definition of a word is not straightforward. We selected all individual words one by one which contains a term and document id. In our experiment we selected 152,341 words from the all documents. For example: "A computer is a general purpose device that can be programmed to carry out a set of arithmetic or logical operations automatically". Which can be extracted like this: "a", "computer", "is", "a", "general", "purpose", "device", "that", "can", "be", "programmed", "to", "carry", "out", "a", "set", "of", "arithmetic", "or", "logical", "operations", "automatically".

Step 2: Stop Words Removal. There are many common words, they do not carry any specific information. These words occur in all texts with approximately the same frequency, and do not relate to the content of text. They could disturbed the similarity calculation, most of them are prepositions, conjunctions or pronouns. We make a possibility to the users to manage the stopwords freely. The users can append new stopword, delete a stopword on the stopwords section. After removing stopwords similarity calculation result would be better than before.

Step 3: Stemming. Stemming aims at identifying the ground from of each word. There are many stemming algorithms are available. We have chosen the algorithm Porter stemming which is most commonly used stemmer without a doubt, also one of the most gentle stemmers.

Step 4: The DocuRank. Since the number of the concepts contained in the document is usually very large, it is important that the concepts are properly ranked when presented to the user. The idea is that weight all terms in each document for retrieving ranked documents. More simply, we ranked he documents by weight of each term. It helps to find

corresponding documents. Various methods for weighting terms have been developed in the field.

In the case of the term frequency *tf(t,d)*, the simplest choice is to use the raw frequency of a term in a document, i.e. the number of times that term t occurs in document d.

$$\mathrm{tf}(t,d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}} \quad (1)$$

The inverse document frequency is a measure of how much information the word provides, that is, whether the term is common or rare across all documents. It is the logarithmically scaled inverse fraction of the documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|} \quad (2)$$

4. Experimental results

In experiments, we used the corpus which contains sentence from the abstract and introduction of 30 scientific articles that have been annotated according to a modified version of the Argumentative Zones annotation scheme. These scientific articles come from three different domains such as PLoS Computational Biology (PLOS), the machine learning repository on arXiv (ARXIV), the psychology journal Judgment and Decision Making (JDM).

Figure 2 shows the recall on the 900 articles. The precision and recall were calculated with different number of articles. Figure 3 shows the precision on the 900 articles. The precision and recall were calculated with different number of articles.

As a result, we can see that if number of articles or data is less than 400, our DocuRank approach shows better result than other 2 libraries. And while increasing number of articles the precision and the recall was proportionally decreased.

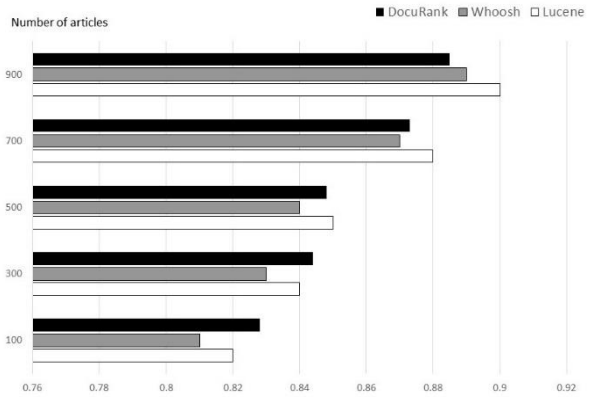


Figure 2. Recall on unlabeled dataset

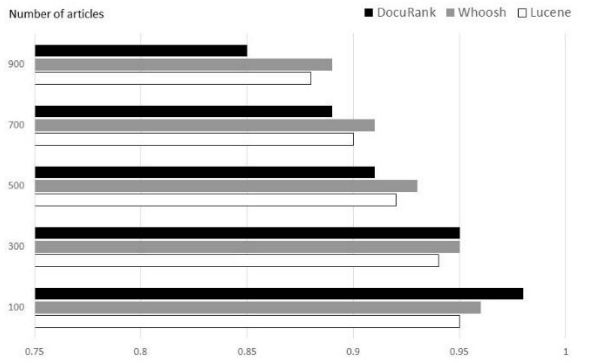


Figure 3. Precision on unlabeled dataset

5. Conclusion

In this paper, we have presented a novel weight-based ranking approach that exploits relevance information using search terms. This technique is presented as a natural extension of weighting methods using information about the distribution of index terms in documents in general. Our experimental results showed that the proposed weight-based indexing approach is more effective than classical keyword-based indexing ones. Moreover, in case of a few number of testing data our approach performs better than Lucene and Whoosh. In future works, we plan first to improve what extend the weighting factor depends or not on the used document collection, and when having huge number of data to improve result of approach.

Acknowledgment

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2015-H8501-15-1013)

supervised by the IITP(Institute for Information & communication Technology Promotion) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No.2013R1A2A2A01068923).

6. References

[1] MCGRAYNE, Sharon Bertsch, "*The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy*", Yale University Press, 2011.

[2] SASARAK, Christopher, et al, "min: A multimodal web interface for math search", In: *Symp. Human-Computer Interaction and Information Retrieval, Cambridge, MA*. 2012.

[3] CHENG, Dean, et al. PolySearch, "a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites.", *Nucleic acids research*, 2008, 36.suppl 2: W399-W405.

[4] REBHOLZ-SCHUHMANN, Dietrich, et al. "EBIMed—text crunching to gather facts for proteins from Medline", *Bioinformatics*, 2007, 23.2: e237-e244.

[5] PINKERTON, Brian, "Finding what people want: Experiences with the WebCrawler", In: *Proceedings of the Second International World Wide Web Conference*. 1994. p. 17-20.

[6] LIN, Simon M., et al. MedlineR, "an open source library in R for Medline literature data mining", *Bioinformatics*, 2004, 20.18: 3659-3661.

[7] JELIER, Rob, et al. Anni 2.0, "a multipurpose text-mining tool for the life sciences", *Genome Biol*, 2008, 9.6: R96.

[8] Center for Machine Learning and Intelligent Systems, https://archive.ics.uci.edu/ml/datasets/Sentence+Classification

[9] Apache Lucene http://lucene.apache.org/

[10] Whoosh Python https://pypi.python.org/pypi/Whoosh/

Mining Association Rules from Educational Data to Improve Teaching and Learning Outcomes

Chunyan Ji, Clement Leung, Junru Zhong
Computer Science and Technology Programme
United International College
P.R. China
{chunyanji, clementleung}@uic.edu.hk, j430003045@mail.uic.edu.hk

Abstract

Data mining plays an important role in many fields, and with most governments allocating substantial resources to education, vast amounts of educational big data is becoming available. Thus mining educational data mining is increasingly important and is pivotal in achieving improved educational outcomes. In this paper, we mine association rules from actual educational data collected. After mining the association rules between majors and students' performances, teachers and students performances, class time and student' performances, we extract knowledge about which factors can affect students' performances. Based on the mining results, we provide suggestions to school administration units to help arrange teaching assignment optimally for team taught courses. Instead of random teaching assignment, the best match teachers and students can work together to reach higher teaching and learning quality. From such results, it is expected the same approach can be used to improve learning outcomes in other educational courses and settings.

Keywords: Educational Data Mining, Association Rules, Teaching Assignment

1. Introduction

The application of data mining and knowledge discovery technologies has proven to be successful in many kinds of areas such as marketing, businesses, entertainment, etc. Although educational data mining is a recent research field, it has developed rapidly. Many researchers have found that educational data mining is a promising field of research. Using data mining techniques, we can discover veiled knowledge from data that is related to a course. The knowledge can help teachers design and manage their course in a more efficient way to improve teaching and learning quality. It can also help administrative unit to build more intelligent management systems. Researchers have done many works on educational data mining. From mining students or course related data, researchers can analyze students' learning behavior, students' performances; evaluate teachers' performances, etc. [1] gave a case study that used data mining to identify the behavior of failing students, and then teachers can warn students if they have the similar behavior. In [2], the author used data mining to predict students' final grade. [3] used data mining to help build the teaching evaluation system. Data mining technology is also used in the educational administration area. In [4][5], the author used data mining to improve educational administration management system. In [6], student performance is viewed as a classification task and the decision tree method is used. Unlike [6], we do not apply decision tree methodology but instead, we focus on association rules. In this study, we use data mining technology to help the administration unit to assign teaching loads to achieve the best learning outcomes. In data mining, the data can be collected from the teachers, for example, the students' course work data. With rapid development of computer technology and the Internet, more and more educational institutes use e-learning system, a vast amount of data can also be collected from the e-learning system. In this study, we collected students' course work data from the Information Technology course offered by United International College, China. The purpose is to seek a better strategy to manage the team taught course in a more efficient way. There are 15 IT classes each semester. Over ten teachers teach the IT courses. Currently, the content for all the students are the same. In the future, the content may be differentiated by students' background, majors, and interests. In both cases, to assign a correct teacher to each class is an urgent task to optimize the

teaching assignment method, hence to improve teaching and learning outcomes. We need to dig out some evidence on how to arrange the teaching assignments smartly instead of the current random teaching assignment.

2. Association Rules mining

Association rules mining is one of the most commonly used techniques in data mining. It was initially used for market basket analysis. It allows finding rules of the form "If A then C where A is the antecedent and C is the consequent, A and C are item set". There are two steps in association rules mining. The first step is to find the item set, and the second step is to produce the association rules from the item set. We use support to define the probability of the item set actually occur. When more than a predefined minimum support occurs, it means the rule occurs frequently. Confidence is the probability of the consequent occurs when the antecedent is true. The association rules can be considered to be strong if the confidence is high.

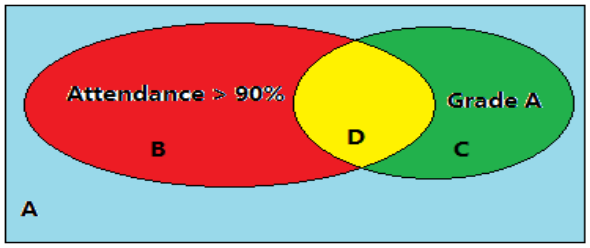


Figure 1. An example of association rule

Figure 1 shows an example of an association rule for student data in a class. Set A represents the total number of students in this class, set B represents the students who attended more than 90% of class time, set C represents the students who received grade A, and set D means students who have good attendance and received grade A. In this example, our dataset are students' attendance records and students' final grades. We want to mine the relationship between students' attendance and their final grades. In the association rule, the antecedent is "If attendance is over 90% of the class time", which means students' attendance is good. The consequent is "grade = A". The support is equal to *count(D)/count(A)*, while the confidence is equal to *count(D)/count(B)*. If the support meets the minimum value and the confidence is high enough, it means students who attend the class frequently will more likely receive high grades, so students should attend classes.

In this study, we look for factors that affect students' final grades in a team taught course, which means a course that has many sessions and different sessions are taught by different teachers. The dataset is student's course work, teacher's information, class time setting in the recent two years. We can mine the following association rules from this dataset:

- Teacher → Grade. Different teachers have different teaching styles. Even if the teaching content is the same, the students' performances can be different.
- Students' background → Grade. Students' background information includes students' date of birth, students' sex, students' home town, etc. These attributes can all be an individual antecedence in an association rule to mine how they affect students' performances in this class.
- Major → Grade. Students' major could have strong association with their grades.
- Students' learning behavior → Grade. Students' own learning behavior can affect their find grades. For example, class attendance, how often students use E-learning system, and the frequency of participation in group discussions, etc.
- Class time setting → Grade. College students tend to sleep late at night. Class time could affect students' performances also.

In this study, we focus on three association rules to find out knowledge to improve teaching and learning outcomes. They are:

Teacher → Grade
Major → Grade
Class time → Grade

3. Analysis of Educational Data

3.1 Data Collection

In this study, the data we collected is from the Information Technology course, a team taught general education course offered to all freshmen. The data includes 4 semesters over 2,500 students' grades, teachers' information, and class time for each class. There are 15 IT sections each semester and the IT team has more than 10 instructors. The current teaching assignment is arranged randomly by the academic registry. There is no method used in this random

process, which means it doesn't care about which teacher matches which major students better, or what class time is better for students to learn. In this study, we found out the following knowledge:

- Although all IT classes teach the same content, different teachers with different teaching styles lead to different students' performances.
- Some teachers are more suitable to teach certain major students.
- Students' major is the most important factor related to their performances.
- Class time also affects students' performances.

3.2 The Association Rule Teacher → Grade

In our association rule, an example of the item is that *If Teacher is Teacher1*, and the consequent has one item that is student *Grade = x* where x can be Grade A, B, C, D, F.

The support of this association rule is equal to

$$\frac{\text{Number of students who received Grade x in Teacher1's cl}}{\text{Total number of the students who took IT course in that sem}}$$

.

Figure 2 shows a sample of the support of Teacher → Grade rule for 2015 spring semester.

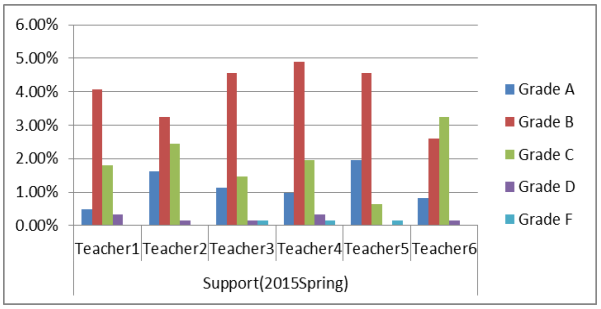


Figure 2. Support of rule Teacher → Grade

We calculated six teachers' support. The blue bars show the support of the teachers associated with grade A, which means the probability of Teacher1' class has grade A students. According to the blue bars, we noticed that Teacher2 and Teacher5 have more top students. Figure 3 shows the confidence of the Teacher → Grade rule. We can see both of Teacher2 and Teacher5 exceed 20% of the confidence. It can be said that these two teachers are good at educating top students.

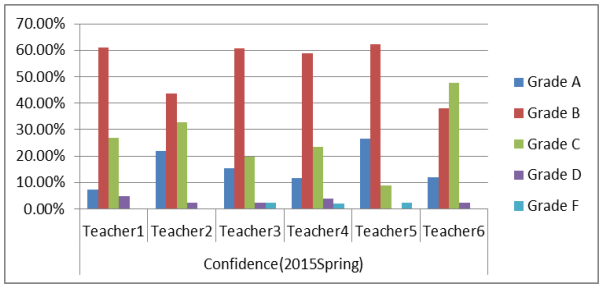


Figure 3. Confidence of rule Teacher Grade

To get an overall picture of teachers' performances, Figure 4 shows the support of the Teacher → Grade rule for Grade A in the latest two years. From the line chart, we can see that Teacher2 and Teacher5 keep the higher support over the years with one exception of Teacher 5 in semester 2.

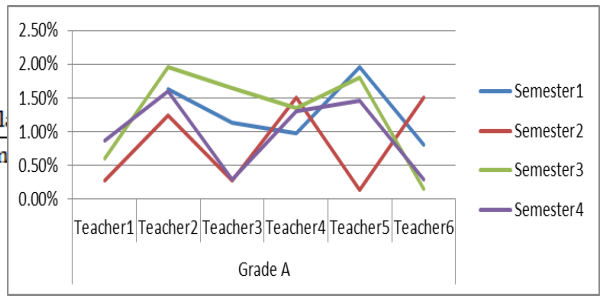


Figure 4. Support of rule Teacher → Grade in four semesters

Looking into the data, we believe it may be related to the students. In semester 2, Teacher5 taught the SWSA (Social Work and Social Administration) major. It also shows that Teacher1 and Teacher6 keep having the lowest support of grade A. One exception is that Teacher6 has highest support of Grade A in semester2. Teacher6 taught the TESL (Teaching English as Second Language) major. From our average score analysis for all majors (see Figure 5), we can see SWSA has weak students while TESL has the best performance these years. We also found out that the major is the most important factor related to the students' performances. Some majors stay at the top all these years while some majors stay at the bottom regardless of which teachers taught them.

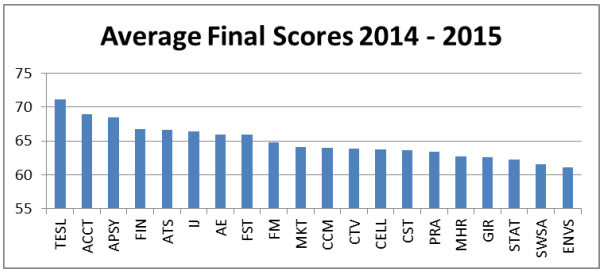


Figure 5. Average scores for all majors 2014-2015

3.3 The Association Rule Major → Grade

After mining the association rule of "If Major is m, then students grades is x" where m is the major name, and x is the Grade A, B, C, D, and F, we can see different major students perform differently even if some of them were taught by the same teachers (see Figure 6).

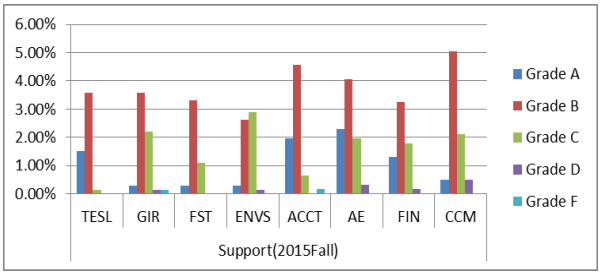


Figure 6. Sample support of rule Major → Grade in 2015 Fall

3.4 The Association Rule Class time → Grade

We also mined the association rule of Class time → Grade. Classes were arranged in four time slots: Early Morning, Late Morning, Early Afternoon, and Late Afternoon. From Figure 7, it is very obvious that early afternoon classes produce the best students' performances. The corresponding confidence we found out is over 30%. That means the best time for students to learn is early afternoon. We started to have evening classes since last semester. The support we found out for evening class is much lower than all other classes. It is suggested that evening classes should be avoided.

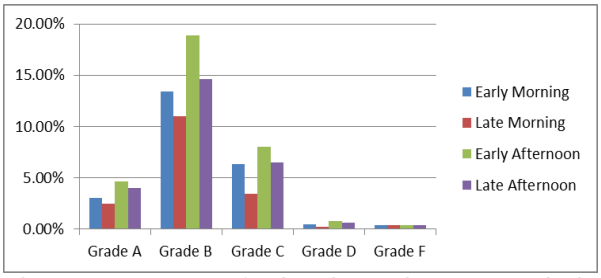


Figure 7. Support of rule Class Time → Grade in four semesters

3.5 Observations and improvement of learning outcomes

Table 1 is a sample of association rules with their support and confidence measures discovered from data for students who received grade A.

Table 1. Association Rules for Grade A students in 2015 Fall

No.	Antecedent	Support	Confidence
1	Teacher=Teacher5	1.95%	26.67%
2	Teacher=Teacher2	1.63%	21.74%
3	Teacher=Teacher1	0.81%	11.90%%
4	Teacher=Teacher6	0.40%	7.32%
5	ClassTime=EarlyAfternoon	4.20%	28.26%
6	ClassTime=EarlyMorning	2.40%	16.30%
7	Major=Accounting	1.95%	40.22%
8	Major=GIR	0.28%	1.85%

The higher support and higher confidence means there is a high correlation between the antecedent and the consequent. Based on the data mining results above, we suggest that academic registry should consider the following rules when assigning teachers to the team taught IT course.

- Assign Teacher2 and Teacher5 to weak majors.
- Assign Teacher1 and Teacher6 to good majors.
- Use early afternoon as the class time as much as possible.
- Arrange early afternoon class time for weak majors.
- Avoid evening classes.

4. Conclusions

This case study in educational data mining showed how data mining can be used to in teaching assignment to help improve teaching and learning outcomes. By collecting the students' grades over the years, we mined how teacher affect students' performances, how major is related to the students' performances, and how class time can affect the students' performances. Unveiling the knowledge can help academic registry

arrange teaching assignment in a smart way. A team course has many teachers and students involved. Teachers have their own teaching styles even if the course content is the same. Certain teachers match certain majors well, so they should be arranged to teach these students. The best teachers and the best class time should be arranged for weak majors. Teachers should pay more attention to the weak major students such as providing more tutorial time and exercises. In the future, we will use other data mining rules to mine the students' data to find more information on how to improve teaching and learning for the IT course.

Acknowledgements

The authors would like to thank International United International College Internal Research Grant for the support to make this project possible.

References

[1] Merceron, A. and Yacef, K., "Educational Data Mining: A Case Study" In Proceedings of the 12th International Conference on Artifitial Intelligence in Education AIED 2005, , IOS Press, Amsterdam, The Netherlands, 2005.

[2] Minaei-Bidgoli B., Kashy, D. Kortemeyer G., Punch W., "Predicting Student Performance: An Application of Data Mining Methods with an Educational Web-Based System". In the Proceeding of 33rd ASEE/IEEE conference of Frontiers in Education 2003.

[3] Zheng Zhang, Xuegang Hu, Yaping Zhang, Data Mining in Teaching Evaluation System in Chinese Universities (in Chinese), Computer Development and Application, 2007, 20(2).42-43.

[4] Fang Deng, Rui Shen, the 2nd Hubei Normal College journal, 2009, 26(8).99-102

[5] Yan Wang, Data Mining in Educational Administration Management, Journal of Tianjin College of Managers (in Chinese), 2014(3):69-70

[6] B. Baradwaj, S. Pal, Mining Educational Data to Analyze Students' Performance, International Journal of Advanced Computer Science and Applications (IJACSA) Vol. 2, No. 6, 2011, pp 63-69.

An Image Retrieval Framework based on Knowledge Ontology

C.H.C Leung[1] and Yuanxi Li[2]
[1]*Department of Computer Science*
United International College (UIC), China
clementleung@uic.edu.hk
[2]*Department of Computer Science*
Hong Kong Baptist University, Hong Kong
csyxli@comp.hkbu.edu.hk

Abstract

We study several semantic concept-based query expansion and re-ranking scheme and compare different ontology-based expansion methods in image search and retrieval. To improve the query expansion efficiency and accuracy, we employ the CYC knowledge base to generate the expansion candidate concepts, while filter and rank the expansion results by calculating concept similarities using the Semantic Relatedness Metrics. Using our knowledge-based query expansion in image retrieval, the efficiency and accuracy has been improved.

Keywords: image retrieval, knowledge ontology, query expansion

1. Introduction

The presence of particular objects in an image often implies the presence of other objects [8]. If term U → V, and if only U is indexed, then searching for V will not return the image in the result, even though V is present in the image. The application of such inferences will allow the index elements Ti of an image to be automatically expanded according to some probability which will be related to the underlying ontology of the application.

In this paper, we mainly focus on CYC Knowledge Base [10, 11] to generate the candidates of query expansion. The CYC knowledge base (KB) is a formalized representation of a vast quantity of fundamental human knowledge: facts, rules of thumb, and heuristics for reasoning about the objects and events of everyday life.

As for redefining image indexing [8], the most popular way is to simplify the semantic knowledge into the semantic similarity between concepts, WordNet [9], is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. It is one of these applications of semantic lexicon for the English language and is a general knowledge base and commonsense reasoning engine. The purpose of the work is both to produce a combination dictionary-and-thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications.

In this paper, we propose a query expansion for image retrieval system using CYC knowledge base as expansion candidate generator, then employ WordNet as the refining tool to filter and re-rank and expanded queries. With the proposed approach, the indexing accuracy and efficiency has been improved. In section 2, related work has been introduced; section 3 mainly describe how the system work; the experimental results are given in section 4 with the conclusion is drawn in the last section.

2. Related Works

To find the proper expansion candidates, it is required to measure the relatedness between the original query and the candidate queries. There are three types of relatedness measurement:

2.1 Topological similarity

Four types of approaches are used to calculate topological similarity between ontological concepts or instances:

Edge-based similarity: Edge-based similarity measures are based mainly on counting the number of edges in the graph to get the path between two terms [1, 2, 3];

For Example, Pekar et al [6] proposed an approach to measure the taxonomic similarity between a and b by

$$T(a,b)=\frac{\delta(root,c)}{\delta(a,c)+\delta(b,c)+\delta(root,c)} \quad (1)$$

where δ(a,b) describes the number of edges on the shortest path between a and b.

Node-based similarity: in which the main data sources are the nodes and their properties.

For example, Resnik [4] proposed an approach to measure the concept similarity based on the notion of information content. The information content of a concept (term or word) is the logarithm of the probability of finding the concept in a given corpus.

$$sim(w_1,w_2)=\max_{c_1\in sen(w_1)\ c_2\in sen(w_2)}[sim(c_1,c_2)] \quad (2)$$

Pairwise similarity: Combining the semantic similarities of the concepts they represent, measure functional similarity between two instances. For examples, the Pairwise Document Similarity considers symmetric similarity measures defined as follows:

$$sim(d_i,d_j)=\sum_{t\in V} w_{t,d_i}\cdot w_{t,d_j} \quad (3)$$

Where sim(di, dj)is the similarity between documents di and dj and V is the vocabulary set.

Groupwise similarity: Compare with Pairwise Similarity, Groupwise Similarity does not combining the semantic similarities of the concepts they represent, it calculates the similarity directly. For example, Jaccard coefficient [7] measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|} \quad (4)$$

2.2 Statistical similarity

This relates to the expectation that certain semantic objects tend to occur together. The relevant weighting is expressed as a conditional probability given the presence of other objects. An expansion to associate an image object Oj given the presence of object Oi is taken to be indexable when

$$Prob\,[Oj\,|Oi]\geqslant h' \quad (5)$$

Many approaches are used to measure the statistical similarity of entities, like GLSA (Generalized Latent Semantic Analysis), NGD (Normalized Google distance), PMI (Pointwise mutual information) and etc.

For Example, The PMI (Pointwise mutual information) [5] PMI of a pair of outcomes x and y belonging to discrete random variables X and Y quantifies the discrepancy between the probability of their coincidence given their joint distribution and their individual distributions, assuming independence:

$$\mathrm{pmi}(x;y)\equiv\log\frac{p(x,y)}{p(x)p(y)}=\log\frac{p(x|y)}{p(x)}=\log\frac{p(y|x)}{p(y)} \quad (6)$$

In Section 2 the Evolutionary Adaptive architecture for EGs is introduced together with the Treasure game and the user metric used as reference. The issues of adapting the genetic operator and representation to the EG context are discussed in the Section 3. Future work is discussed and conclusions are drawn in Section 4.

3. Knowledge-based Query Expansion System

Using CYC, certain objects in an image may be linked to related objects [2]. Such inferences will entail examination of the conditional probabilities P [Ji | Jj], where Ji, Jj are objects and Jj is given to be present in an image. Common sense association and ontology in CYC are used to construct an inference tree, which allows the index elements Xi's of an image to be automatically expanded according to given probability linked to the underlying ontology of the domain.

3.1 A Framework of CYC Based Query Expansion Image Search Model

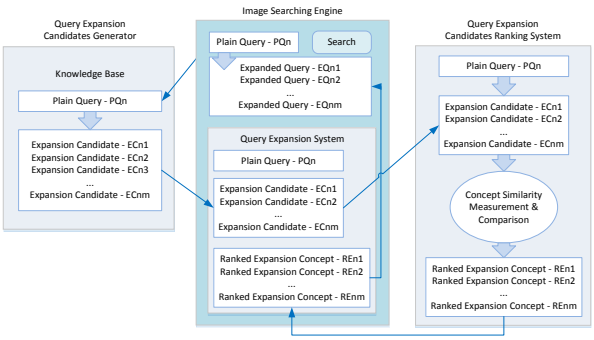


Figure 1. Knowledge-based Query Expansion System

As shown in Fig.1, the conceptual design of our CYC-WordNet based query expansion image retrieval and re-ranking model mainly contains three modules as follows:

- Image Searching Engine: It contains user interface which use could pass the plain query to the image searching engine, as well as the Query Expansion System which connect to the Query Expansion Candidate Generator and Query Expansion Candidates Ranking System.
- Query Expansion Candidate Generator: Based on CYC knowledge base, it generates the expanded candidates for further processing.
- Query Expansion Candidates Ranking System: Based on the similarity measurement results, it filters and re-ranks the expanded queries.

3.2 CYC-WordNet Based Query Expansion System Work

When user input the keyword (Plain Query PQn) in Image searching Engine, the query PQn is automatically passed to Query Expansion Candidate Generator. Query Expansion Candidate Generator, which is based on CYC knowledge base. It check if PQn has multiple concepts and relationships with other concepts and passes the generated linked concepts Expansion Candidates Array (ECn1, ECn2…ECnm) to Query Expansion Candidates Ranking System. In Query Expansion Candidates Ranking System, it processes the Expansion Candidates Array (ECn1, ECn2…ECnm) by calculating the concept similarity of PQn and (ECn1, ECn2…ECnm) based on WordNet Similarity. It filters to keep the top T similar concepts among (ECn1, ECn2…ECnm), give the filtered and re-ranked results Ranked Expansion Concept Array (REn1, REn2…REnm) by sorting the similarity measurement results. Then it passes the Ranked Expansion Concept Array (REn1, REn2…REnm) back to the Image Searching Engine. The Image Searching Engine gives the expanded queries back to user via the user interface.

3.3 Concept Distance Measurements

To measure the similarity between PQn and each candidate in Expansion Candidates Array (ECn1, ECn2…ECnm), we propose to use WordNet Similarity Measurements [3]. The similarity is calculated as follows:

$$d(c_1, c_2) = \frac{2 \times logp(lso(c_1, c_2))}{logp(c_1) + logp(c_2)} \quad (7)$$

In this calculation, we use PQn as c1 and each candidate in Expansion Candidates Array (ECn1, ECn2…ECnm) as c2.

4. Preliminary Experimental Results

Measures of performance are taken between the unaided approach similar to that in searching engines and the proposed approach for each individual query as well as collectively for their union. A set of representative semantic queries, which usually contain different confusing concepts, is designed for the experiments. The following are used to measure system performance of both approaches on the same image collection:

$$\text{Precision} = \frac{|\{relevant\_images\} \cap \{retrieved\_images\}|}{|\{retrieved\_images\}|} \quad (8)$$

A subset of the tested queries (here only shows the plain queries) is contained in Table 1. For each query, top five expanded concepts are tested to get the precision P1, P2 …P5. Then we take the mean value of them ((P1+P2+…P5)/5) as the precision of proposed approach for each plain query.

- average precision

The average precision is the mean value of all individual queries' search precisions.

For the unaided approach, we pass the original plain queries the searching engine. As we can see, the column of increased precision obviously indicates the significant advantage of our proposed approach over the unaided approach.

As we can see in Fig. 2, The average precision of the unaided approach is around 51% while it rises up to approximately 77% for the proposed approach, which is about 27% higher than the unaided approach. These results indicate that significant improvement in performance may be attained from using the proposed approach.

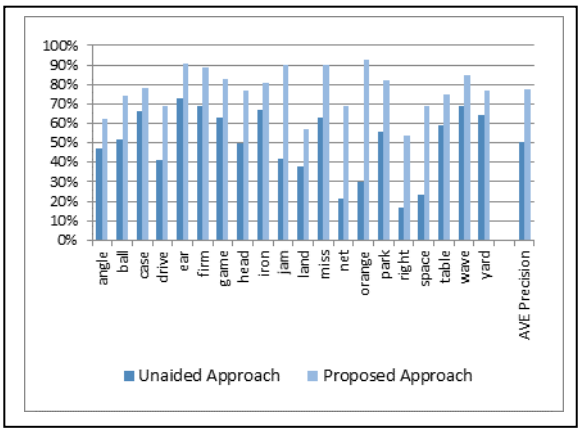


Figure 2. Experimental Results

5. Conclusion

As shown from the results of our experiments, with the proposed approach of the knowledge based query expansion system, the accuracy of web image searching has seen significant improvements. With the Query Expansion Candidate Generator and Query Expansion Candidates Ranking System, the expanded queries has been selected and refined by the combination system of CYC knowledge base and WordNet similarity. The semantic meanings and concepts of web images are significantly enriched. These experimental results clearly indicate the feasibility of the proposed framework.

There still exist some limitations in our proposed methods, such as the setting of the weights for expanded query words. It is also advantageous to minimize and optimize the processing time among the modules. Our future work consists of developing algorithms and approaches for optimizing to achieve more accurate and effective indexing results.

6. References

[1] MAZANDU, Gaston K, MULDER, Nicola J. "A topology-based metric for measuring term similarity in the gene ontology", *Advances in bioinformatics*, 2012, pp. 17

[2] WU, Zhibiao, PALMER, Martha, "Verbs semantics and lexical selection" *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1994, pp. 133-138.

[3] PEKAR, Viktor; STAAB, Steffen, "Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision", *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 2002. pp. 1-7.

[4] RESNIK, P, "Using information content to evaluate semantic similarity in a taxonomy", *arXiv preprint cmplg/9511007*. 1995.

[5] Kenneth Ward Church and Patrick Hanks (March 1990). "Word association norms, mutual information, and lexicography". Comput. Linguist, pp. 22–29.

[6] CHURCH, Kenneth Ward, HANKS, Patrick, "Word association norms, mutual information, and lexicography", *Computational linguistics*, 1990, pp. 22-29.

[7] JACCARD, Paul. The distribution of the flora in the alpine zone. *New phytologist*, 1912, pp. 37-50.

[8] LEUNG, C. H. C.; LI, Yuanxi, "CYC based query expansion framework for effective image retrieval.", *Image and Signal Processing (CISP), 2011 4th International Congress on*. IEEE, 201, pp. 1353-1357.

[9] MILLER, George A. WordNet, "a lexical database for English". *Communications of the ACM*, 1995, pp. 39-41.

[10] CYC Knowledge Base Official Website: http://www.cyc.com/

[11] OpenCyc Organization Official Website: http://sw.opencyc.org/

[12] WONG, Chun Fan. "Automatic semantic image annotation and retrieval*". Hong Kong Baptist University (People's Republic of China)*, 2010.

[13] BUDANITSKY, Alexander; HIRST, Graeme. "Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures.". *Workshop on WordNet and Other Lexical Resources*. 2001. pp. 2.2.

[14] JIN, Yohan, et al, "Image annotations by combining multiple evidence & wordnet.", *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 2005, pp. 706-715.

[15] ANDREOU, Agissilaos, "Ontologies and Query expansion. *Master of Science", School of Informatics, University of Edinburgh*, 2005.

[16] NATSEV, Apostol Paul, et al, "Semantic concept-based query expansion and re-ranking for multimedia retrieval", *Proceedings of the 15th ACM international conference on Multimedia*, ACM, 2007, pp. 991-1000.

[17] LI, Yuanxi; LEUNG, C. H. C, "Multi-Level Semantic Characterization and Refinement for Web Image Search.", *Procedia Environmental Sciences*, 2011, pp. 147-154.

[18] WONG, Roger CF; LEUNG, Clement HC. "Automatic semantic annotation of real-world web images.", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2008, pp. 1933-1944.

[19] REED, Stephen L., et al, "Mapping ontologies into Cyc", *AAAI 2002 Conference Workshop on Ontologies For The Semantic Web*, 2002, pp. 1-6.

[20] MATUSZEK, Cynthia, et al, "Searching for common sense: populating Cyc™ from the web.", *AAAI*. 2005, pp. 1430-1435.

[21] GAO, Yul; FAN, Jianping. "Incorporating concept ontology to enable probabilistic concept reasoning for multi-level image annotation.", *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, ACM, 2006, pp. 79-88.

[22] TORRALBA, Antonio; FERGUS, Rob; FREEMAN, William T, "80 million tiny images: A large data set for nonparametric object and scene recognition", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2008, pp. 1958-1970.

[23] TAM, A. M.; LEUNG, Clement HC, "Semantic content retrieval and structured annotation: Beyond keywords.", *ISO/IEC JTC1/SC29/WG11 MPEG00/M5738, Noordwijkerhout, Netherlands*, 2000.

Traffic Flow Analysis on Public Transport Access Data

Amarsanaa Ganbold[1], Tsolmon Zundui[2], Purev Jaimai[3]
Department of Information and Computer Science,
School of Engineering and Applied Sciences,
National University of Mongolia,
{amarsanaag[1], tsolmonz[2]}@num.edu.mn, purev[3]@seas.num.edu.mn

Abstract

An efficient data analysis of traffic flow plays an important role in achieving better transportation services. The aim of this work is to find out passengers' travel pattern from incomplete transport access data. Our proposed big data analytical model predicting endpoints of travel regularity gives significantly improved representation of live traffic behavior. We investigated nearly 38.3k patterns in three months data recorded 35M boarding actions.

Keywords: Public transit, Traffic flow, Data analysis

1. Introduction

Approximately 67.9% of automobiles in Mongolia were registered and running in Ulaanbaatar [1]. Therefore it is major cause leading to traffic congestion. Public transit has long been considered to provide an effective way to reduce congestion, air pollution. To improve public transportation services and encourage more people to use public transit, transit agencies have been striving to identify the key factors that attract transit riders [2] through studying their travel patterns. Traditional transit travel pattern analysis largely relies on rider satisfaction survey or travel diaries, which is very time consuming, costly, and difficult to implement due to low response rate and accuracy [3]. Ulaanbaatar public bus transit system has many problems and even cause of traffic congestion in Ulaanbaatar due to many bus routes overlapped. Current situation of Ulaanbaatar transportation as following [4]:

- Travel speed decreased over years
- Inadequate supply of public transport services and lack of experts and efficient management
- Inappropriate design for transport facility and maintenance

Smart card automated fare collection systems/Transport access data are being used more and more by public transit agencies [5][6]. While their main purpose is to collect fare, they also produce large quantities of very detailed data on onboard transactions. The use of, transport access data to track passengers' long term travel activities and patterns, such as the number of typical daily trip chains, common boarding or alighting stops, offers a far more convenient and efficient data source. Smart card\Transport access data records both temporal and spatial information for each riders, making it feasible to conduct individual travel pattern analysis.

The most of previous researches based on public transport access data extracted travel behavior information macroscopically rather than by analyzing individual passenger travel pattern and not optimized for a large dataset [2].

The purpose of this paper is whether data analysis can be used to study passenger pattern from transport access data. Extracting bus passengers' travel patterns from transport access data can be particularly challenging because of the pricing policy passengers don't scan when they alight even it is required. To deal with the data issue, this paper proposes a robust and comprehensive data analysis to extract grouped passengers' travel pattern within any two bus stops and regularly from a large dataset with incomplete information. Specifically, two major issues are examined in this study.

First, the spatial travel patterns for a particular transit passengers are investigated. Here "spatial travel pattern" means that transit passengers repeatedly visits the same or adjacent bus stops on multi-day basis. Then we move onto determine the regularity of a transit passengers' travel pattern, which refers to frequency of the similar trips for this transit passenger, and the frequency of the similar trips can be considered an effective measurement of travel regularity.

The objective of this study are to assist both transit agencies and transportation researchers by developing

data analytics procedure to extract individual passengers' travel patterns and travel regularity; and ensuring these data analytics are capable of processing huge transport access datasets.

The rest of the paper is organized as follows. Section 2 introduces backgrounds of our analysis datasets, hypothesis and data problem. Section 3 introduces the traffic flow model design. The model implementation and data analysis task is described in Section 4. Section 5 shows experimental result of the data analysis which used traffic flow model. Finally, we conclude the paper.

2. Background

Ulaanbaatar public transit incorporated began to issue smart card on August 1, 2015. There is only one type of system in Ulaanbaatar: flat fares. Transit passengers pay a fixed rate for buses by tapping their smart cards on the card reader when entering and passengers need to hold their cards near card reader device to complete transactions. The smart card scan system does store information on boarding and alighting locations and time.

The key information stored in databases therefore includes card ID, route number, driver ID, transaction time, remaining balance, boarding stop, and alighting stop. However, it is required tapping their cards when entering and exiting buses, due to design flaw in the system, tapping the card in exiting buses isn't affect later usage of cards or passengers. Hence, the most of passengers don't tap card when they exits or scan their cards right after the entrance. This would pose a data problem to be considered to find out the missing data of passengers' get off activities. Until November 1, 2015 approximately 550k cards sold and generated 170k card transactions in every day. These characteristics of the Ulaanbaatar smart card system create additional challenges for those seeking to process the data and useful knowledge from it. From these data problem, we hypothesized that

- it is possible to recognize individual passenger pattern from the only passengers' get-on access data
- it is possible to find out bus stops along the passenger travel.

To demonstrate and check the hypotheses for public transit bus passengers in Ulaanbaatar, we considered a typical travel time (in this case August 1, 2015 to November 1, 2015) and the transaction data from public transit agency of the city of Ulaanbaatar. The all card holder names are anonymised and using ID which generated for every card.

3. Modeling Traffic Flow

The passengers' travel pattern for between each boarding locations is likely to show certain travel pattern during a multi-day period. To retrieve these hidden and repeated travel pattern, we developed an algorithm that finding each passenger's boarding the bus stop chains in certain time and grouped by passengers that same travel pattern any two bus stops as a travel regularity. As explained before, in this context regularity means "frequency of the similar trip for each bus stops".

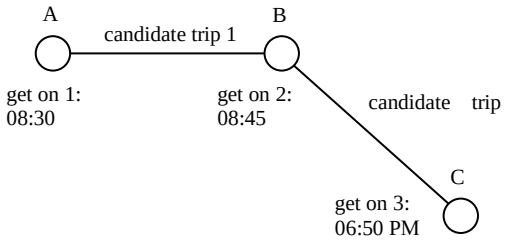


Figure 3. A sample spatial representation of get on record for an individual passenger on a certain day

A candidate trip shown in Fig.1 is a timely sequence of a passenger's get on records on two bus stops. If the passenger get on the bus stop A, he or she might get off a bus stop near B. Or, anyone who usually takes bus on the bus stop B in the morning and next get on event is occurred on the C in the evening, he or she goes between B and C. Therefore, a candidate trip is not exact trip between two points but a number of same trips can be a travel pattern because bus passengers usually repeat their travels.

Input: Bus Stop List (BSL), Passengers' Get On Records (PGOR)
Output: Travel Pattern List (TPL)
1. *sequence of list (SOL) = 0*
2. **for each** *passenger* in PGOR **do**
3. Create *a list of pairs* of two bus stops in BSL
4. Add the *list* to SOL
5. **end**
6. Extract *similar pairs* from SOL
7. Count *frequency* of *similar pairs*
8. Add *similar pairs*, *frequency* to TPL
9. **return** TPL

Figure 4. Pseudocode to extract the travel pattern of bus passengers

This model can retrieve passenger travel pattern from get on access data only.

To identify travel pattern regularity, we used clustering passengers with similar travel pattern and place them into different spatial characteristics. This information would help transit agencies evaluate and optimize transit services and bus routes. For instance, using K-means clustering algorithm [7] to retrieve to

where passengers travels from a bus stop which heavy loading. In other words, in order to find centroids, clustering passengers' ending points of travel patterns can be done. Then those centroids can spatially be visualized by lines which are connected to the bus stop coordinate. For a given dataset d_i, at (d_1,d_2,...d_n) and a given point (starting bus stop coordinate) p_{d_i} for dataset d_i, ending points of clustered patterns are returned from the vector function k_j which performs K-Means clustering algorithm.

$$C = \sum_{i=1}^{N} \sum_{j=1}^{M_{d_i}} p_{d_i} \times k_j(d_i) \quad (1)$$

This computation shown in the formula (1) gives C, total regularities of travel patterns where M_{d_i} , the number of centroids in dataset d_i and N, the number of datasets are.

4. Data Analysis Implementation

We have built 3 node Hadoop[3] platform where 1 name node server organizes all directory tree on the HDFS file system and tracks where across the cluster the file data is kept, and 2 physical data node servers stores cluster files.

Data analysis is conducted on 35M boarding data records from Ulaanbaatar public transport system. We used HIVE Data warehouse software as a tool to organize data and query from the data. This tool storing huge amount of data in distributed storage, divided into map-reduced jobs and processing on the jobs. We used HiveQL commands to get our main result which is passenger travel patterns and some other results. In table 1, five travel patterns that most passengers travel these two points out of 38387 travel patterns.

Table 3. A sample data for travel pattern

pattern	busstop_from	passengers	busstop_to
38343	Central Stadium	14244	Bayangol Hotel
38344	Yonsei Hospital	14915	Officers' palace
38345	120 Myangat	15340	Zaisan
38346	Bayangol Hotel	17299	Central Stadium
38347	Officers' palace	17454	Yonsei Hospital

Column Busstop_from is bus stop name which passengers' starting point of travel pattern, column busstop_to bus stop name which the passengers' ending point of travel pattern and column passengers is number of passengers who have same pattern between these two points.

After retrieve travel pattern data, we used IPython Notebook[4] interactive computational environment to data analytics. We extended this tool by many Python package for machine learning, spatial data visualization, math, and data analysis [8].

5. Experimental result

On travel pattern data, we created 26 intractive javascript graphics using data visualization tools highcharts, bokeh, folium, and tableau. For example, visualization of travel pattern on Ulaanbaatar map, clusters of travel patterns, transfer travel pattern, most loaded bus routes, volume of passengers in each bus routes, number of passengers in every time, etc.

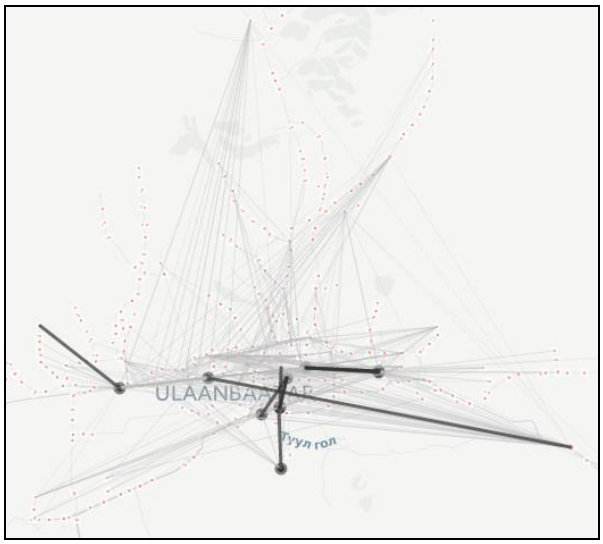


Figure 5. Passenger travel pattern

Fig. 3 shown most loaded 6 travel patterns with black line and a gray line represents a travel pattern of between 2400-18000 similar travel patterns of passengers.

In the result of this work we have found out 38.3k travel patterns from the passengers' travel behavior Ulaanbaatar public transportation.

[3] http://hadoop.apache.org/

[4] http://ipython.org/notebook.html

Figure 6. A partial of travel pattern clusters

Fig.4 shown 4 clusters of the most popular travel patterns. It includes between 263-302 travel patterns lines.

6. Conclusion

In this study, we calculated traffic flow of public transport passengers of Ulaanbaatar from UB Smart card system data. Therefore, K-means clustring of travel patterns shows that it can be good case of extracting knowledge from data. Furthermore, this study would help transit agencies planning, optimize bus routes and retrieved many informations that supporting decision making on many public transport related problems. Also it can provide useful informations for business sectors and public administrations.

In the future, evaluation method should definitely be developed for validation of our proposed model. Moreover, integrating data from transportation network, automobile congestion, population density and private businesses that would be give us more valuable information for smart city. Transit service performance [9] would be evaluated as well.

7. References

[1] "Number of vehicles, by region, aimags and the Capital," *Mongolian Statistical Information System*, 2016. .

[2] X. Ma, Y. J. Wu, Y. Wang, F. Chen, and J. Liu, "Mining smart card data for transit riders' travel patterns," *Transp. Res. Part C Emerg. Technol.*, vol. 36, pp. 1–12, 2013.

[3] K. K. A. Chu and R. Chapleau, "Augmenting Transit Trip Characterization and Travel Behavior Comprehension," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2183, no. -1, pp. 29–40, 2011.

[4] N. Tsevegjav, "Urban Transport System in Ulaanbaatar city," Ahmedabad, India, 2014.

[5] B. Agard, C. Morency, and M. Trépanier, "Mining public transport user behaviour from smart card data," in *In: The 12th IFAC Symposium on Information Control Problems in Manufacturing (INCOM*, 2006.

[6] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transp. Res. Part C Emerg. Technol.*, vol. 19, no. 4, pp. 557–568, 2011.

[7] E. M. C. E. Services, *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Wiley, 2015.

[8] Donne Martin, "Data Science IPython Notebooks," 2015. .

[9] J. K. Eom, J. Y. Song, and D.-S. Moon, "Analysis of public transit service performance using transit smart card data in Seoul," *KSCE J. Civ. Eng.*, vol. 19, no. 5, pp. 1530–1537, 2015.

Finding Prognostic Factors to MACE in Patients with Myocardial Infarction

Young Joong Kim[1)], Ho Sun Shon[2)], Man Geun Jeong[3)], Kyung Ah Kim[4)], Jong Yung Lee[5)]
Dept. of Computer Science[1), 3)] , Medical Research Institute[2)] ,
Dept. of Biomedical Engineering[4] , Dept. of Software Engineering[5)]
Chungbuk National University, Cheongju, Korea
{rex, kimka, jongyun}@chungbuk.ac.kr, {shon0621, jmg621}@gmail.com

Abstract

CVD (Cardiovascular disease) is the leading cause of death in the world. Of these, 50% of death rate has myocardial infarction. However, many researchers are studying prevention of cardiovascular disease for a temporarily healthy person. Namely, research on the cardiovascular disease is performing for the main purpose of primary prevention. Therefore, the purpose of this paper was to investigate significant prognostic factors associated with MACE (Major Adverse Cardiac events) in Korean patients with myocardial infarction. Through this research, we could find major prognostic factors to Mace in patients with myocardial infarction. The results of this paper were as follows. As the age increased, the more the mace rate increased. In the regression analysis, significant predictors of the MACE were age, Heartrate, NT-proBNP(<390/mL), Class III and Class IV of Killip classification, history of ischemic heart disease, history of diabetes mellitus, and complications. Consequently, our result will contribute to prognosis estimations through proper prognostic factors of myocardial infarction patients.

Keywords: Prognostic Factors, Myocardial Infarction, Cardiovascular disease

1. Introduction

Cardiovascular disease is the leading cause of death in the world. In 2015, over 50,803 deaths in Korea attributed to major cardiovascular diseases. CVD includes all the diseases of the heart and circulation including coronary heart disease, angina, heart attack, *congenital heart disease and stroke [*1]. Research on the cardiovascular disease is perfoming for the main purpose of primary prevention. Related work is representative of Framingham Heart Study [2,3] and Estimation of Cardiovascular Risk in an individual patient [4]. The Framingham Heart Study was initiated by the USA Public Health Service to study the epidemiology and risk factors for Cardiovascular disease. The concept of "risk factors", coined by Framingham Heart Study, involved gaining an understanding of factors predisposing to the occurrence of CVD. Nowadays, we define risk factor as a measurable element or characteristic that is causally associated with an increased rate of disease and that is an independent and significant predictor of the risk of presenting disease [3].

Cardiovascular risk estimation is one of a number of scoring systems used to determine an individual's chances of developing cardiovascular disease [5-7]. Those studies are studying prevention of cardiovascular disease for a temporarily healthy person. There is also lack of studying prevention for patients with cardiovascular diseases.

Therefore, the purpose of this paper is to investigate significant prognostic factors associated for MACE in Korean patients with myocardial infarction.

2. Population and Methods

Among 14,885 participants in Myocardial Infarction data, 11,831 participants aged <30 or >80 years were excluded. Patients who did not have all the variables to calculate the risk factors excluded at baseline. We also excluded patients with missing data on events.

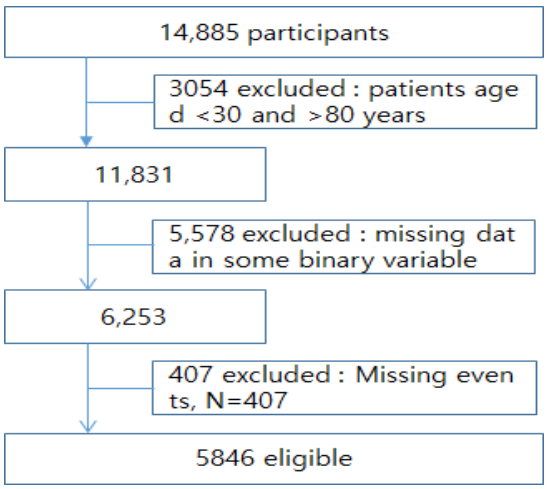


Figure 7. Patient inclusion flowchart

We summarized details concerning inclusion and exclusion of studies in Figure 1. Clinical end-points (events) defined as Cardiac death, Re-PCI, MI, and CABG. Predictors included age, current smoking, systolic blood pressure, body mass index, glucose intolerance, total cholesterol, and so on. The risk factors grouped into some categories.

Logistic regression analysis performed to identify the independent predictors of primary end-points. A p-value <0.05 was considered as statistically significant. All statistical analysis performed using the Statistical Package for Social Sciences (SPSS version 18).

3. Experiment Results

The total number of 5,846 patients with Myocardial Infarction for a mean of 1.01 years was included. Table 1 shows details of the characteristics of the eligible patients.

Table 4. Characteristics of baseline

Symbol		Men N=4310 Mean±SD	Women N=1536 Mean±SD
Age, year		58.3±11.1	67.5±8.9
Body Mass index		24.3±3.3	22.7±2.1
Systolic BP		130.0±27	131.2±28.6
Diastolic BP		80.0±16.3	78.9±16.4
Total cholesterol		180.7±39.4	188.3±41.9
Triglyceride		128.5±88.4	118.9±72.2
HDL-cholesterol		44.2±16.4	47.8±22.3
		%	%
Smoking	No	20.9	85.9
	Current	58.7	11.4
	EX	20.4	2.7
History of Hypertension	No	56.9	38.5
	Yes	42.3	60.9
History of Ischemic Heart Disease	No	47.7	78.4
	Yes	65.8	11.6
History of dyslipidemia	No	79.1	78.4
	Yes	10.3	11.6
Family history of heart disease	No	84.8	88.1
	Yes	8.1	4.8
Complications	No	90.4	89.4
	Yes	9.5	10.6
Killip classification	I	79.9	72.7
	II	10.6	13.2
	III	5.4	9.6
	IV	2.5	3.1

Logistic regression analysis was used to evaluate the prognostic factors associated with MACE for MI patient. Therefore, the major and independent prognostic factors for MACE are age, Heartrate, NT-proBNP(<390/mL), Class III and Class IV of Killip classification, history of ischemic heart disease, history of diabetes mellitus, and complications. (Table2)

Table 5. Results for proper prognostic factors of Myocardial Infarction patients

	OR	95% CI		p-value
		Lower	Upper	
Age	1.014	1.005	1.024	0.001
HeartRate	1.006	1.002	1.010	0.006
NT-porBNP(<390 pg/mL)	1.694	1.378	2.082	0.001
History of Ischemic Heart Disease	1.389	1.118	1.726	0.003
History of diabetes mellitus(yes)	1.369	1.127	1.662	0.002
Complications(yes)	1.423	1.095	1.850	0.008
Killip classification(III)	1.668	1.228	2.265	0.001
Killip classification(IV)	1.693	1.087	2.265	0.020

The study did not include whether or not using risk factors generally. The reason is why it is the basis of the patients with myocardial infarction.

4. Conclusion and Future Work

Research on the MACE is performing for the main purpose of primary prevention. In this paper, we were able to investigate significant prognostic factors associated with MACE in Korean patients with myocardial infarction. The results of this paper were as follows. As the age increased, the mace rate increased. In the regression analysis, significant predictors of the MACE were age, Class III and Class IV of Killip classification, history of ischemic heart disease, history of diabetes mellitus, and complications. Consequently, our result will contribute to prognosis estimates through proper prognostic factors of myocardial infarction patients. Based on this study, it is to identify

a more accurate prognosis, plus develop a prediction model for estimating the risk of patients with myocardial infarction.

5. Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No-2013R1A1A206518).

References

[1] S Mendis, P Puska, and B Norrving, "Global atlas on cardiovascular disease prevention and control.", World Health Organization, 2011.

[2] S.S. Mahmood, D. Levy, R.S. Vasan, and T.J. Wang, "The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective." *The Lancet*, 2014; Vol.383(9921), pp.999-1008.

[3] C.J. O'Donnell, R. Elosua, "Cardiovascular risk factors. Insights from Framingham Heart Study", *Rev Esp Cardiol*, Vol.61(3), 2008, pp. 299-310.

[4] Wilson, P. W., and B. F. Culleton. "Estimation of cardiovascular risk in an individual patient without known cardiovascular disease.", *UpToDate Textbook of Medicine. Waltham, MA: Massachusetts Medical Society, and Wolters Kluwer publishers*, 2010.

[5] Sr. D'Agostino, B. Ralph, S. Grundy, L.M. Sullivan, and P. Wilson, "Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation.", *Jama*, Vol.286(2),2001,pp.180-7.

[6] J Hippisley-Cox, C. Coupland, Y. Vinogradova, J. Robson, R. Minhas, A. Sheikh, and P. Brindle, "Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2.", *Bmj*, Vol.336(7659), 2008, pp.1475-82.

[7] R.B. D'Agostino, R.S. Vasan, M.J. Pencina, P.A. Wolf, M. Cobain, J.M. Massaro, and W.B. Kannel, "General cardiovascular risk profile for use in primary care the Framingham Heart Study. Circulation.", Vol.117(6), 2008, pp.743-53.

Automated Detection of Outliers in Cardiovascular Database

Man Geun Jeong[1], Young Joong Kim[2], Jong Yun Lee[3] Ho Sun Shon[4]
[1,3,4]*Dept. of Computer Science, Chungbuk National University, Cheongju, Korea*
{jmg621, shon0621}@gmail.com, jongyun@chungbuk.ac.kr
[2]*Medical Research Institute, Chungbuk National University, Cheongju, Korea*
rex@chungbuk.ac.kr

Abstract

Due to advent in cardiovascular disease over the world, research on this has been performed actively. In Korea, data of patient with cardiovascular disease have been collected, and the data has been stored into database and utilized in a number of studies. However, there is problem in anaylzing data, because there are many noises and missing values in collecting and managing the data. I*n this paper, we will detect outlier in preprocessing and make analysis of clinic data. This method intended to automate the preprocessing using box-plot and 2 Standard Deviation(2SD) and to analyze the data efficiently.* According *to the analysis result, box-plot has better results than 2SD affected by outlier in cardiovascular patient data. Hereby, we will detect outlier automatically and can be used for further clinic* study *and* diagnostic research.

Keywords: Cardiovascular database, Outlier detection, Box-plot, 2 standard deviation

1. Introduction

Recently, Big-Data has become widespread in the world [1]. Therefore, there have been requested methods to get valuable information by using the data in various fields [2]. In the medical field, clinical studies and diagnostic studies have been researched based on a vast amount of information data [3].

There are a lot of difficulties in preprocessing of error data to clinical studies and diagnostic studies, because patients data used in research contains noise, outlier, and blank. Therefore, the purpose of this research is to develop program that searches for errors and removes them.

2. Related Work

An application of data mining in health care management to detect abnormal values presented in medical databases [4]

Outliers in medical databases can be caused by measurement errors or may be the result of inherent data variability. The abnormal value of mitoses, for instance, could lead to the diagnosis of malignant cancer or it might just be due to human mistake or execution error. The results of the experiment show that outlier mining i.e. outlier detection and analysis has a great deal of potential to find useful information from health care databases which consequently helps decision makers to automate and quickens the process of decision making in clinical diagnosis as well as other domains of health care management [4].

Box plots were used to detect univariate outliers directly whereas the box plotted Mahalanobis distances identified multivariate outliers. The removal of outliers increased the descriptive classification accuracy of discriminant analysis functions and nearest neighbour method, while the predictive ability of these methods is reduced somewhat. Outliers were also evaluated subjectively by expert physicians, who found most of the multivariate outliers to truly be outliers in their area.

The informal method may be used for straightforward identifying suspicious data or as a tool to collect abnormal cases for an in-depth analysis [5].

3. Methods

In this paper, data from 14,885 patients with acute myocardial infarction are used. Representatively, heart rate of attributes will be used as an example.

3.1 Outlier detection by using box-plot

Box-plot is a general method for detecting outliers. [8]. This is graph expressing figures and created by "lower extreme, lower quartile(Q1), median(Q2), upper quartile(Q3), and upper extreme" from data [6,7]. Also, distance from Q3 to Q1 is called Interquartile range (IQR). The formulas of the Low thresholds and Upper thresholds are as follows.

$$Low\ thresholds = lower\ quartile - (1.5 * IQR) \quad (1)$$

$$Upper\ thresholds = upper\ quartile + (1.5 * IQR) \quad (2)$$

If the result is less than Low thresholds or greater than upper thresholds, then it is an outlier.
The following code is algorithm derived from formulas as stated above.

Algorithm1: Box-plot

```
Input:   (1) rows : continuous variable data
Output:  list
FOR i = 0 to number of rows Do
  IF row[i] is not empty then
    IF row[i] is not noise then
      Insert into list and order the list
Q_3 = list's upper quartile
Q_1 = list's lower quartile
MINVALUE = Q_1 - (1.5 * IQR)
MAXVALUE = Q_3 + (1.5 * IQR)
FOR i = 0 to number of rows Do
  IF list[i]'s value < MINVALUE then
    drop list[i]
  IF list[row's size − i − 1] > MAXVALUE then
    drop list[row's size − i − 1]
  IF both case Not drop then
    FOR loop out
RETURN list
```

Figure 8. Outlier detection algorithm using Box-plot

3.2 Outlier detection by using 2 SD

Standard deviation is an indicator showing how far they are dispersed from mean. Small standard deviation means that most data are concentrated in the nearby mean. And, supposing that normal distribution is exist within $\bar{x}\pm2\sigma$, 95% of data are included. Also generally, if $|x_i - \bar{x}| / \sigma > 2$ the data is regarded as outlier [8].

However, we repeatedly detected until outlier doesn't appear because the standard deviation is also influenced by outlier. The following algorithm2 is for detecting outlier used by 2SD.

Algorithm2: 2SD

```
Input:   (1) rows : continuous variable data
Output:  List
FOR i = 0 to number of rows Do
  IF row[i] is not empty then
    IF row[i] is not noise then
      Insert into list and order the list
WHILE outlier existed
  means = list's Means
  std_dev_2 = multiply list's Sample Standard Deviation by 2
  FOR i = 0 to number of rows Do
    IF list[i]'s value - means < -std_dev_2
      drop list[i]
    IF std_dev_3 < list[row's size − i − 1]'s value - means
      drop list[row's size − i − 1]
    IF both case Not drop
  FOR loop out
WHILE out
RETURN list
```

Figure 9. Outlier detection algorithm by 2SD

4. Result

4.1 Description of raw data

The total number of original patient data is 14,885 and 1,115 of them are blank. Mean is 78.32, min is 0, max is 999.lower quartile is 65, upper quartile is 88, and median is 76.

Following figure 3 shows the distribution of raw data. 9 outliers of data containing 720(1), 740(1), 941(1), 999(6) don't appear in the histogram.

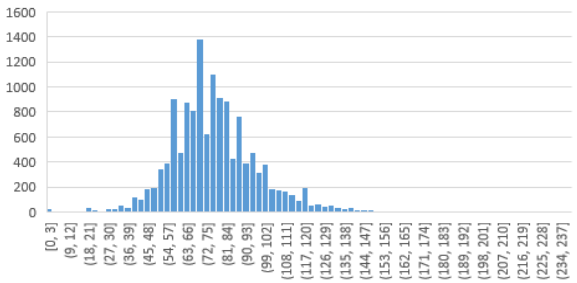


Figure 10. Histogram of raw data

4.2 Experimental results of proposed algorithm

When using a Box-plot, 476 outliers are detected and mean of data is 76.74, min of data is 29, max of data is 124. Also, lower quartile is 65, upper quartile is 88, and median is 76.

When using 2SD algorithm, 196 outliers are detected and mean of data is 77.15, min of data is 18, max id

data is 138. lower quartile is 64, upper quartile is 88, and median is 76.

We compared original data with things as follows from outlier detection used by 2 algorithms and data through figure 4 and figure 5.

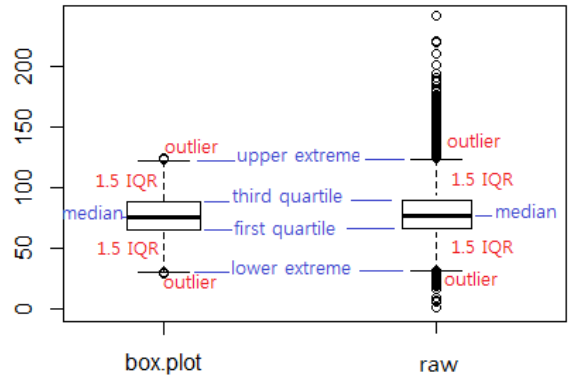


Figure 11. Comparison of box-plot and raw data

In Figure 4, the right side is raw data and the left side shows the result of using boxplot.

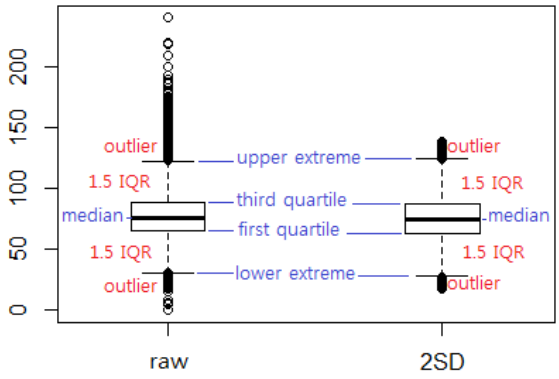


Figure 12. Comparison of raw data and 2SD

In Figure 5, the left side is raw data and the right side shows the result of using 2SD by boxplot. While box-plot is finding 476 outliers, 2SD find the 196 outliers. As a result, box-plot showed better performance than 2SD.

5. Conclusion

There are a lot of difficulties in preprocessing error data to clinical studies and diagnostic studies. There are problems when collecting and managing data because many noises and missing values appear. Therefore, we will detect outlier in preprocessing and analyze clinic data. We made automatic detection of outlier program using box-plot and 2SD. Consequently, box-plot has better results than 2SD affected by the outlier in cardiovascular patient data. Based on this research, we detect outlier automatically and this research can be used for further clinic study and diagnostic study.

6. Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No-2013R1A1A206518).

7. References

[1] Ben-Gal, Irad. "*Outlier detection", Data mining and knowledge discovery handbook*. Springer, US, 2005, pp. 131-146.

[2] Gupta, Manish, et al. "Outlier detection for temporal data." *Synthesis Lectures on Data Mining and Knowledge Discovery 5.1*, US 2014, pp. 1-129.

[3] Collins, Gary S., and Douglas G. Altman. "Predicting the 10year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2." *BMJ*, 2012, e4181, pp. 1-12

[4] Kumar, Varun, Dharminder Kumar, and R. K. Singh. "Outlier mining in medical databases: an application of data mining in health care management to detect abnormal values presented in medical databases." *International Journal of Computer Science and Network Security 8.8*, 2008, pp. 272-277.

[5] Laurikkala, Jorma, et al. "Informal identification of outliers in medical data." *Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology*, 2000, pp. 20-24

[6] WALFISH, Steven. "A review of statistical outlier methods". *Pharmaceutical technology*, 2006, pp.11- 82.

[7] BEYER, H. Tukey, John W, "Exploratory Data Analysis. Addison-Wesley Publishing Company Reading, Mass", *Menlo Park, Cal., London, Amsterdam, Don Mills, Ontario*, Sydney 1977, XVI, 688 S. *Biometrical Journal*, 1981, pp.413-414.

[8] SEO, Songwon, "*A review and comparison of methods for detecting outliers in univariate data sets*", *PhD Thesis. University of Pittsburgh*, 2006.

Online Motivation Analysis Model over Cloud Computing Environment

Hai Jing Jiang, Zhi Yuan Chen, Wei Ding, Tie Hua Zhou, Ling Wang
Department of Computer Science and Technology, School of Information Engineering, Northeast Dianli University, Jilin, China
Haijing_103702@126.com, 64999329@qq.com, thzhou55@163.com, smile2867ling@163.com

Abstract

In recent years, with the development of Smart Grid, its running process power data is becoming more and more diverse, complex. How to process the data, the traditional technical incompetence, only use cloud computing. Cloud computing will use a variety of resources through virtualization technology, form abstract service model, thus mining data at depth from power source fusion. The purpose of our paper is to establish an online motivation analysis model.

Keywords: Smart Grid, Cloud computing, Online motivation analysis model

1. Introduction

Smart Grid is the development direction and trend of electric power industry. Smart Grid using advanced information and communication, computer, control technology and other advanced technologies, realization of power generation, grid operation, use of terminal and the needs of all stakeholders in the electricity market and functional coordination, maximize reliability, self-healing capacity and stability of the system [1].

Emerging areas such search engines, E-Commerce, social network, data type and size is growing an un-precedent rate, and in Smart Grid system, data from all aspects of the system as a whole, was one of the key tech areas of application[2].

With the development of Smart Grid, mass development of smart meters and sensor technology widely used, electric power industry generated a lot of varied and complex data sources, how to store and use such data, is a difficult problem faced by power companies; and the emergence of cloud computing has occurred an upheaval.

Cloud computing is a model based on Internet related services, primary via a network to provide dynamic, scalable, and virtualized resources constantly. It could use software, storage, networks, servers, and other types of resources through virtualization technology, establishing an abstract service model.

The perfect combination of cloud computing and grid systems, all pending massive power data and computing resources for effective integration to shape a versatile data-processing and information sharing platform, which solve bottleneck that Smart Grid encountered in the era of big data

2. Related works

Cloud computing technology for the application of the power industry in China is still at the exploratory stage, roughly divided into the following several aspects: system building, data storage, and the complex relationship mining. About power system, Xie Huacheng et al. design the cloud computing resources management platform framework and part of the module, commit to achieving ERP data backup of the electric power enterprises. Still has no specific implementation [3]. Li Feng et al. address design architecture and subordinate levels about the simulation of cloud computing center: device power cloud, the data management cloud, simulation of cloud [4].

In regard to data storage, abundant data from smart grid could take advantage of the distributed file system to store, such as utilize HDFS of Hadoop, However, these systems could store large data, hardly satisfy real-time response [5].

Unknown found model in time series models, Heng Tang et al. proposed a novel k-Motif-based algorithm to solve the problem of existing data mining, in addition by summary the motivations which be found, to generate the original model. This method does not require an increase ‘w’ evaluation in the original k-Motif algorithms and identify different lengths of motif needs to be run only once [6]. J. Lin et al. address the

problem of finding repeated patterns in time series, and introduced an algorithm to efficiently locate them. In addition, introduce the first discrete representation of time series that allows a lower bounding approximation of the Euclidean distance [7].

Smart Grid research issues mainly reflected in the complex structured of data, fragmented, and real-time:

(1) Massive data, multidimensional data and multiple types of data.

(2) Scattered placed, distributed management of the data source

3. Big data fusion of multiple source in Smart Grid

The prospective of Smart Grid require the transfixion the generation, transmission, transformation, distribution, electricity, dispatch and many other areas, achieve overall collection, fluid transmission, efficient handling of information, supporting power, information and business flows of highly integrated. Therefore, the primary function will be to achieve large-scale integration of multisource heterogeneous information, provided resource intensive configuration to Smart Grid data center.

Northeast area of China has achieved massive cover age of smart meters in 2015. It has completed the data collection infrastructure basically, major data source is the ubiquitous sensor network that through high-speed communications networks concentrated on master control center, make the power use measurable and controllable to be come true.

Moreover for abundant heterogeneous data, the power industry has not yet established a master control center to deal with such big data, and existing systems and methods cannot satisfy the growing demand for data processing, distributed processing of cloud computing can only be used to handle multisource data fusion.

Figure 1 Tri-flows integration in Smart Grid

4. Online motivation analysis model

With the widely deployment of terminals such as smart meters, smart power plugs, etc, a large-scale Smart Grid could be constructed. Intelligent substations, EV power stations and other projects put into operation as well as wind, solar and other intermittent sources of energy access, the electric power industry is at a critical turning point in the information age, the data resulting from grid will be more and more complicated. For electric flow and information flow aspects, demand to achieve real-time processing to reduce the uncertainties of power random access. On the basis of collection that big data, develop an online motivation analysis model, intended for range beyond plans, suddenly appeared of electricity peak and renewable energy in some factors that caused power insufficient [8], it could find out the cause of the event trigger and based on data analysis fast and real-time to reduced power supply of pressure, ensure that level off supply and use. We build an online motivation analysis model to achieve the following functions:

(1) **Power segment:** continuous, high degree of automation, require real-time monitoring of the entire process, high-speed, real-time data processing and long-term storage of historical data and the productions of information integration and sharing. Future solutions of Smart Grid would require real-time response, even if a node failure. The monitoring and control of mass small distributed power will be the next challenge for power systems. Online motivation analysis model own strong scalability, but also according to the dynamic size of power system enhanced computing ability at any time

(2) **Transmission and distribution segment:** condition monitoring with high requirement on data storage and processing platform of performance or real-

time, future systems requires real-time processing of several orders of magnitude more than the current monitoring data. Online motivation analysis model could facilitate every control center of information sharing and collaboration.

(3) **Power segment:** the future environment of the Smart Grid, families may be equipped with a variety of electrical, power monitoring equipment to achieve low-cost electricity, and matches the load on the grid. With the growing interaction between the company and user, real-time data become more and more significant. Online motivation analysis model through online data analysis, we can better understand the behaviors of electricity customers, comprehend power demand, provided support for short-term load forecasting.

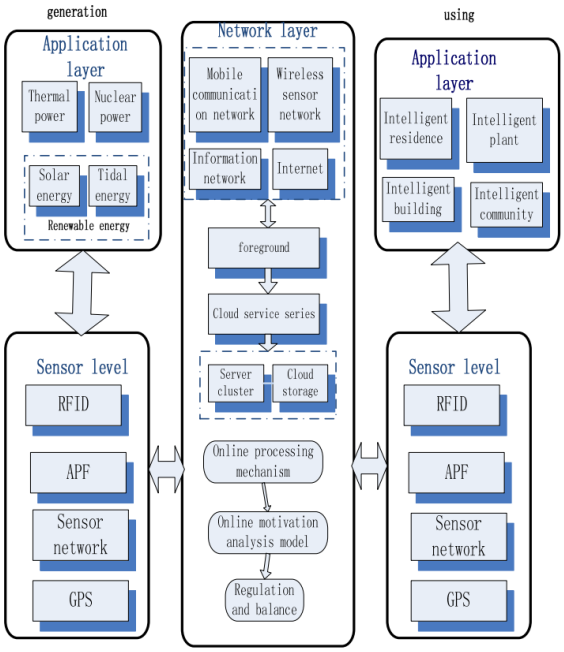


Figure 2 Online motivation analysis model

4. Conclusion

Cloud computing platform with powerful computing and storage capacity, through integration of abundant heterogeneous distributed computing resources [9], provide new approaches to solve complex computational problems in power systems. Online motivation analysis model based on cloud platforms contribute to online operation and optimization of control of power system analysis.

5. Acknowledgements

This work was supported by the Education Department Foundation of Jilin Province (No.201698), by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, and by the Science and Technology Plan Projects of Jilin city (No.201464059).

6. References

[1] Fang, X., Misra, S., Xue, G. and Yang, D., 2012. Smart grid—The new and improved power grid: A survey. Communications Surveys & Tutorials, IEEE, 14(4), pp.944-980.

[2] The Chairman of electrical engineering society informati zation of China. China's white paper on the development of electric power data [S]. 2013.

[3] Huacheng, X. and Xiangdong, C., Unstructured data access for cloud storage [J]. Computer Application, 2012, 32(7): 1924-1928.

[4] Feng, L., Jun, X. and Jinbo, L., The outlook and discussed of Intelligent substation relay protection configuration [J]. Electric Power Automation Equipment, 2012, 32(2):122-126. [5] Apache. Apache Hadoop core [EB/OL], 2012-08

[5] Apache. Apache Hadoop core [EB/OL], 2012-08

[6] Tang, H. and Liao, S.S., 2008. Discovering original motifs with different lengths from time series. Knowledge-Based Systems, 21(7), pp.666-671.

[7] Lonardi, J.L.E.K.S. and Patel, P., 2002. Finding motifs in time series. In Proc. of the 2nd Workshop on Temporal Data Mining, pp. 53-68.

[8] Guangbin, Z., Hongchun, S. and Jilai, Y., Using the modulus of generalized current traveling wave measured book a semi-supervised clustering screening[J], Proceedings of the CSEE, 2012, 32(10):150-158.

[9] Thusoo, A., Sarma, J. and Jain, N., HIVE: a warehousing solution over map-reduce framework [C], VLDB, 2009: 1626-1629

Horse stamp detection in real nomadic environment

Gantuya Perenleilkhundev, Bold Zagd, Suvdaa Batsuuri
School of Engineering and Applied Sciences, National University of Mongolia
gantuya@seas.num.edu.mn, bold@num.edu.mn, suvdaa@num.edu.mn

Abstract

Horse stamp recognition is important task for nomadic life. In real nomadic environment, the horse detect is one problem for stamp recognition. In this paper, we propose a detection method based on Mongolian horse-color. Some researchers are determined many horse-colors. In this time we studied the most common 12 horse-colors and determined RGB color scheme for detection. Our proposed method achieved reasonable in the experimental results.

Keywords: Horse detection, horse-color

1. Introduction

Nowadays, horse-color research is becoming interesting topic. L.Munkhtur et al determined more than 300 horse-colors and named basic colors are white, black, bay, brown, roan, tan, beige, grey and Isabella. [3]
Mongolian horse naming can be divided into 2 categories: Dark and light color. When identifying the color of a horse, main body, hair, skin, shrike, mane, tail, fetlock and leg/stem color are taken as main considerations. Horse naming include: A black horse: The main body, mane, tail, fetlock and stem colors are black. But if the main body is faded and pale black, while the mane, tail and stem are black, it is identified as grullo horse. A horse is called khaltar if main body is black and the eyes, muzzle, breast and flank is beige. A horse is named sorrel horse if the main body is of pure reddish tone; forelock, mane and leg color is similar or slightly brighter in color than the main body. Sorrel color is divided as brown, red, fawn, and reddish-yellow, depending on the hue. A bay horse-alezan: Main body color is burgundy brown. Mane, tail, fetlock and cannon are black. Depending on the hue, a bay horse color can be divided into yellow, brown, black and red bay. But if the eyes, muzzle and inner thigh are brighter beige- it is called a bay-khaltar. A chestnut horse-umber: main body is of dark color. Forelock, mane, tail consists of brown and black mix. A dull grey: Main body is brown and sorrel. But forelock, crest, tail, fetlock is of bright and dark mix. Muzzle, eyes, inner thigh is usually bright pink. Depending on the hue, dull color can be divided as chestnut dull and sorrel dull. Grey horse: Main body hair color is mainly white with a mix of dark hair. The skin of a grey horse is darker. Grey horse is born with darker color, but becomes brighter as it grows-up. Depending on the main body color, grey horse can be named as red-grey, navy grey and black-grey. A red dun horse: Main body color is similar to that of sorrel horse. Mane, tail, fetlock color is brighter. They usually have a darker line on the back and darker lines on the legs. There are some red dun horses who don't have these dark linings also. Depending on the tail and mane colors, they can be named as red-dun, dun, ukhaa-dun etc. A dun horse: Main body color is yellowish and of orange tone but mane, tail, cannon and the back is black. Shoulders and upper part of the leg has black stripes. A buckskin horse: Main body is of pale sorrel brown color mixed with black. Breast and inner thigh parts are comparably brighter; mane, tail, cannon and the back is black. A palomino horse: Main body is golden yellow or bright. Mane, tail is of same color of the main body or brighter. A palomino horse is sub-identified as white palomino and dun palomino. [1,2]
A grullo horse: Main body is ash-greyish. Mane, tail, fetlock, cannon is black, with black lining on the back; inner thigh is slightly brighter than the main body color. A spotted horse: Main color of the body is bright, with darker spots on it. Depending on the spot color, a horse can be named as brown-spotted, black-spotted etc. A roan horse: Main body consists of white and dark mixed colors. Head, mane, tail and legs are in darker color. Main body color is taken into consideration, when identifying roan horse. Mongolians do not brand their horses on random spots, but rather stamp the branding on a visible spot, preferably on the right or left side of the hip, thigh or the buttocks. On rare occasions, branding can also be done on parts nearby the shoulders.

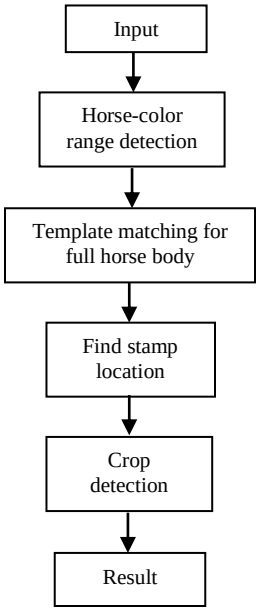
Generally, Mongolians do not brand and earmark their livestock on random spots, especially their fast horses trained for races. Branding is usually done on the right side of the thigh. If there's a nearby herder with similar branding symbol, they choose a different branding spot. For example, if the herders have similar moon shaped

branding, one of them puts the stamping on the right side, while the other stamps on the left side of the thigh. [3,4]

In general, there are many methods using color range for object detection. One of the most popular work is skin-color detection for face recognition or hand gesture recognition task [8, 9]. Our idea originated from the works [6, 7]. In the next chapter we show a horse-color detection method.

2. Method

We apply proposed method for the most popular 12 horse-colors. First of all, we determined RGB color range from sample images by manually. After than we a system



that detects horse based on horse-color range. Figure 1 has shown the system structure.

Figure.1. System structure of the proposed method.

2.1. Horse body parts and their color range

We divide horse image into two parts for color range. First one is the main body part and the other one includes mane, tail, fetlock and cannon. Figure 2 has shown the two parts of the horse image.

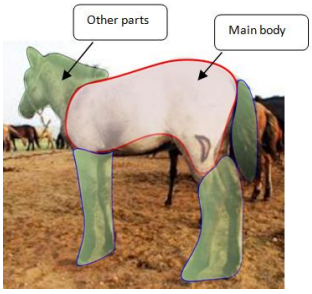


Figure. 2. Roan horse body parts

2.2. Horse color range dtection

In this part we detect horse based on the horse colors ranges. Table 1 has shown the average color range of 12 horse-colors.

Table 1. The average color range of 12 horse-colors.

	Horse-color	**RGB color range**	**Color**
1	Black	R-20, G-19, B-53	
2	Chestnut	R-140, G-91, B-77	
3	Bay	R-138, G-49, B-9	
4	Brown	R-126, G-52, B-7	
5	Reddish Yellow	R-159, G-135, B-135	
6	Dark Brown	R-197, G-204, B-210	
7	Tan	R-185, G-153, B-114	
8	Beige	R-224, G-181, B-139	
9	Isabella	R-227, G-196, B-150	
10	Golden	R-196, G-128, B-53	
11	Grey	R-213, G-212, B-218	
12	Roan	R-157, G-156, B-172	

2.3. Template matching for full horse body detection

As a result of color range detection of the input image, usually some parts are not detected. Therefore, we need to improve the shape of the full body.

In this part we use Template matching algorithm [5].

2.4. Stamp detection

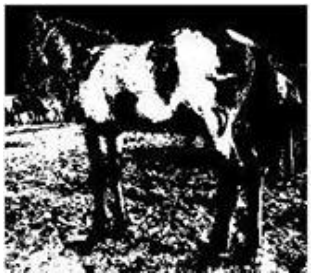
In formally, Mongolian horse stamp is located in the upper part of the right leg. After detection full body of the horse, it is needed to find out right leg and the stamp. [4]

3. Experimental Results

3.1. Result 1. Object and background colors are similar case.

In this experiment, we implement the proposed method in Matlab 2015b. The selected color range was R>100, B>G, B>R for Roan color.

3(a)



3(b)

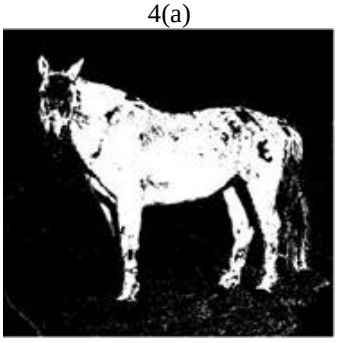
3(c)

Figure 3. Horse-color based stamp detection. (a) Input image (b) Horse detection result (c) Stamp location detection

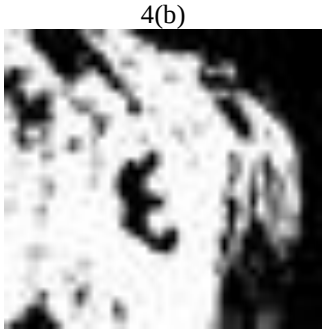
3.2 Result 2. Object and background are different color range.

In this experiment, the selected color range was R>100, B>G, B>R for Brown color. In some case, horse detected successfully. So we don't need to improve result by template matching.

4(a)



4(b)



4(c)

Figure 4. Horse-color based stamp detection without Template matching. (a) Input image. (b) Horse detection result. (c) Stamp location detection

The detection results depend on the environment background. In Figure 4, the object has different color range from background. Therefore, there is no need to improvement parts.

3.3. Result 3. Total detection with different background complexities. Table 2 has shown the result.

Table 2. Total detection

Input images	Background color	Detection results
120 images (12 colors x 10 images)	simple	92.3%
120 images (12 colors x 10 images)	complex	84.1%

4. Conclusion

In this paper, we propose a detection method based on Mongolian horse-color. We studied the most common 12 horse-colors and determined RGB color scheme for detection. Our proposed method achieved reasonable in the experimental results.

5. Reference

[1] S.Dulam. "Mongolian Symbolism" 2010

[2] Kh.Perlee. "Бүтээлийн чуулган" 2012

[3] L.Munkhtur, "Words' mean and structure of Mongolian horse color feature", *PhD thesis*, 2007

[4] J.Saruulbuyan and A.Davaasambuu, "Mongolian horse stamp explanation", 2011

[5] Duc Thanh Nguyen, "A Novel Chamfer Template Matching Method Using Variational Mean Field", *In CVPR*, 2014.

[6] Weimeng Chu1, Fang Liu, "An Approach of Animal Detection Based on Generalized Hough Transform", *International Conference on Computer, Networks and Communication Engineering*, 2013

[7] Fernando Rendo, Mikel Iriondo, Carmen Manzano, and Andone Estonba, "Identification of horse chestnut coat color genotype using SNaPshot", *BMC Res Notes*, 2009, pp. 255

[8] Bahare Jalilian1, Abdolah Chalechale, "Face and Hand Shape Segmentation Using Statistical Skin Detection for Sign Language Recognition", *Computer Science and Information Technology*, 2013, pp. 196-201..

[9] A.R.Chandra Suresh, R.Upendar Rao, and Ramakrishna P, "Real-time Hand Gesture Detection and Recognition Robot Using ARM7", J*ournal of Research in Electrical and Electronics Engineering (ISTP-JREEE*), 2014, Volume 3.

Survey on 3D model based pose estimation methods

E.Tsetsegjargal[1], R.Javkhlan[2], D.Usukhbaatar[3], B.Suvdaa[4]
School of Engineering and Applied Sciences, National University of Mongolia
{tsetsegjargal, javkhlan, suvdaa}[1,3,4] @seas.num.edu.mn, osohoo02@gmail.com[2]

Abstract

Many applications require tracking of complex 3D objects. These include visual implementing of robotic arms on specific target objects, Augmented Reality systems that require real-time registration of the object to be augmented, and head tracking systems that sophisticated interfaces can use. Computer Vision offers solutions that are cheap, practical and non-invasive.

This survey reviews the different techniques and approaches that have been developed by industry and research. Those techniques are given of the numerous approaches developed by the Augmented Reality and Robotics communities, beginning with those that are based on point or planar marks and moving on to those that avoid the need to engineer the environment by relying on natural features such as edges, texture or interest. The survey concludes with the different possible choices that should be made when implementing a 3D tracking system and a discussion of the future of vision-based 3D tracking.

Keywords: Object tracking, object detecting, pose estimation. 3D tracking

1. Introduction

To make things easier, some 3D knowledge is often used for pose-estimation. The 3D knowledge usually comes in the form of a CAD model of the object which can be created using either reconstruction methods or commercially available software. This type of pose-estimation is called Model-Based Pose-Estimation.

Within the model-based pose-estimation techniques, we can distinguish between five families of approaches depending on the image features used to evaluate the similarity of the pose [12]. The first one relies on edges of both the image and the projection of the target 3D object. The second one includes techniques that use pixel-wise information, including statistical information, optical flow, interest points or template matching.

A 3D model-based pose-estimation algorithm usually follows the path shown on Figure 2.1. The main purpose is to find the correct pose of the 3D model which rendered will match as best as possible with the input image or frame of a sequence. The evaluation to determine if the current pose is similar enough to the image and the different techniques are described in 2.

2. State of the Art

As mentioned 3D model-based pose-estimation can be divided into two different types of methods depending on the image features used to evaluate similarity between the rendered model and the input image. In this Section, those methods are described, including some example figures and important facts.

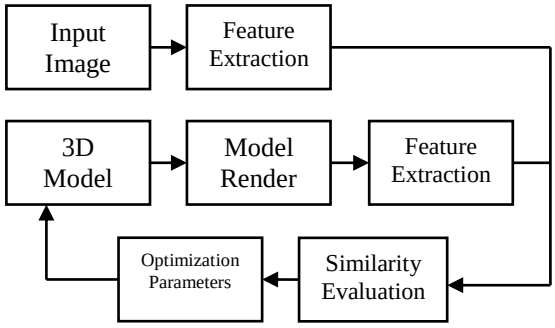


Figure 2.1: Model-Based Pose-Estimation Process

2.1. Edge-based methods

The first approaches for 3D pose-estimation were all edge-based. Mostly because of the simplicity of implementation but also because these methods are computationally efficient. Edge-based algorithms have proved their efficiency and robustness to lighting changes and specular highlights. These methods can be divided into two main groups:

- Extraction of strong gradient without explicitly extracting contours.

- First extract image contour to fit the model to this contour.

A quick overview of these methods is shown in the following paragraphs.

2.1.1. RAPiD[10]

Basically this algorithm uses a set of 3D control points sampled along the 3D model edges and in the areas of high-contrast or rapid albedo change. The 3D motion of the object between two consecutive frames can be recovered from the 2D displacement of the control points. Recent systems include optimization algorithms as well as coarse-to-fine implementations such as [16].

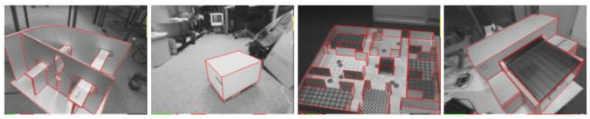


Figure 2.2: Example of algorithm that uses RAPiD basics. [11].

2.1.2. Edge Extraction

Another approach is to globally match model primitives with primitives extracted from the image. For each image, straight line edge segments are extracted, while the 3D model edges are projected with respect to the estimated pose. The matching is based on the Mahalanobis distance of line segment attributes. Figure 2.3 refers to an algorithm [14] that implements edge extraction technique.

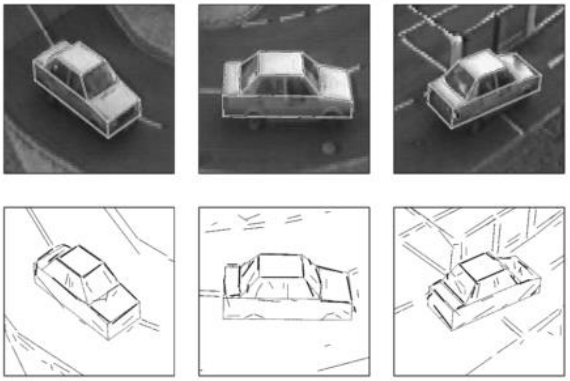


Figure 2.3: Example of Edge Extraction [14].

2.1.3. Direct optimization on Gradients

This approach tries to fit the model projection directly on the gradient image from the input frame. Obviously it is not guaranteed that the strong gradients will match with the model's edges.

As mentioned, these algorithms are robust for changes in the lighting of the scene as well as specular highlights on objects since the information used is based on pixel relations with their neighbors, rather than information from itself. This may seem to work in many circumstances, but edge-based techniques can suffer from the normal drawbacks that arise from using local image features like high sensitivity to noise or missing information, and a multitude of local minima that result in poor segmentations.

Cluttered scenes, due to their nature, can produce the above conditions that will make the segmentation and pose-estimation fail.

2.2. Optical Flow-Based Methods

Optical flow is the apparent motion of the image projection of a physical point in an image sequence, where the velocity at each pixel location is computed under the assumption that projection's intensity remains constant. It can be expressed as

$$m' = m + \begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} dt, \qquad (1)$$

where m the projection of a point in an image I at time t, m' its corresponding location in the next image, captured at time $t + dt$, and $(\dot{u}, \dot{v})^T$ the apparent speed of the 2D motion at m. The vector field of the $(\dot{u}, \dot{v})^T$ is the optical flow. It can be computed using the Lucas-Kanade method [17] for example, which adopts a multiscale approach and assumes that the optical flow varies smoothly.

2.2.1. Using Optical Flow Alone

The Optical Flow calculates the apparent motion of a single pixel within two consecutive frames of a sequence, based on the assumption that the intensity or colour of the used pixels is not varying from frame to frame, but specially, on the assumption that the initialization of the system is correct. Basically the optical flow between the current and the previous frame is computed continuously, and then, the algorithm tries to find the best pose parameter that matches with the obtained flow.

There are also some algorithms that have tried to combine the edge-based techniques with the optical flow like [9] which uses a Kalman Filter to merge both cues and is illustrated on Figure 2.4.

Figure 2.4: Left image: original, center: edges, right: optical flow [9].

2.2.2. Combining Optical Flow and Edges

Several authors combine edge and optical flow information to avoid error accumulation. For example, [19] uses a Kalman filter to combine the two cues.

This approach yields impressive results especially because it estimates not only the motion but also the deformations of a face model. Nevertheless, it still depends on the brightness constancy assumption during optical flow computation, and major lighting changes can cause tracking failure. Because of the linearization in the optical flow equation cue, the range of acceptable speeds is also limited.

2.3. Template matching

Template Matching techniques are based on the Lucas-Kanade algorithm, which you can find an excellent review in [1]. Its goal is to find the parameters p of a deformation f that warps a template T into an image i. This deformation can be either a simple affine transformation or even a homography in 3D context as performed in [15] and shown on Figure 2.5. It is interesting because it does not necessarily rely on local features such as edges or interest points, but on global region tracking, using the entire pattern of the object. [1] also shows more developments that have been made over the last two decades regarding this algorithm.

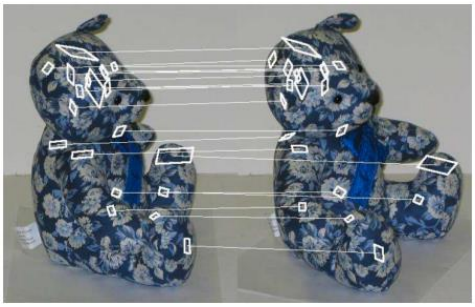


Figure 2.5: 3D application of Template Matching. [15].

2.3.1. 2D Tracking

The general goal of the Lucas-Kanade algorithm is to find the parameters p of some deformation f that warps a template T into the input image I_t, where the f deformation can be a simple affine warp as well as a much more complex one. This is done by minimizing

$$O(p) = \sum_j (I_t\left(f(m_j;p)\right) - T(m_j))^2, \qquad (2)$$

the sum of squared errors computed at several m_i locations.

The Lucas-Kanade algorithm assumes that a current estimate of p is known, which is reasonable for tracking purposes. It iteratively solves for p by computing Δi steps that minimize

$$\sum_j (I_t\left(f(m_j;p_j + \Delta)\right) - T(m_j))^2. \qquad (3)$$

2.3.2. Jacobian Formulation

The general goal of the Lucas-Kanade algorithm is to find the parameters p of some deformation f that warps a template T into the input image I_t, where the f

2.4. Interest Points

Selecting Interest Points from an image is an interesting task. Usually an interest point should have the following properties [7]:

- They should be different from immediate neighbors to avoid edge-points that lead to erroneous matches.
- On repetitive patterns, interest points should be either rejected or given less importance to avoid ambiguous matches.
- The patches surrounding then should be textured so that they can be easily matched.
- The selection should be repetitive in different images.

After selection has been correctly performed, the algorithm looks in a region around the location of the interest point in the previous image, and tries to fit the feature. To perform this, a similarity measure is obtained between an interest point and all its possible matches. The 3D coordinates of the corresponding interesting points in both images can be obtained by back-projecting them to the 3D model. [4] shows how to find the correspondence of 3D points between to images of the same scene using Viewpoint Invariant

Patches and Figure 2.6 shows an example of interest points in a 3D tracking application.

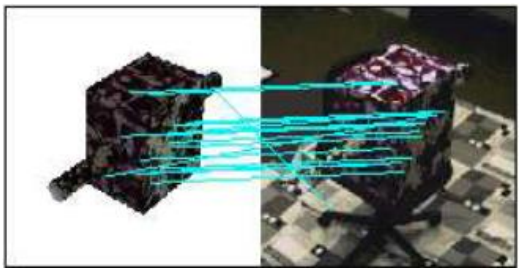


Figure 2.6: Interest Point of textured target. [8]

2.4.1. Pose Estimation by Tracking Planes

An alternative way [26] to use interest points is to track 3D planar structures as opposed to full 3D models. This choice is justified by the fact that it is a common special case that makes the 3D model acquisition problem trivial. For example, in man-made environments such as the one of Figure 4.5, the ground plane and one or more walls are often visible throughout the scene. Furthermore, the resulting method is efficient and precise.

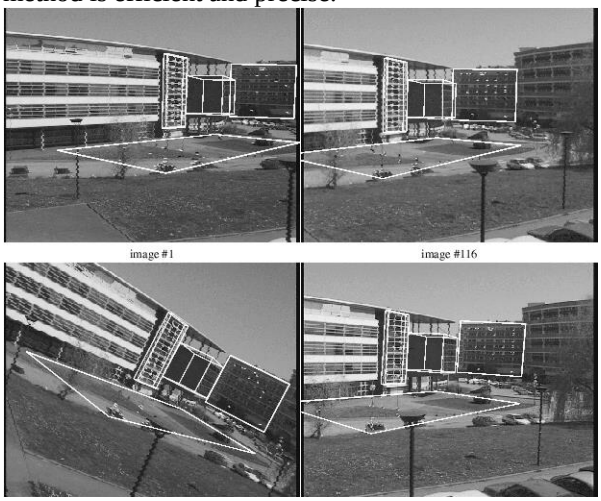


Figure. 2.7 Tracking planar structures using interest points. (From [25], figure courtesy of G.Simon and M.-O. Berger.)

2.5. Tracking Without 3D Models

All the methods presented up to here estimate the pose given an a priori 3D model. However, it is possible to simultaneously estimate both camera motion and scene geometry, without any such model. The recovered trajectory and 3D structure are expressed in an arbitrary coordinate system, for example the one corresponding to the initial camera position. This problem is known as Simultaneous Localization and Mapping (SLAM) by roboticists who are mainly interested in the self-localization of a moving robot. We present here two different classes of approaches that both rely on interest points.

2.5.1. n-Images Methods

The first class of approaches relies on projective properties that provide constraints on camera motion and 3D point locations from 2D correspondences. While such approaches have long been used for offline camera registration in image sequences [23], only recent improvements in both algorithms and available computational power have made them practical for real-time estimation.

For example, [21] shows how to recover in real-time the trajectory of a moving calibrated camera over a long period of time and with very little drift. As direct application of this approach would quickly in drift, two techniques are used in [21] to mitigate this problem. First, the pose is refined once in a while by minimizing the re-projection error of the already reconstructed 3D points over sets of frames. Second, the system is made to occasionally "forget" the 3D coordinates and to re-compute them from scratch to avoid error accumulation.

This system has been tested with a vehicle-mounted camera and yields results very close to that of a GPS, even for trajectories of several hundreds of meters. In other words, error accumulation is not completely avoided, but considerably reduced.

2.5.2. Filter-Based Methods

Pose and structure can also be recursively estimated using the Extended Kalman filter [22, 24]. In particular, [22] shows that it can yield very good results in real-time. While [21] proposes a bottom-up approach – interest points are tracked in 2D then reconstructed to 3D, here the pose estimation is done in a top-down manner. The camera is supposed to move smoothly, with unlikely large accelerations. The filter state therefore contains the camera pose parameters, and the linear and angular velocities used to predict the camera pose over time. The filter state also contains the 3D locations of some points detected in the images. In each coming frame, the position of a feature point is predicted and its uncertainty is estimated using uncertainty propagation, using the 3D location stored in the filter state, the predicted camera pose and its uncertainty. This constraints the search for the point position in the current image, retrieved using sum-of-squared difference correlation. This position is then given to the Kalman filter to update the point 3D

location. These hypotheses are tested in subsequent images by matching them against the images, and their probabilities are re-weighted.

3. Experimental resources

Polygonal Objects For objects that have strong contours and are silhouetted against relatively simple backgrounds, the RAPiD-like methods are a good place to start. They give good results while being fast and relatively simple to implement. By relying on fast optimization techniques and on a fast and reliable top-down feature extraction, it is possible to process images at more 50 Hz on a modern PC. They are also naturally robust to light and scales changes, and specular effects. They run very fast even on older, slower computers. As a result, they have actually been used for visual implementing in industrial environments where reliable edges can be found.

Textured Objects If the target object is textured, such image cue can replace, or complement, the contour information. If the target object is planar and occlusions are unlikely to occur, template-based methods such as those discussed in Subsection 2.3 have been reported to perform very accurately in the presence of an agile motion. This method was reported to take less than 10 ms on a by now old computer (an O2 Silicon Graphics workstation with a 150MHz R5000 processor). It is unfortunately very sensitive to occlusions and hard to extend to fully 3–Dimensional objects.

Methods	Rely on	Suitable when	Manual initialization	Accuracy	Failure modes	3D model required
RAPiD	Edges	Strong edges, simple background	Yes	Can jitter	Very fast motion Cluttered background	Yes
Template matching	Template matching	Small planar object	Yes	Highly accurate	Occlusion Very fast motion	No
Pose Estimation by Tracking Planes	Interest Points	Planar textured objects	Yes	Can drift	Very fast motion	No
Eliminating Drift	Interest Points	3D textured objects	Yes	Avoid drift by using keyframes, reduce iitter	Very fast motion	Yes
n-Images Methods	Interest Points	3D textured scenes	No	Limited drift	Very fast motion	No

Edge-based methods have a fairly low computational complexity because they only work for a small fraction of the image pixels. However, they can become confused in the presence of texture or of a cluttered background. In such cases, area-based methods come into their own and justify their increased computational requirements, which remain quite manageable on modern computers. Nevertheless, optical flow-based methods depend on brightness constancy assumption during optical flow computation, and major lighting changes can cause tracking failure. Because of the linearization in the optical flow equation cue, the range of acceptable speeds is also limited. The template-based approach is attractive because it has low computational requirements, and is simple to implement. But it has some disadvantages. It loses some of its elegance when occlusions must be taken into account, and handling illumination changes requires an offline stage where appearance variations are learned. The class of objects that can be tracked is also limited and a 3D object of general shape under general perspective view have never been handled in this way. All these drawbacks disappear when using a local, feature-based approach. Interest points give information similar to optical flow, but with no need for assumption on the brightness constancy or linearity assumptions. Computers have now become powerful enough to make them practical for real-time applications. As a result, they are now popular and yield the most successful 3D tracking techniques.

4. Conclusion

To conclude this survey, a more generic and desirable approach is therefore to develop purely image-based methods that can detect the target object and compute its 3D pose from a single image. If they are fast enough, they can then be used to initialize and reinitialize the system as often as needed, even if they cannot provide the same accuracy as traditional recursive approaches that use temporal continuity constraints to refine their estimates. Techniques able to do just this are just beginning to come online. And, since they are the last missing part of the puzzle, we expect that we will not have to wait for another twenty

years for purely vision-based commercial systems to become a reality.

5. References

[1] S. Baker and I. Matthews. Lucas-Kanade, "20 Years On: A Unifying Framework" *International Journal of Computer Vision*, 2004, pp. 221–255.

[2] C. Bibby and I. Reid, "Robust Real-Time Visual Tracking using Pixel-Wise Posteriors", *ECCV*, 2008.

[3] T. Chan and L. Vese, "Active contours without edges.", *IEEE TIP*, 2001, pp. 266–277.

[4] B. Clipp, J.-M. Frahm, and M. Pollefeys, "3D model matching with Viewpoint-Invariant Patches (VIP).", *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1-8.

[5] D. Cremers, M. Rousson, and R. D. A, "review of statistical approaches to level set segmentation Integrating color, texture, motion and shape*", International Journal of Computer Vision*, 2007, pp. 195– 215

[6] S. Dambreville, A. Yezzi, M. Niethammer, "A. Tannenbaum, and A. variational framework com- bining level-sets and thresholding.", *BMVC*, 2007, pp 266-280

[7] W. Foerstnet, "A feature-Based Algorithm For Image Matchin", *Archives of Photogrammetry and Remote Sensing*, 1986.

[8] J. Gall, B. Rosenhahn, and H.-p. Seidel, "Robust Pose Estimation with 3D Textured Models.", *Image (Rochester, N.Y.)*, 2006.

[9] M. Haag, "Combination of Edge Element and Optical Flow Estimates for 3D-Model-Based Vehicle Tracking in Traffic Image Sequences.", *International Journal of Computer Vision*, 1999, pp. 295–319,

[10] C. Harris, "Tracking with Rigid Objects.", *Active Vision*, 1992.

[11] G. Klein and D. Murray, "Full-3D Edge Tracking with a Particle Filter.", *Proc. of BMVC*, 2006.

[12] V. Lepetit and P. Fua, "Monocular Model-Based 3D Tracking of Rigid Objects: A Survey", *Com- puter*, 2005, pp. 1–89,

[13] V. A. Prisacariu and I. D. Reid, "PWP3D: Real-time segmentation and tracking of 3D objects", *Proceedings of the 20th British Machine Vision Conference*, 2009, pp. 1–10,

[14] D. Roller, K. Daniilidis, and H. H. Nagel, "Model-based object tracking in monocular image sequences of road traffic scenes" *International Journal of Computer Vision*, 1993, pp. 257–281,

[15] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, "3D Object Modeling and Recognition from Photographs and Image Sequences", *Lecture Notes in Computer Science*, 2006.

[16] C. Wiedemann, M. Ulrich, and C. Steger, "Recognition and Tracking of 3D Objects.", Proceedings of the 30th DAGM, 2008, pp. 132–141.

[17] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision", *International Joint Conference on Artificial Intelligence*, 1981, pp. 674–679.

[18] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 20, no. 2, 2004, pp. 91–110.

[19] M. Haag and H.-M. Nagel, "Combination of edge element and optical flow estimates for 3-D model-based vehicle tracking intraffic image sequences," *International Journal of Computer Vision*, vol. 35, no. 3, 1999, pp. 295–319.

[20] D. DeCarlo and D. Metaxas, "Optical flow constraints on deformable models with applications to face tracking," *International Journal of Computer Vision*, vol. 38, 2000, pp. 99–127.

[21] D. Nister, O. Naroditsky, and J. Bergen, "Visual odometry", *Conferenceon Computer Vision and Pattern Recognition*, June 2004, pp. 652–659.

[22] A. Davison, "Real-time simultaneous localisation and mapping with a single camera,", *Proceedings of International Conference on Computer Vision*, 2003, pp. 1403–1410.

[23] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: A factorization method," *International Journal of Computer Vision*, vol. 9, no. 2, 1992, pp. 137–154.

[24] A. Azarbayejani and A. P. Pentland, "Recursive estimation of motion, structure and focal length," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 6, 1995, pp. 562–575.

[25] G. Simon and M.-O. Berger, "Pose estimation from planar structures*," Computer Graphics and Applications*, vol. 22, 2002m pp. 46–53.

[26] G. Simon, A. Fitzgibbon, and A. Zisserman, "Markerless tracking using planar structures in the scene," , *International Symposium on Mixed and Augmented Reality*, 2000, pp. 120–128.

Feature selection in Intrusion detection datasets

Ugtakhbayar.N, Usukhbayar.B, Ganbayar.U and Nyamjav.J
National University of Mongolia
44911.n@gmail.com

Abstract

In the last few years increasing suspicious connections rapidly in computer network due to internet of things and content based business. In addition, traditional solutions such as firewall, traditional intrusion detection systems, URL blocker, content filter and security gateways are mostly untrustworthy in new type of attacks. So, they cannot throughout all time detecting the whole data and there is a need for decreasing processing data size. Accordingly, proposing an adaptive intrusion detection system which operates in open datasets and efficiently disclose new type of attacks with feature selection method. In this paper, we are executing some feature selection method in open datasets such as Kyoto 2006+ and KDD 99.

Keywords: feature selection, open dataset, data mining

Design and implementation of 32 bit MIPS processor

Battogtokh.J, Batpurev.M, Bold.Z
Department of Electronics and Communication Engineering,
School of Applied Science and Engineering,
National University of Mongolia
Ulaanbaatar, Mongolia
jtogtokh@yahoo.com

Abstract

Energy efficient, space efficient and optimized microcontrollers are the need of the day. We have implemented a modified MIPS architecture that leads to significant power reduction by effectively reducing unwanted clock transitions of various blocks utilizing techniques such as clock gating and stall power reduction.This paper presents the design and development of a high performance and low power MIPS microprocessor and implementation on FPGA. The proposed high performance microprocessor is modeled and verified using FPGA and simulation results. Application of low power methodology to MIPS processor reduces 30% of the total power and makes it more efficient. Though there is 27% increase in area the efficiency of the processor in terms of power makes it an ideal system.The functions of these modules are implemented by pipeline without any interlocks and are simulated successfully on Modelsim 6.3f and Xilinx 12.3i.

The number of non-trivial solutions in Quadratic Sieve

Gantulga.G, Bayarpurev.M, and Garmaa.D
National University of Mongolia,
School of Engineering and Applied Sciences
gantulgag@seas.num.edu.mn

Abstract

Fast and efficient integer factorization algorithms are crucial in analyzing, designing and breaking cryptographic schemes. The well-known quadratic sieve algorithm utilizes non-trivial solutions of the equation $x^2 - y^2 = 0$ *in* $\boldsymbol{Z}_n$*. It uses non-deterministic method which called seiving to construct the non-trivial solutions. Naturally, the running time of the quadratic sieve algorithm for a given integer n depends on the ratio of the cardinality of non-trivial solutions and trivial solutions, i.e.* $x = \pm y$ *in* $\boldsymbol{Z}_n$*. In this paper, we propose an efficient algorithm for counting the non-trivial as well as trivial solutions.*

It's said heuristically that the number of non-trivial solutions are at least half of the solutions satisfying congruent of squares. But we'll show that it's not the case for the numbers in RSA cryptosystem in which the product of two distinct primes are used, by calculating the exact number of trivial and non-trivial solutions using our developed algorithm. It means more relations are needed to ensure success in the case of RSA.

If the number of distinct primes are increased, the number of non-trivial solutions are more than the number of trivial solutions.

Keywords: quadratic sieve, integer factorization, number ring

SDN design for Enterprise Network

Ganbayar.U, Ugtakhbayar.N, Naranbaatar.B, Usukhbayar.B
Mobile and Embedded Technology Research Center
School of Engineering and Applied Science
National University of Mongolia
{Ganbayar, Ugtakhbayar,Usukhbayar}@num.edu.mn

Abstract

Recently, SDN /Software-Defined Networking/ is becoming a state-of-the-art topic among both researchers and organizations because computer networking is being altered significantly by using new architectures and techniques because of demands in which comprise plenty of users, applications, tons of either main devices or mobile, ever-growing video streams and big data in network environment. Besides, Enterprise networks are facing problems, those are demanded financial support, management simplification, powerful security and policy, mobile flexibility, and application aware network. SDN architecture, however, gives us the opportunity to tackle with problems wisely. In our research work, we offer a possibility of how to adapt with new SDN network that is suitable for medium enterprise networks. First, we present concerning main characteristic and benefits of SDN. We then illustrate its implementation on traditional network structure. End goal of our research is to find a possible solution for the future enterprise network with demonstrating it on National University of Mongolia's network.

Key words: *Sofware-defined Networking, Openflow, application layer, controller, infrastructure layer, inbound interface, southbound interface, language-based virtualization, data plane, data path.*

1. Introduction

It is reported that the worldwide network infrastructure will accommodate nearly three networked devices and 15 gigabytes data per capita in 2016, up from over one networked device and 4 gigabytes data per capita in 2011. Such an expansion of network infrastructure would result in an increase in complexity. Due to this factor major problems and limitations of surface in the traditional architecture of Enterprise networks. First, networks are enormous in size with continues growth every day and are highly heterogeneous, especially when equipment**,** apps, and services are provided by different manufacturers, vendors, and providers. Faced with flat or reduced budgets, enterprise IT departments try to squeeze the most from their networks using device-level management tools and manual processes. Second, networks are very complex to manage. The situation could be made worse as legacy network platforms does not have inbuilt programmability, flexibility and support to implement and test new networking ideas without interrupting ongoing services. [1]

To express the desired high-level network policies, network operators need to a very time consuming configuration each individual network device separately, requiring them to work on different CLI's. Automatic reconfiguration and response mechanisms that can solve this problem are virtually non-existent in current IP networks. The control plane (that decides how to handle network traffic) and the data plane (that forwards traffic according to the decisions made by the control plane) are bundled inside the networking devices, reducing flexibility and hindering innovation and evolution of the networking infrastructure. Software-Defined Networking (SDN) is an emerging network architecture that gives hope to change the limitations of current network infrastructures. [2] The key idea of SDN is to decouple the control plane from the data plane and allow flexible and efficient management and operation of the network via software programs. The separation of the control plane and the data plane can be realized by means of a well-defined programming interface between the switches and the SDN controller. The controller exercises direct control over the state in the data plane elements via this well-

defined application programming interface (API). Moreover, SDN allows logical centralization of feedback control with better decisions based on global network view and cross-layer information. The separation of the control plane and the data plane can be realized by means of a well-defined programming interface between the switches and the SDN controller. The controller exercises direct control over the state in the data plane elements via this well-defined application programming interface (API). Moreover, SDN allows logical centralization of feedback control with better decisions based on global network view and cross-layer information. [5]

The fully automated provisioning and orchestration of IT infrastructures has been recently named Software-Defined Environments (SDEs) by IBM.
The four essential building blocks of an SDE are:

- Software-Defined Networks (SDN)
- Software-Defined Storage (SDS)
- Software-Defined Compute (SDC)
- Software-Defined Management (SDM)

Definition: The ongoing selection and use of resources by a server to satisfy client demands according to optimization criteria. Google, for example, has deployed a software-defined network to interconnect its data centers across the globe. This production network has been in deployment for 3 years, helping the company to improve operational efficiency and significantly reduce costs. While commercial SDN deployment started within data-centers and the WAN, the roots of today's SDN arguably go back to the policy management needs and the scalability of enterprise networks. Enterprises stand to benefit from SDN on many different levels, including: (i) network policy can be declared over high-level names and enforced dynamically at fine levels of granularity, (ii) policy can dictate the paths over which traffic is directed, facilitating middle box enforcement and enabling greater network visibility, (iii) policy properties can be verified for correctness, and (iv) policy changes can be accomplished with strong consistency properties, eliminating the chances of transient policy violations. [3]

2. SDN structure

The new SDN architecture defines the topology in 3 layers.

1. Application layer.
2. Control layer.
3. Infrastructure layer.

Each layer has its own specific functions and sub layers. While some of them are always present in an SDN deployment, such as the southbound API, network operating systems, northbound API and network applications, others may be present only in particular deployments, such as hypervisor- or language-based virtualization. [2]

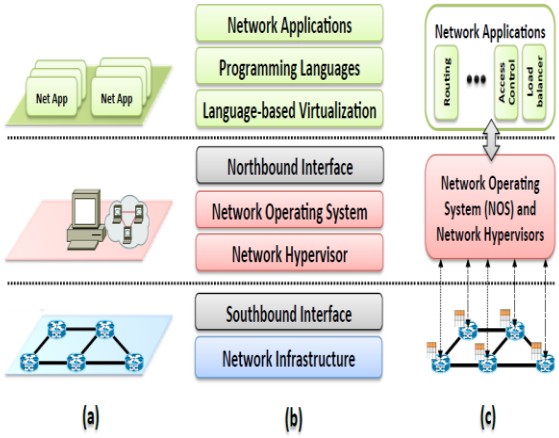


Fig.1. Layers and sublayers.

(a) – Layers.
(b) - Sub layers.
(c) – System design architecture.

1.1 Infrastructure

A SDN infrastructure, similar to a traditional network, consists of a set of network devices (host, switch, router …). The main difference lies in the fact that these network devices function as a forwarding device, rather than making autonomous decisions on their own.

Layer 1. Network Infrastructure

The network intelligence of the network is removed from the devices and gathered at one place to make a logically centralized control system i.e. the SDN Controller. More importantly, these new networks are built (conceptually) on top of open and standard interfaces (such as OpenFlow), a crucial approach for ensuring configuration and communication compatibility and interoperability among different data and control plane devices.

Layer 2. Southbound Interface

The Southbound interfaces are the connecting bridge between the Data plane and the Control plane. The most used protocol for southbound interface is OpenFlow.

Layer 3. Network Hypervisor

Hypervisors enable distinct virtual machines to share the same hardware resources. In a cloud infrastructure-as-a-service (IaaS), each user can have its own virtual resources, from computing to storage. This enabled new revenue and business models where users allocate resources on-demand, from a shared physical infrastructure, at a relatively low cost.

Layer 4. Network Operating System

SDN is promised to facilitate network management and ease the burden of solving networking problems by means of the logically-centralized control offered by a network operating system (NOS). The crucial value of a NOS is to provide abstractions, essential services, and common application programming interfaces (APIs) to developers.

Layer 5. Northbound Interface

The northbound interface is mostly a software ecosystem, not a hardware one as is the case of the southbound. The main purpose of this interface is to interaction between the control plane and the application plane.

Layer 6. Language-based Virtualization

Another state-of-the-art research topic is virtualization. Using virtualization on networks with SDN architecture can open up many new ways to use the network fully.

One form of language-based virtualization is static slicing. This a scheme where the network is sliced by a compiler, based on application layer definitions. The output of the compiler is a monolithic control program that has already slicing definitions and configuration commands for the network. In such a case, there is no need for a hypervisor to dynamically manage the network slices. Static slicing can be valuable for deployments with specific requirements, in particular those where higher performance and simple isolation guarantees are preferable to dynamic slicing.

Layer 7. Programming Languages

One of the defining functions of SDN is allowing the network to be automated by programming. Abstractions provided by high level programming languages can significantly help address many of the challenges of these lower-level instruction sets.

In SDNs, high-level programming languages can be designed and used to:

1) Create higher level abstractions for simplifying the task of programming forwarding devices;

2) enable more productive and problem-focused environments for network software programmers, speeding up development and innovation;

3) promote software modularization and code reusability in the network control plane;

4) Development of network virtualization.

Layer 8. Network Applications

Network applications can be seen as the "network brains". They implement the control-logic that will be translated into commands to be installed in the data plane, dictating the behavior of the forwarding devices. Take a simple application as routing as an example. [2]

3. Controller

Because SDN introduces a new structure of ecosystem with a larger number and more types of devices, an increased focus on such models and mechanisms is required to facilitate interoperability. The SDN Controller is required to be able to exert programmatic direct control of forwarding behavior, thereby actually defining the network (e.g. its logical topology), not merely influencing/configuring it. When deploying multiple instances of SDN Controllers those will need to communicate to distribute work, synchronize state, or otherwise coordinate cooperation, e.g. each SDN Controller instance may be responsible for a portion of the network. [4]

- *Support for Protocol Independent Forwarding and Datapath Programs.*
- *Remain Data Plane Agnostic.*
- *Support for Diverse Hardware / Software Platforms and Datapath Models.* [4] [5]

A SDN controller can have these base services:

- Model-Driven Service Abstraction Layer (MD-SAL)
- Topology Manager
- Statistics Manager
- Switch Manager
- Forwarding Rules Manager
- Host Tracker
- ARP Manager

With an end-to-end view of the network, the SDN controller is also uniquely positioned as a platform where network applications and services can reside. [2] Meaning that the application layer of SDN network can be merged with the controller or can be virtualized on server connection on northbound interfaces of SDN controllers.

4. Related work

Around the world, researchers and organizations are attracted on how to implement SDN on traditional enterprise network like we are intrigued in it. In 2013, ONF introduced "SDN in the Campus Environment" research work. The work was concerning attributes and

challenges of today's campus network and how to adapt SDN in campus network. Moreover, they worked on traffic isolation use case and role of SDN in campus network design. Besides, in Erric Murray's presented solution titled "Challenges in Enterprise Networking and How SDN can help", he demonstrated SDN solution for enterprise network that demonstrated as Kindred HealthCare center. Therefore, plenty of organizations and researchers are working on this challenging topic.

5. Proposed method

We demonstrate our research on the National University of Mongolia's network that refers medium enterprise network.

Table.1. Network components of NUM network.

Building	*In campus*	**13 buildings**
	Remote	**2 locations**
Devices	*Virtual server*	**53 servers**
	Backup server	**2 servers**
	managed devices	**30 devices**
	unmanaged	**453 devices**
	ip camera	**200 cameras**
	Voip phone	**350 phones**
Inbound traffic		**530 Mbps**
Users	*wired /per day/*	**4000 users**
	wireless /per day/	**9000 users**

Critical applications: "SiSi" web application used for students grading system, web server, network management applications in order to push commands, monitor devices, security applications, the backup system in the data center.

Requirements:

- Separate network bandwidth for students, workers, teachers, and guests efficiently and without complex configurations on each devices.
- Isolate services including university grading system, web, network services, data center and wireless network.
- Support high availability, virtualization, security and management.
- Reduce cost in the future
- Require less staff to maintain network
- Fast adoption of new service on existing network
- Demand flexibility management on Big Data.

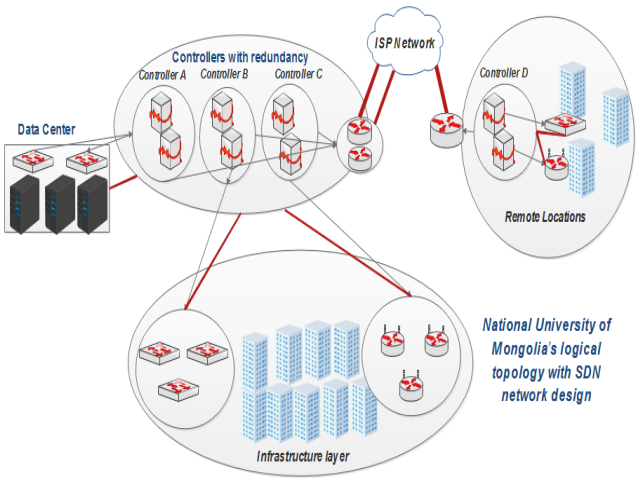


Fig.2. National University of Mongolia's network with SDN in the future.

In the Figure.3, the Controller A /for DataCenter/, Controller B /for Data network/ and Controller C /for wireless network/ are connected each other with redundancy and failover system. Controller D is used for remote locations.

We predict that we will see significant advantages after implementing SDN in the future.

- Network virtualization: Isolate network traffic into various logical flows.
- Improving performance: SDN allows for a centralized control with a global network view and a feedback control with information exchanged between different layers in the network architecture.
- Cost effectiveness: After exchanging into SDN architecture we do not need to change devices fully. Instead we can change interface with higher speed one and controller can do other improvements.
- Security: Dynamic security policies can improve data integrity, availability, and prevent internal and external threats.
- Enhancing configuration: It is the biggest problem for enterprise network that always require static configuration on each new device. SDN gives us the opportunity to configure all devices automatically by using controller.
- Management simplification: Controller with management tool to set policies and provide a more dynamic view of the entire network, no touch level configuration.

- More granular network control with the ability to apply comprehensive and wide-ranging policies at the session, user, device, and application levels

6. Conclusion

In our research, we studied about SDN environment and characteristic, especially focused on how to implement for enterprise network. Moreover, we investigated current issues of National University of Mongolia's network that could be refer medium enterprise network. Based on our study of traditional enterprise network and research of Software-defined networking, we found a solution for future network architecture and regulation. As a consequence, we see significant advantages after implementing SDN for imminent future network. For instance, since we regulate SDN on our network, the performance, security, manageability, control, and cost will be powerful and effective as well as more distinct.

7. Future proposal

Currently, we are motivated in both SDN virtualization and security as well as more accurate and compatible design for enterprise network. We will base on our recent proposal and continue our research with efficient and effective way.

8. Reference

[1] ONF White Paper "Software-Defined Networking: The New Norm for Networks" , 2012

[2] ONF White Paper "Framework for SDN: Scope and Requirements", June 2015

[3] Eric Murray. "Challenges in Enterprise Networking and How SDN Can Help"

[4] Rong Gu, Chen Li, China Mobile. "SDN Controller Requirement"

[5] Diego Kreutz, Member, IEEE, Fernando M. V. Ramos, Member, IEEE, Paulo Verissimo, Fellow, IEEE, Christian Esteve Rothenberg, Member, IEEE, Siamak Azodolmolky, Senior Member, IEEE, and Steve Uhlig, Member, IEEE. "Software-Defined Networking: A Comprehensive Survey", 2015

[6] Dan Levin, Technische Universität Berlin; Marco Canini, Université catholique de Louvain; Stefan Schmid, Technische Universität Berlin and Telekom Innovation Labs; Fabian Schaffert and Anja Feldmann, Technische Universität Berlin. "Reaping the Benefits of Incremental SDN Deployment in Enterprise Networks", April-2013

[7] ONF White Paper. "SDN Architecture"

[8] Wenfeng Xia, Yonggang Wen, *Senior Member, IEEE*, Chuan Heng Foh, *Senior Member, IEEE*, Dusit Niyato, *Member, IEEE*, and Haiyong Xie, *Member, IEEE. "*A Survey on Software-Defined Networking*", 2015*

[9] Keisuke Kuroki, Nobutaka Matsumoto and Michiaki**,** Hayashi. Integrated Core Network Control And Management Laboratory KDDI R&D Laboratories, Inc. Saitama, Japan. "Scalable OpenFlow ControllerRedundancy Tackling Local and Global Recoveries**.**", Afin 2013

[10] Carlos j. Bernardos, Antonio de la oliva, Pablo Serrano, Albert Banchs, Luis m. contreras, hao jin, and Juan Carlos Zúñiga. "An architecture for software defined wireless networking.", June-2014

[11] Dmitry Drutskoy *Elysium Digital*, Eric Keller *University of Colorado*, Jennifer Rexford *Princeton University*. "Scalable Network Virtualization in Software-Defined Networks.", 2013

[12] https://tools.ietf.org/html/rfc7426

[13] http://searchsdn.techtarget.com/definition/SDN-application-software-defined-networking-application

A finite-state morphological transducer for Khalkha Mongolian nominal

Zoljargal Munkhjargal[1], Altangerel Chagnaa[2], Purev Jaimai[3], Nanzadragchaa Dambasuren[4]
Department of Information and Computer Science, National University of Mongolia
{zoljargal, altangerel, purev, nanzadragchaa}@num.edu.mn

Abstract

Our research describes the development of a finite-state morphological transducer for Khalkha Mongolian nominal. The transducer has been developed for morphological generation for use within a search engines to enrich query, but has also been extensively tested for analysis. The finite-state toolkit used for the work was the Helsinki Finite-State Toolkit (HFST). The poster describes some issues in Khalkha Mongolian morphology, the development of the tool, some linguistic issues encountered and how they were dealt with, and which issues are left to resolve. We directly used 8340 head words and 47 morphological classification (29 of nominal, 18 of verb) from Ts.Damdinsuren's Mongol Usgiin Durmiin Toli (Mongolian Morphological Dictionary). The two-level rules we have described deal with vowel harmony, vowel drop, insertion of vowel/consonant and soft signs. An evaluation is presented which shows that the transducer has medium-level coverage.

Improving the result of the model for predicting the class fault proneness using data mining anomaly detection techniques

Batnyam Battulga[1], Lkhamrolom Tsoodol[1], Erdenetuya Namsrai[2], Purev Jaimai[1]
School of Engineering and Applied Sciences
National University of Mongolia[1]
School of Business Administration and Humanities,
Mongolian University of Science and Technology[2]
(batnyam, lhamrolom, oyunerdene, purev)@seas.num.edu.mn

Abstract

Software maintenance cost is more the 75 percent of total cost of system development life cycle and software maintenance is modification process of a software product after delivery to correct faults, to improve performance or other attributes or to adapt the product to modified environment. From this it is possible to decrease software maintenance cost by developing faultless software and predict fault structure using model design when design phase of SDLC.

The main purpose of this research is to improving the result of the model for predicting fault proneness of the class. To achieve this we doing experiments using anomaly detection technique before predict the class fault proneness.

Experimental result shows that most suitable method is combination of Resampling and Correlation Based Subset Selection algorithm and Random Forest. That approach is predicted the class fault proneness approximately 90%.

Keywords: OOD metrics, class fault proneness, data mining technique, classification technique, anomaly detection technique

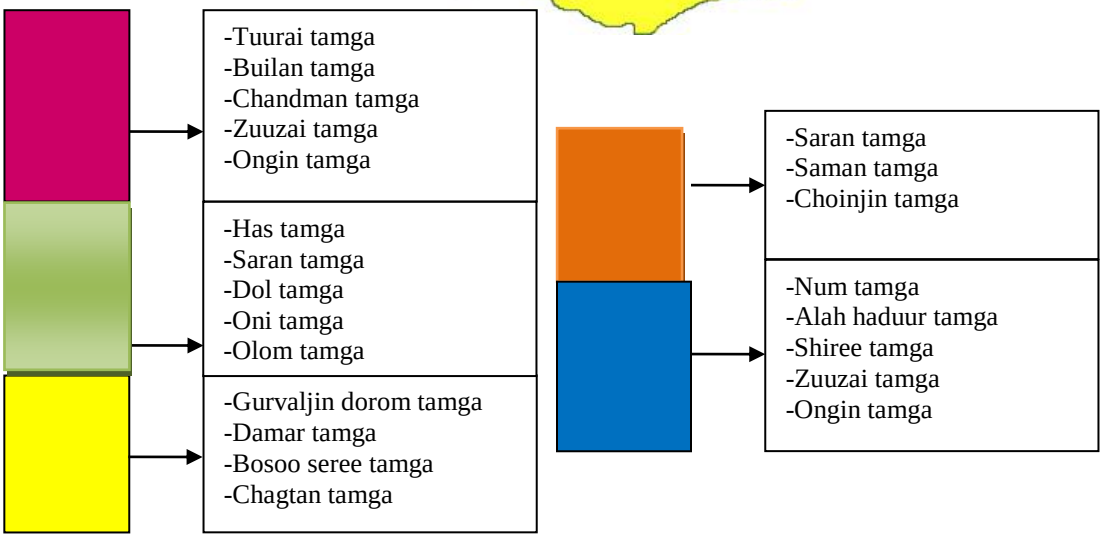
Modern Trend of Mongolian Horse Stamp

Gantuya Perenleilkhundev, Shaariibuu Setev, Suvdaa Batsuuri*
School of Engineering and Applied Sciences, National University of Mongolia
gantuya@seas.num.edu.mn, {shaariibuu, suvdaa}@num.edu.mn

Abstract

Researchers are studied that Mongolian horse traditional stamp means property of the herdsmen. Therefore the stamps depend on the geographical distribution. Also the shape and symbolism of the stamp are related to archelogical findings and they are meaningful historically. In this study, we find out that same shape stamps are not related to same geographical location. Stamp includes more than 2 shapes, that a basic shape and additional shapes. Therefore we compared the collected stamps by its basic shape with geographical location, additional shapes with geographical and other means. As a result, we conclude that new stamp can be created basic shapes by geographical region and then add other shapes by owners interesting and family valuable symbols. For example, nowadays some horse's stamp includes Toyota, lexus logo etc.

Keywords: Horse, stamp



Differential wheeled mobile robot real time self-localization and path planning method for microcontroller

Bold Zagd
Batbayar Unursaikhan

Abstract

Purpose of this paper is simplifying real time self localization and path planning algorithm of differential driving indoor mobile robot in order to make robot's function smarter. Using the rotary encoder is one of the reliable, cheap way in terms of room condition. The robot movement can be more efficient for energy, speed and their mechanical parts if it smoothly. Most of surface paths of mobile machines are based on Bezier curve. But this solution requires solving the quadratic or cube equation. And most of algorithms for a self localization are based on a fully floating point operation and trigonometry functions. Big equation is not suitable for real time operation of moving robots. Because of these reasons high-performance controllers should be applied. We propose self-localization calculating and path planning algorithm for low-performance microcontroller without using operations mentioned above. The self-localization algorithm fails more in long distance. Experimental results for the simulation presented.

Key words: Differential wheeled mobile robot, self localization, localization, autonomous robot, path planning, trajectory estimating

Self-tuning PID controller for dynamic systems

Battur Ganbat, Lodoiravsal Choimaa
Batbayar Unursaikhan

Abstract

Proportional-Integral-Derivative (PID) controllers have utilized most commonly in the control of DC motors industrial automation. In this article, we have been presented auto-tuning method for PID controller. PID controller has a problem which is overflow from inertia of mechanical process when change set point value. Because of standard PID controller's K_p , K_i and K_d values are constant. Therefore, we developed adaptive PID control design improved by Ziegler-Nichols tuning rule. PID controller gain parameters supplied to responded motor with variable load when increase and decrease DC motor shaft speed. The self-tuning PID's values are varied with the shaft speed variations. Experimental results for the simulation presented.

Key words: PID controller, self-tuning, adaptive PID controller

An Improved Medical Decision Support System for Predicting the Stages of Chronic Obstructive Pulmonary Disease

Solongo Khurts[1], Nasantuya Namsrai[2], Erdenetuya Namsrai[3], Otgonnaran Ochirbat[4]
Health Sciences University of Mongolia[1]
General Hospital of Defense and Law Enforcement[2]
Mongolian University of Science and Technology[3]
National University of Mongolia[4]
{*sun_solongo, nnasantuya*}*@yahoo.com, otgonnaran@seas.num.edu.mn*

Abstract

Air pollution has major health impacts on people living in Ulaanbaatar. As written in the WORLD BANK report: Ambient annual average particulate matter concentrations in the capital of Mongolia are 10–25 times greater than Mongolian air quality standards and are among the highest recorded measurements in any world capital. Chronic obstructive pulmonary disease (COPD) induced by air pollution and smoking was found to be a major cause of illness in Mongolia. For a medical doctor, diagnosing the condition of Chronic Obstructive Pulmonary Disease and starting treatment in the lower stages is very crucial. Therefore, software system for assisting the doctor for determining the medical condition of a patient is required. In this paper, we have demonstrated the possibility of predicting the stages of COPD patients using classification techniques. To improve classification result, we have applied anomaly detection method to eliminate anomalies in a data preprocessing stage.

The findings of the present study suggest that antioxidant capacity reflected by COX and the lipid peroxidation products MDA in erythrocyte's membrane are linked to the severity of COPD.
Body composition is an important non-pulmonary impairment that modulates the risk of functional limitation in COPD, even after taking pulmonary function into account. Body composition abnormalities may represent an important area for screening and intervention in COPD.

Key words: Chronic obstructive pulmonary disease, free radical scavenging activity, cytochrome c oxidase, Lipid peroxides products, Malondialdehyde, anomaly detection, classification, ID3, decision support system, data mining

Land Management System with Instant Area Estimator

Oktyabar Enkhtaivan[1], Nasanbat Namsrai[2], Oyun-Erdene Namsrai[1]
National University of Mongolia[1]
Mongolian University of Science and Technology[2]
{oktyabar,oyunerdene@seas.num.edu.mn, nasanbat@gmail.com}

Abstract

A geographic information system (GIS) is a system designed to capture, store, manipulate, analyze, manage, and present all types of spatial or geographical data. In this research work, we have implemented fractional calculation based land management GIS system. The main goal of this information system is to estimate user selected area instantly when user requires to know exact size of any part of their land; software system should fulfill that requirement.

We have developed geographical land management system using Dotspatial data. Although the software ArcGIS 9,3 used at the Metropolitan Land Department, is fitted to requirements, the study proved that it is expensive, difficult to install and its application is time consuming. The Land management exemplary software, we have developed during this research work, shows that these problems can be solved easily.

Keywords: Fractional Calculation, Land Management system, GIS

Virtual lab management using Citrix

Ankhzaya Jamsrandorj, Sodbileg Shirmen
Department of Electronics and Communication Engineering,
School of Engineering and Applied Sciences,
National University of Mongolia,
Ulaanbaatar, Mongolia
ankhzaya@seas.num.edu.mn; sodbileg@seas.num.edu.mn

Abstract

Virtual labs provide to access remote users into centralized resources and to solve complex calculus using any portable or desktop devices anywhere on the Internet. Recently, virtual labs have been become crucial IT infrastructure at higher education. This paper reports comparison result between virtual labs and implementation method of the virtual lab management at the National University of Mongolia using Citrix. The report is intended to help faculties and administrators considering a similar implementation.

Keywords: network virtualization, remote lab, virtual lab, e-learning

1. INTRODUCTION

The Virtualization is a proven software technology that makes it possible to run multiple operating systems and applications on the same server at the same time. It's transforming the IT landscape and fundamentally changing the way that people utilize technology. [3] Virtual lab is one of the implementations of it at higher education organizations.

The Virtual lab is allowed to use application software which is installed on remote lab computers or servers managed by server software remotely. It is useful when students need to use an application software which is not installed on their home computer. The applications are not actually downloaded to student's home PC. Computing takes place on remote virtual servers; however, from their perspective, it appears as though students are running the applications on their own PC.

According to this paper we did some experiments related with hardware requirements of the virtual lab depends on loads of application software and number of clients. We created Citrix cluster using personal computers as servers and installed application software for computer networking courses. While students and staffs were using it for their daily jobs, the cluster was monitored and measured by performance on the results.

2. METHODOLOGY

This subsection explains about the test-bed deployed to collect the experiment results, the methodology used to build the virtual lab and its performance management.

2.1. Test-bed

Mongolia's largest and most prestigious university, the National University of Mongolia (NUM) has about 20'000 students and 1,500 full-time professors. Also NUM has 8 buildings and each of these buildings has at least 5-8 computer or advanced technology laboratories. Currently NUM has no virtual labs.

2.2. Citrix

We used XenServer, XenApp, XenDesktop and XenReceiver of Citrix systems to build Virtual lab at NUM.

Citirix XenServer is a hypervisor platform that enables the creation and management of virtualized server infrastructure. It is developed by Citrix Systems and is built over the Xen virtual machine hypervisor. Citrix XenServer manages the allocation and distribution of physical server computing resources among virtual machines and administers their performance and use. [1]

Citrix application virtualization technology isolates applications from the underlying operating system and

from other applications to increase compatibility and manageability. As a modern application delivery solution, XenApp virtualizes applications via integrated application streaming and isolation technology. This application virtualization technology enables applications to be streamed from a centralized location into an isolation environment on the target device where they will execute. [1]

Citrix XenDesktop is a suite of desktop virtualization products from software provider Citrix Systems. Also Citrix Receiver is the easy-to-install client software that provides access to XenDesktop and XenApp installations. [1]

2.3. Building virtual lab using Citrix at NUM

As we mentioned before, we used Citrix XenServer 7.8 to build a server on general purpose PC.

Computer that installed virtual server: 64-bit×86 Intel Core i5 3.1GHz CPU, 8GB RAM, 500GB SATA storage disk and 100Mbps network interface card.

Current virtual lab programs: Microsoft Office, MATLAB, Notepad++, Packet tracer, GNS3 and VMware.

Figure 1 shows graphical user interface for managing Citrix XenServer at NUM.

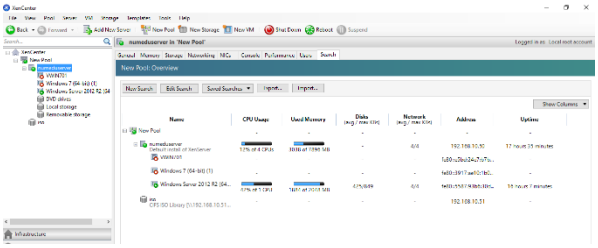


Figure 13. Virtual lab server at NUM

3. RESULT

Currently most of the NUM's laboratories have capacity of 15-20 students. Thus, we selected the number of users accessing server between those numbers at same time.

Students have to download and install the "Citrix Receiver" and they login to virtual lab using their own user ID and password. They also visit the https://vlabs.num.edu.mn to direct access the virtual lab.

CPU performance of the server during 5 students running Packet tracer software is shown in Figure 2.

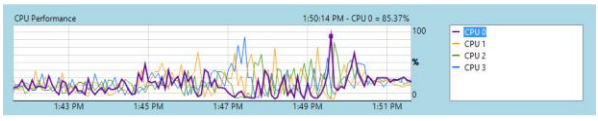


Figure 14. CPU performance of the Virtual Server

This indicates that CPU usage is full while 5 users access to virtual lab and student's operation is becoming slow.

Figure 3 shows memory usage during that span of time.

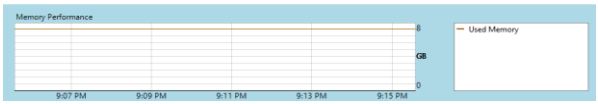


Figure 15. Memory usage of the Virtual Server

We created just 3 virtual machine on our virtual server: including controlling system of desktop application virtualization, our master image, and operating system that users can access. However, it required more memory capacity to launch all three machines. So we shut down our virtual machine that master image works on it.

Figure 4 shows network usage of the server. Since network connection is 100M, network usage is being relatively high.

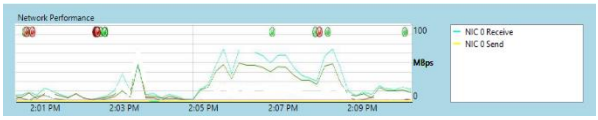


Figure 16. Network usage of the Virtual Server

4. CONCLUSION

In this work, we created a virtual laboratory and organized its management. Since we used normal PC as a server, performance of the server was not good.

Firstly, when multiple users access the server, server was quickly run out of resources.

Secondly, Due to insufficient speed of network interface, server was becoming slower when multiple users access the server.

Thirdly, Because of these issues, users were not able to work flawlessly on the server.

Even though, the performance of the system was poor, the remote laboratory is still useful and needed for the university.

In future works, we are planning to replace server computer as a proper powerful server PC.

5. References

[1] http://citrix.com

[2] N. M. Boutaba, *A survey of network virtualization*, 2009.

[3] Fernando Terroso S´aenz, R. F.-P. (2009). *Virtualization technologies: An overview*.
[4] J. E. Nair,. *Virtual Machines. Versatile plataforms for systems and processes*, 2005.

[5] P. Barham, B. D. , *Xen and the Art of Virtualization*, 2003.

[6] C. A. Salem, *Virtualization and Databases*, 2000.

[7] D.Shackleford, *Virtualization Security*. 2000.

[8] Solutions, D. Power, *Virtualization technologies*, 2005.

[9] Technical white paper. *Best practices for deploying Citrix XenApp on XenServer on HP ProLiant servers*, 2010.
[10] The Xen project. *Xen Architecture Overview*, 2008.

Building OpenWRT Embedded Linux in Atheros

Ankhzaya Jamsrandorj; Sodbileg Shirmen
Department of Electronics and Communication Engineering,
School of Engineering and Applied Sciences,
National University of Mongolia,
Ulaanbaatar, Mongolia
ankhzaya@seas.num.edu.mn; sodbileg@seas.num.edu.mn

Abstract

Most of the future digital services for home and office users will be deployed and delivered through the wireless connectivity. People use wireless network because of the added convenience and productivity to tasks. As the technology emerges, more wireless network devices are being designed and developed. OpenWRT is open source and the GNU/Linux distribution for embedded systems. This paper illustrates the architecture and implementation of a wireless router that used OpenWRT and how to build the system on device with Atheros architecture. The end goal of our research is to build integrated devices in Mongolian.

Keywords - OpenWRT; Wireless router; Embedded system.

An Augmented Reality Integrated Pseudo-3D Map and Optical Tracking Application

Phuong Tien Nguyen, Tung Duong Vu, Hue Thi Le
Institute of Information Technology, Vietnam Accademy of Science and Technology
phuongnt@ioit.ac.vn

Abstract

Augmented reality is a live direct or indirect view of a physical, real-world environment whose elements are augmented by computer-generated sensory input such as sound, video, graphics or GPS data. Information about the environment and its objects is overlaid on the real world. Pseudo-3D is term used to describe either 2D graphical projections and similar techniques used to cause a series of images to simulate the appearance of being three-dimensional when in fact they are not.

The integrated the pseudo-3D map in augmented reality system will bring the best experiences for users. This paper presents some our research results of augmented reality and pseudo-3D map. The involved technical issues as perspective projection, near-far display of additional information, File Tuning, Parallel Tracking And Mapping are also research to bring more "real" effect for the users. An augmented reality integrated pseudo-3D map and optical tracking application are also introduced in this paper.

Keywords: 3D, Pseudo-3D, Augmented Reality, GPS, LBS

1. Introduction

Augmented reality (AR) is a live direct or indirect view of a physical, real-world environment whose elements are augmented by computer-generated sensory input such as sound, video, graphics or GPS data. As a result, the technology functions by enhancing one's current perception of reality. Augmentation is conventionally in real-time and in semantic context with environmental elements. Unlike virtual reality, which creates a totally artificial environment, augmented reality uses the existing environment and overlays new information on top of it. With the help of advanced AR technology the information about the surrounding real world of the user becomes interactive and digitally manipulable. Information about the environment and its objects is overlaid on the real world. Pseudo-3D (2.5D) is term used to describe either 2D graphical projections and similar techniques used to cause a series of images to simulate the appearance of being three-dimensional (3D) when in fact they are not.

The integrated the pseudo-3D map in augmented reality system will bring the best experiences for users. This paper presents some our research results of augmented reality and pseudo-3D map. The technical issues involved as perspective projection, near-far display of additional information, File Tuning, Parallel Tracking And Mapping (PTAM) are also research to bring more "real" effect for the users. This paper is organized as follows. Section 2 briefly reviews the augmented reality and pseudo-3D map. Section 3 proposes the model of the augmented reality integrated pseudo-3D map and optical tracking application and also some technical issues and results. Finally, section 4 summarizes and concludes this paper.

2. Augmented Reality and Pseudo-3D Mapping

2.1. Augmented Reality

There are many definitions about Augmented Reality, but overall it is a technology that the virtual objects can be part of real world. Some definitions of Augmented Reality emphasized that virtual objects are 3-D models, but most of the current definition is accepted by AR virtual object is 2-dimensional models such as text, logos, images... enhanced reality is the interface between the real world and the virtual world. It is illustrated by the Paul Milgram's Reality-Virtuality Continuum diagram as shown below [1].

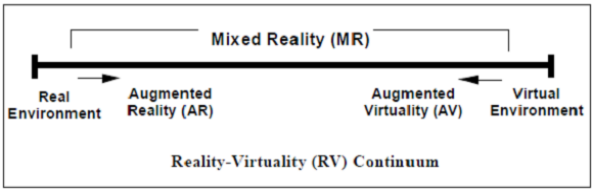


Figure 17. Reality-Virtuality (RV) Continuum (Paul Milgram)

Augmented Reality technology use the object in real world (collecting from GPS, camera, micro…). After processed, will display and provide more additional information about this object. The additional information are descriptions, images, sounds, or animations... In Augmented Reality, real objects can be displayed in 2D or 3D.

Virtual Reality technology can be regarded as the predecessor of Augmented Reality technology with many similarities. The biggest difference is Virtual Reality do not use data from camera directly, all data in Virtual Reality are photos, videos... that recording before.

2.1.1. Augmented Reality System. There are three main components of Augmented Reality system, showing in figure 2 below [2]:

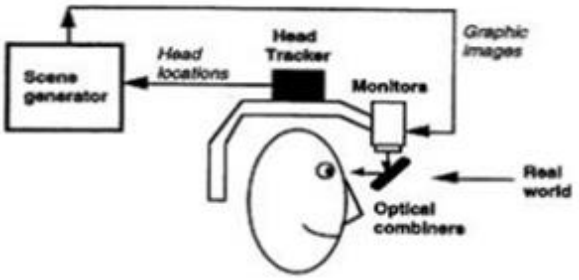


Figure 18. Augmented Reality System

Scene Genegator

The scene generator is the device or software responsible for rendering the scene. Rendering is not currently one of the major problems in AR, because a few virtual objects need to be drawn, and they often do not necessarily have to be realistically rendered in order to serve the purposes of the application.

Tracking System

The tracking system is one of the most important problems on AR systems mostly. The objects in the real and virtual worlds must be properly aligned with respect to each other, or the illusion that the two worlds coexist will be compromised.

There are two tracking object techniques before overlaying them into display screen:

- Object tracking is a techniques using location based tracking (GPS), digital compass, accelerometers.
- Optical tracking is a techniques using image identification.

We can combine both techniques to get better efficacy in many cases.

Display

Most of the Displays devices for AR are HMD (Head Mounted Display), but other solutions can be found (screen of smart phone, VR device…). When combining the real and virtual world two basic choices are available: optical and video technology. Each of them has its own advantages and disadvantages depending on the factors as resolution, flexibility, field-of-view, registration strategies...

Display technology continues to be a limiting factor in the development of AR systems. There are still no see-through displays that have sufficient brightness, resolution, field of view, and contrast to seamlessly blend a wide range of real and virtual imagery. Furthermore, many technologies that begin to approach these goals are not yet sufficiently small, lightweight, and low-cost. Nevertheless, the past few years we have seen a number of advances in see-through display technology, as discussed below.

2.1.2. Augmented Reality in the World. Augmented Reality technology is using in life over the world [4]. The most applications are entertainment, movies, video games (Oculus Rift) or simulations. The user can be experience the video game with storyline from the real world and can be interact with the objects that found in game, and the player can be moved, mobilize during gameplay. In advertising, media, the viewer can be see virtual advertising billboard appear immediately on the stadium, when the match was going on. Using Augmented Reality technology will increase the effectiveness of advertising. Microsoft developed HoloLens to display 3D object through Windows Holographic technology. When we wear HoloLens, the real world was change to virtual world with many 3D objects.

Figure 19. Augmented Reality application (Microsoft HoloLens)

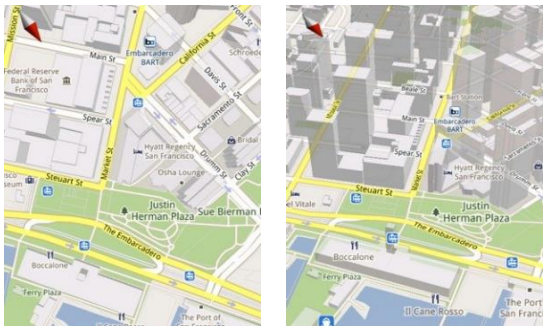
Nowadays, smartphones are growing very fast. Application that using Augmented Reality technology is workable when smartphone have many sensors: GPS, accelerometer, gyroscope… So we can get the position, movement direction and the tilt of smartphone [3, 5]. With the growing of smartphone's

hardware: touch HD screen, cpu, gpu, ram... Foreseeing this trend, vendors of mobile software in the world were soon exploiting enhanced reality technology in the location-based services (LBS-Location Based Services). Previously, when user looking for convenient around a position, user must used pure map application like Google Maps, Nokia Here... but now, user only need to slide the camera and find convenience that display on screen. E.g.: Junaio on Android, Here City Lens on Windows mobile.

Applying augmented reality technology in smart phone application bring a lot of convenience to users. With the vendor, some problems arise: the location data of Google Maps or Bing Maps are very big, but depending the areas, there are not enough to users. That is a reason that we can develop an application similar like Google Maps or Bing Maps and providing a local dataset in a specific area. In the next session, we will introduce about technology of pseudo-3D (2.5D) map and integration capabilities its in augmented reality application on smartphone to enhance experience of users.

2.2. The pseudo-3D map

The pseudo-3D map (or 2.5D map) is term to describe the 2D map using the 2D projections and other similar techniques to simulate 3D objects on map. Different from standard 2D map have the angle of vision is 90º (figure 4a), by changing the angle (figure 4b), the map was more realistic. Beside that, the pseudo 3D map have an advantage than 3D map that do not require high performance of hardware. The pseudo 3D map is the most basic technique, using the projection to change angle of vision while observing the map [7,8].



a. 90º *b. 45º*

Figure 20. Map with different views

When developing augmented reality application, combining the real images from camera and the pseudo-3D map will providing more information and increasing experiences of users. The figure below illustrates an enhanced reality applications is projected onto the windscreen in the car to give the driver information about the upcoming route an extremely intuitive way to add additional useful information.

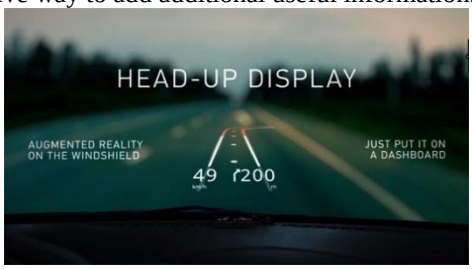


Figure 21. Combination with pseudo-3D map and augmented reality technology

Displaying map under different angles will bring "real" feeling to users. While combining pseudo-3D map in augmented reality application, users can get more information than the reality.

3. Develop an Augmented Reality integrated pseudo-3D map application

In this session, we will introduce the pseudo-3D map development model and some techniques as perspective projection, File Tuning, Parallel Tracking And Mapping are also researched to bring more "real" effect for the users techniques. In the end of session, the results and evaluations are presented.

3.1. Model for Developing the Pseudo-3D Map's Application

Combining the pseudo-3D map and map with Google base layer in web platform [6] required many processing steps. The figure below describes a diagram of developing the pseudo-3D map, combining Google Map API and Google Earth API:

- Google Map API is responsible for providing base map services.
- Google Earth API is used to providing 3D object data.

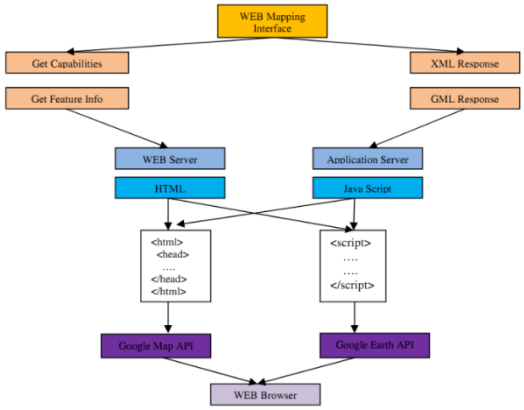


Figure 22. Diagram of develop the pseudo 3D map's application on Web browser

The diagram of developing the mobile application using the pseudo 3D map is described as figure 7 below:

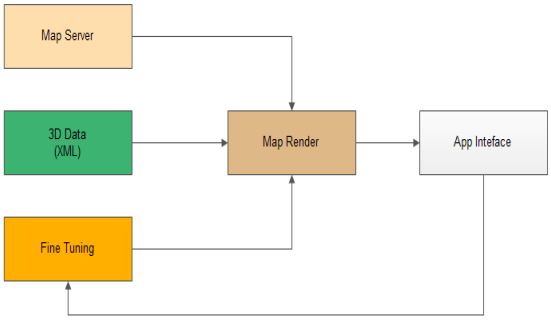


Figure 23. Diagram of develop the pseudo 3D map's application on cellphone

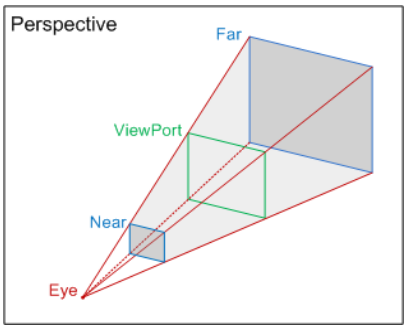
Map Rendering component is used to:

- Receiving and processing map data from Map Server. Map Server is location base services system through web services.
- Processing 3D data in XML format.
- Based on location information (position, direction, inclination), using the project to calibrating display (angle, distance)

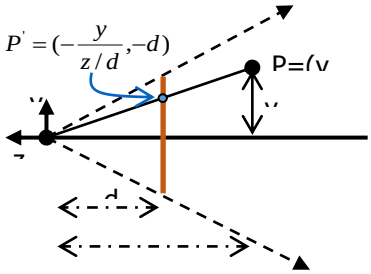
The application interface display the real world through device's camera, the pseudo-3D map data will be overlay the screen, providing information to users. While device have updating of location, direction, the application will calibrate the map and updating on device's screen.

3.2. Fine Tunning

Displaying object in the pseudo-3D map have different with displaying object in the standard 2D map. Depending on the location of the user, the angle of the device and the user's perspective, we need to continuously calibrate the display of objects on the screen including near, far using projection perspective and calculate the position displayed on the camera screen (Figure 8).



a. Perspective projection



b. Calculate display position

Figure 24. Perspective projection and near-far display

The formula for calculating the coordinates of the point P' of P in perspective projection:

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1/d & 1 \end{bmatrix}$$

$$P' = \begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -\frac{1}{d} & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} x/(-z/d) \\ y/(-z/d) \\ -d \\ 1 \end{bmatrix}$$

Where, M is transformed matrix.

3.3. Parallel Tracking and Mapping

To improve experiences of using, the augmented reality application need capability of object identification on the real world, this capability will provide more information for user when observer the real world. This is a part of augmented reality identification system, using optical tracking. There are

some techniques to do this like using SLAM (Simultaneous Localization And Mapping) [9] with Kalman filter and Particle filter or using PTAM (Parallel Tracking And Mapping) [10]. In this paper, we are using PTAM when developing the application.

Parallel Tracking and Mapping Algorithm. PTAM is a part of study that monitoring the parallel-mapped for cell phone. The PTAM trackers are used to tracking the position of 3D camera in real time. PTAM be applied in tasks such as guiding robot is the main, but it is also useful for augmented reality. PTAM originally developed as a research system in the New Vision Laboratory of the University of Oxford [10]. In 2007 this system was first introduced at ISMAR. PTAM was released in 2008 by Isis under a license suitable for use in academic and commercial.

PTAM Components

PTAM has 5 main components:

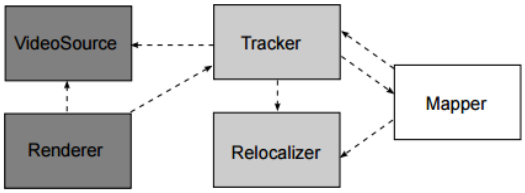


Figure 25. PTAM: video capture and image render are the different components: VideoSource and Renderer

The monitoring is collected from Tracker, using Relocalizer when lost (grey). Using Mapper to create, expand, and improve (white).

Video Source is used to loading video frames from camera. These frames are analyzed by Tracker and overlay an augmented reality layer by Renderer.

Renderer: Each camera frame is rendered on the screen at an overlay of 3D objects. The 3D objects are organized according to the camera position given by Tracker through the device's sensor.

Tracker: finding the point mapped from video frames, these frames are sending to Mapper.

Relocalizer: when do not find map points enough from video frame, Tracker called Relocalizer, estimate camera locations using the image approaches.

Mapper received video frames from Tracker. Used to creating and expanding mapped. Mapper is used to calibrate points on the map.

Figure 26. Augmented Reality application monitoring point specifies in video frames (right) to activate the coating 3D objects (left)

In Figure 10, the right is a video frame displayed gray with the spectific points being monitored, from which the camera position is estimated. The loeft picture shows the coating results with a 3D object and a white border around a book is detected.

3.4. Some Results

Based on studies of augmented reality and the pseudo-3D map, we developed the demo system using the pseudo-3D map and augmented reality technology with regional in scope VAST at 18 Hoang Quoc Viet street, Cau Giay District, Hanoi, Vietnam.

Figure 27. The camera with more information while viewing around

The base layer and data are using Google's platform. 3D object and POIs (point of interest) data were built and added to the application (Figure 12).

Figure 28. Shape data in the pseudo-3D map and related information in Augmented Reality Application

This application also allows to identification objects in the real world and providing more related information. In figure 13, when user slide device's camera to object (e.g. book), device's screen will display more related information of the object. This has many advantages in fact, such as used in supermarket to find information of products.

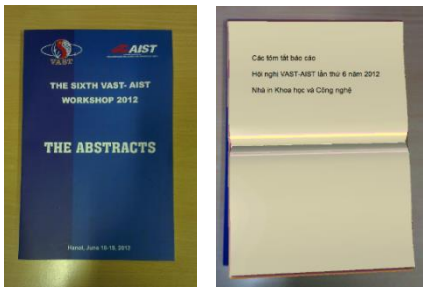


Figure 29. Camera object identification in real world (left) to provide more information (right)

In short, the integration of the pseudo-3D map into augmented reality application will increase user experiences. Users can interact with 3D objects in application. However, to getting the effect of display and interact, the system must have a large of simulation objects (buildings, trees,..) and GIS related data (lat, lon,...). The construction of a data source this simulation requires much time and effort. Besides, the use of pseudo-3D map will set higher requirements than the base map using conventional 2D performance user equipment, especially mobile devices to be able to ensuring that the user experience the best efficiency.

4. Conclusion

The paper presents some results of studies on augmented reality, the pseudo-3D map and and optical tracking. The integrated the pseudo-3D map and and optical tracking in augmented reality system will bring more "real" experiences for users. We have developed a demo system that was testing in a specific area (VAST). It has a great potential in transport and tourist. Some issues still need to be resolved as built 3D data sets, research the effects algorithms perform overlay map on the background of reality, enhancing the ability to recognize objects...

Acknowledgment

This research has been funded by the Research Project, VAST01.04/14-15, Vietnam Academy of Science and Technology.

5. References

[1] Milgram, Paul; H. Takemura, A. Utsumi, F. Kishino, "Augmented Reality: A class of displays on the reality-virtuality continuum", *Proc. of Telemanipulator and Telepresence Technologies*, vol. 2351, pp. 282-292

[2] Rolf R. Hainich, "Augmented Reality and Beyond", *The End of Hardware*, 3rd Edition, 2009

[3] Frank Ableson, Charlie Collins, Robi Sen, "Unlocking Android – A Developer's Guide", *Dick Wall*

[4] Raghav Sood, "Pro Android Augmented Reality", *Apress*

[5] Ben Butchart, "Augmented Reality for Smartphones - A Guide for developers and content publishers", *TechWatch Report*

[6] Fatih S., Hakan K., "3D Gis Application by Implementing 3D City Model Google Earth and Google Map Integration"

[7] Marcus Apel, "From 3d geomodelling systems towards 3d geoscience information systems: Data model, query functionality, and data management"

[8] A. Abdul Rahman, M. Pilouk, "Spatial Data Modelling for 3D GIS"

[9] Georg Klein và David Murray, "Improving the Agility of Keyframe-based SLAM", *Proc. ECCV*, 2008

[10] Georg Klein và David Murray, "Parallel Tracking and Mapping for Small AR Workspaces", *Proc. ISMAR*, 20

Listener's Preference Based Bayesian Learning for Recommendation in Music Site

Young Sung Cho[1], Song Chul Moon[2], Seon-Phil Jeong[3] , Keun Ho Ryu[*]
[1,2,*)]*Database and Bioinformatics Laboratory, School of Electrical and Computer Engineering, Chungbuk National University, South Korea*
[3.]*Division of Science and Technology, BNU-HKBU United International College, Zhuhai, China*
[1]*youngscho@empal.com,* [2]*moon@nsu.ac.kr,* [3]*spjeong@uic.edu.hk,*
[*]*khryu@dblab.chungbuk.ac.kr*

Abstract

Along with the recent growth in the digital music industry and increasing the demand of mobile music, the number of smart phone users are increasing to listen to music based on music site under ubiquitous computing environment. The selection criteria of listener's preferred music has gotten more diverse and complicated as the range of popular music has gotten wider. These days, research to find intelligent methods to customized recommend music on listener preferences under the digital music environment is actively being conducted. However, existing music recommendation systems do not reflect listeners' preferences due to recommendations simply employing listeners' listening log. In this paper, we propose music recommender system via learning listeners΄prefererece based on music sites, such as Melon, Billboard, Bugs Music, Soribada, and Gini, with most popular current songs across all genres and styles. It is also necessary for us to make the task of calculating the preference with weight to reflect the preference of most popular current songs with its popular music charts on trends. We evaluated the proposed system on the data set of music sites to measure its performance. We reported some of the experimental result, which is better performance than the previous system.

Keywords: BN, Clustering

1. Introduction

These days, listeners have increasingly preferred to digital real-time streamlining and downloading to listen to music because this is convenient and affordable for the listeners. The online digital music has become a new communication channel to listen musics, where digital files can be delivered over various online networks to people's computing devices. Then, they can enjoy listening to great music from these free online music streaming sites for listening to Free Music. The demands for music portal sites and many different digital music pieces on music portal site are increasing rapidly. The week's most popular current songs across all genres and styles, ranked by radio airplay audience impressions. A music recommender system has been actually processed the researches to satisfy the needs for listeners and even help you to discover new artists. However, existing music recommendation systems do not reflect listeners' preferences due to recommendations simply employing listeners' listening log. In this paper, we propose a new music recommendation method in music site through listener's preference based Bayesian learning to reflect most popular current songs across all genres and styles on music portal sites, which have its popular music charts on trends, such as Melon, Billboard, Bugs Music, Soribada, and Gini. It is neccessay for us to take the task of preprocessing of calculating lisrener's preference to reflect preferred weight based online music sites with its popular music charts on trends in music database in order to reflect probably-preferred pieces from the database by estimating listener's preferences using listener's user profile. We can improve the performance of recommender system in music site using learning listener΄s preference based Bayesian learning. The next section briefly reviews the literature related to studies. Section 3 is described a new method for music recommender systemin detail, the algorithm for proposing system, and the procedure of processing the recommender. Section 4 describes the evaluation of this system in order to prove the criteria of logicality and efficiency through the implementation and the

experiment. In section 5, finally it is described the conclusion of paper and further research direction.

2. RELATED WORKS

2.1 Clustering

Clustering is the process of grouping physical or abstract objects into classes of similar objects. Its techniques[1,2] fall into a group of undirected data mining tools. The principle of clustering is maximizing the similarity inside an object group and minimizing the similarity between the object groups. Its algorithm is a kind of user's segmentation methods commonly used in data mining, can often use to k-means clustering algorithm. This algorithm uses as input a predefined number of clusters that is the k from its name. Mean stands for an average, an average location of all the members of a particular cluster. The euclidean norm is often chosen as a natural distance which user a between k measure in the k-means algorithm[2]. There are two part of k-means algorithm. The 1st part is that partition the objects into k clusters. The 2nd part is that iteratively reallocate objects to improve the clustering. The system can use Euclidean distance metric for similarity.

2.2 Bayesian Network (BN)

BN model is well known that classic machine learning methods like Hidden Markov models (HMMs), neural networks. BN became extremely popular models in the last decade. They have been used for applications in various areas, such as machine learning, text mining, natural language processing, speech recognition, signal processing, bioinformatics, error-control codes, medical diagnosis, weather forecasting, and cellular networks[2]. Specific types of BN models were developed to address stochastic processes, known as dynamic BN, and counterfactual information, known as functional BN[3]. With the BN, we formulate a item preference model in the form of a joint probability distribution. BN became extremely popular models in the last decade. They have been used for applications in various areas, such as machine learning, text mining, natural language processing, speech recognition, signal processing, bioinformatics, error-control codes, medical diagnosis, weather forecasting, and cellular networks. In general, the case of learning with known structure and partial observability, one can use the EM (expectation maximization) algorithm to find a locally optimal maximum-likelihood estimate of the parameters[2]. There are two main applications of the EM algorithm. The first occurs when the data indeed has missing values, due to problems with or limitations of the observation process. The second occurs when optimizing the likelihood function is analytically intractable but when the likelihood function can be simplified by assuming the existence of and values for additional but missing (or hidden) parameters. The latter application is more common in the computational pattern recognition community.

3. Our proposal for recommendation in music site

3.1. Clustering forlistener's preference based Bayesian learning

In this section, we suggest recommender system in music site using learning listener's prefererece based Bayesian learning. We prepare the experimental data with most popular current songs across all genres and styles for recommendation in music site. We have 1,000 listeners for user profile, who have had the experience to listen songs and have downloaded the mp3 music files from online music site with its popular music charts on trends in the music data reflected by most popular current songs across all genres and styles. That is, we have 1,000 listeners who have listened or downloaded and we use 500 songs data. There are 5 rates weighted by listener's prefered music site to reflect most popular current songs across all genres and styles on music portal sites, which have its popular music charts on trends, such as Melon, Billboard, Bugs Music, Soribada, and Gini. The following Fig. 1 show the result of weight by each music site listener prefered. In case of each rate, it is shown that the result of weight by each music site listener prefered.

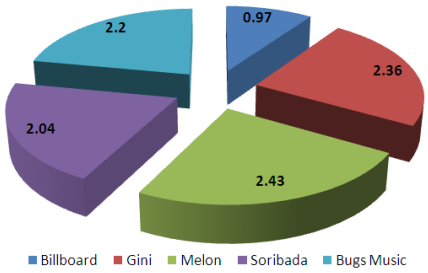


Figure. 1. The result of weight by each music site listener preferred

It was necessary for us to take the task of pre-processing to build the music database for structure of content based listeners' music. It was grouped and

ordered by each music genre and music style based on big/medium/small category as the following table 1.

Table 1. The list of category for music site

music sites with rank	Melon realtime Top 100, Billboad Hot 100, Sorobada popular chart, Gini Top 100, Bugs Daily Top 100
genre	R&B, Ballad, Dance, Folk, Electronica, Drama, Pop, Rock, Hip Hop, Ani, Pop, Country, Rap, Soul, Soft Pop
Style	R&B Ballad, 00' Ballad, Urban, Soft Pop/Rock, 00' Dance, Pop Rap, Medium, Folk Pop, Soft Dance, Electronica, Rap/Hip-Hop, Pop, R&B Dance, Pop Rock, Neo Soul, Urban, Club Dance, Punk Rock, Modern Folk, Dance Pop, Country Pop, Reggae, Korea TV Drama, Soul, CM Song, Alternative Pop, Indiem etc.

The big category is based on music sites with rank. The medium category is based on genre of music. The small category is based on style of music. The music database created after suitable preprocessing for structure of content based listener's music. The system can create the cluster of music data sorted by music genre and music style for the preprocessing task on the analytical agent. The system can compute listener's probability of preference of all categories of music genre and music style in clustering data which is selected by social variable such as age, gender, occupation, and music propensity. As a result of that, the system has finished the ready to recommend songs with high probability in music category belonged to brand songs. As a matter of fact, we use clustering for music genre and music style for music recommender system with Bayesian suggestion via learning through weight of listener's prefered music site with most popular current songs across all genres and styles, to adjust the result through Bayesian learning with weight of listener's prefered music site. We apply to make the task of clustering for music genre and music style to recommendation in the music site based on Bayesian learning.

4. The environment of implementation and experiment & evaluation

4.1. Experimental data for evaluation

We used 1,000 listeners of user profile, who had had the experience to listen songs and had downloaded the mp3 music files from online music site with its popular music charts on trends in the music data reflected by most popular current songs across all genres and styles. They had listened or downloaded from online music site and had used 500 songs data. It was necessary for us to make the task of clustering listener′s preference with weight of online music site using user profile. For doing that, we made the implementation for prototyping of music recommender system[5]. The experimental dataset for music recommender system was collected by each 5 online music sites for proving of the proposed. We have finished the system implementation about prototyping music recommender system. We'd try to carry out the experiments in the same condition with dataset collected in online portal music sites such as Melon, Billboard, Bugs Music, Soribada, and Gini. The 1st system is proposing system using BN learning through weight of listener's prefered music site called by "proposal", the 2nd system is the existing system called by "Previous".

4.2. Experiment and Evaluation

The mean absolute error(MAE) between the predicted ratings and the actual ratings of users within the test set. The proposing system's overall performance evaluation was performed by MAE. The mean absolute error is computed the following expression-1 over all data sets generated on purchased data.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | \varepsilon_i | \quad (1)$$

N represents the total number of predictions, ε represents the error of the forecast and actual phase i represents each prediction. The performance was performed to prove the validity of recommendation and the system's overall performance evaluation.

Table 2. The result of MAE based on Music Site

Site	Existing	Proposal
Billboard	0.09	0.04
Genie	0.1	0.05
Melon	0.09	0.04
Soribada	0.09	0.05
Bugs Music	0.11	0.06

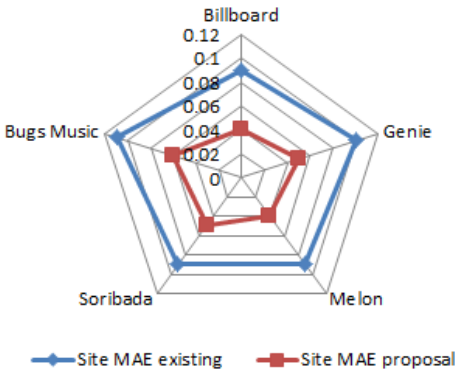


Figure. 2. The result of MAE based on music site

Figure. 3. The screen of music recommending site

Above Table 2 presents the result of music recommender system by evaluation using MAE. The proposing system is improved better performance than the existing system. Our proposing system using learning listener's prefererece is lower to 0.05 in average MAE of site than existing system**,** i.e., The MAE for the proposing system is lower average 2 times than the MAE of existing system. The proposing system is improved more 2 times than the existing system. We can improve the performance of recommender system in music site using learning listener's preference based Bayesian learning. As a result, we could have the music recommender system in music site to be able to recommend the songs with an immediate effect. The Figure. 3 is shown in the screen of music recommending site on a smart phone under ubiquitous computing environment. The proposing system is better performance than the previous system.

5. CONCLUSION

Along with the spread of digital music and the development of music source, the demands for music recommender are increasing. We proposed a new music recommender system using learning listener's prefererece. We reflected most popular current songs across all genres and styles in order to improve the accuracy of recommender, to reduce listeners' searching effort to find out the songs with an immediate effect. We carried out experiments with dataset of collecting from online portal music sites to measure its performance. We reported some of the experimental results. The existing system did not yet reflect the importance of listener's prefered music site with most popular current songs across all genres and styles, then it did not consider these dynamic changes in different songs. It was crucial to have different value for weight of listener's prefered music site and adjust the results by reflecting the important songs based on listener's prefered music site in order to improve the accuracy of music recommender, to meet the needs of listeners changing according to the trend of music genre and music style. We carried out experiments with dataset of collecting from online portal music sites to measure its performance. We reported some of the experimental results. We improved the performance of recommender system in music site using learning listener's preference based Bayesian learning. It is meaningful to present a new music recommender system using BN learning through weight of listener's prefered music site in online portal site environment.

6. ACKNOWLEDGEMENTS.

This researc[1)] h was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No.2013R1A2A2A01068923). It was supported by BNU-HKBU United International College, Zhuhai, China.

7. References

[1] Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. MIT press.

[2] Steyvers, M., Griffiths, T. L., & Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences*, *10*(7), 327-334.

[3] Pearl, Judea, and Stuart Russell. *Bayesian networks*. Computer Science Department, University of California, 1998.

[4] J.L. Herlocker, J.A. Konstan, A. Borchers, J. Riedl. 1999. *An Algorithmic Framework for Performing Collaborative Filtering.* Proceedings of the 22nd International Conference on Research and Development in Information Retrieval, 230-237.

[5] Y.S. Cho, S.C. Moon, I.B. Oh, J.H. Shin, K.H. Ryu. 2013. *Incremental Weighted Mining based on RFM Analysis for Recommending Prediction in u-Commerce*. Ubiquitous Information Technologies and Applications, vol.7, No. 6, 133-144

Competitiveness Enhancement of Home IoT Service by Smart Home Mirror

Yeong Real Kim[1], Tae Gu Kang[2], Kyung Mun Kang[3]

[1]*Dept. of Management Informations, .Professor, ,Chungbuk National University*
361-763,Gaeshin-dong Heungduk-gu Cheong-ju, Chungbuk , Korea

[2]*Dept. of Management Informations, .Ph.D Cours ,Chungbuk National University*
361-763,Gaeshin-dong Heungduk-gu Cheong-ju, Chungbuk , Korea

[3]*SOMA System, Inc. President & CEO 5, Seongsuil-ro 8-gil, Seongdong-gu, Seoul, Korea*
yrkim@chungbuk.ac.kr snookerk@hanmail.net soma@somacns.co.kr

Abstract

Smart home business is in growing trend under recognition as one of the future businesses but its obstacles are calling for countermeasures and its functions are not enough to satisfy customer's needs for houses that can be easily controlled by all-in-one control. Currently, its commercialization is not fully developed due to different OS by manufacturers of home appliances, lack of killer-contents, limits in sharing, insufficient user profile management, and burden of new investment. Thus, it is urgent to expand smart home focusing on the expansion package and evolve to personalized service, so smart mirror product is now regarded as the most powerful alternative to solve all the problems above which already reached a market scale of billion dollars in 2013. With social reputations and through satisfaction of previously impossible desires, smart home mirror is suggested as an alternative to overcome obstacles on the road of the smart home business and its potential will be contemplated.

Keywords: Smart Home, Smart Mirror, IoT

1. Current Status and Issues of Smart Home Business

1.1. Smart Home?

It is a human-centered smart life environment to realize convenience, improvement of welfare, and safe lives of people through convergence of IT with residential environment. It refers to all products, services, solutions, etc. which monitor, control, and operate household elements in its network: home appliances such as TV, air conditioner, refrigerator, etc.; energy consumption devices such as water, electricity, heating and cooling, etc.; and security devices such as door-lock, surveillance camera, etc[1].

1.2. Development Stage of Smart Home

· As High-Speed Internet service was widely provided in the late 1990s, domestic smart home business had a chance for invigoration of home network business but it was slipped away and stuck in gridlock.

· Establishment of early stage of smart home due to popularization of smartphones in late 2000s

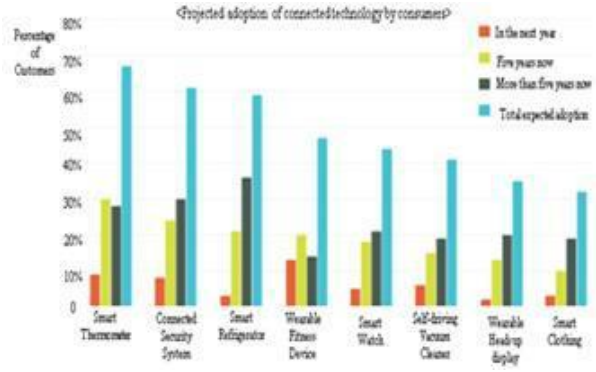
· As IoT was commercialized in late 2010s, Situation Aware smart home is being introduced.

1.3. Domestic Smart Home Market and Prospect

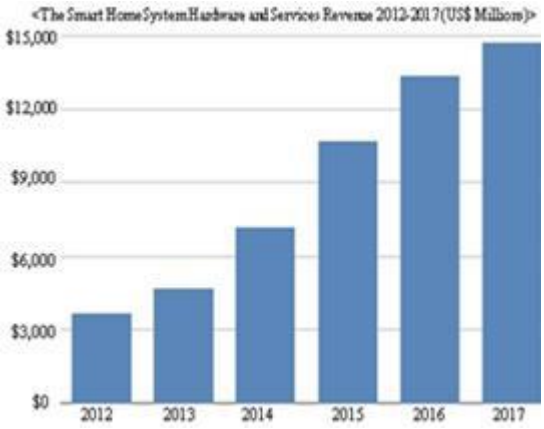
Domestic smart home market is expected to show 27.5% of annual average growth in turnover until 2017, and currently manufacturers such as Samsung and LG and network providers such as SKT are consolidating their status as the Big 3[2]

1.4. Global Smart Home Market and Prospect

As smartphones become popular now a days, the customers‟ needs for smart devices such as „security, home appliance, health care, etc.‟ which have interlinkage with smartphones are in increase trend, and the prospect of the market size for smart home is gradually expanding according to changes of consumption environment, expecting the growth of global smart home market scale up to 14.7 billion dollars by 2017[3].



Source : Acquity Group, 2014.08



Source : NextMarket Insights

2. Issues in Smart Home Business

2.1 4 Home Needs by Consumers

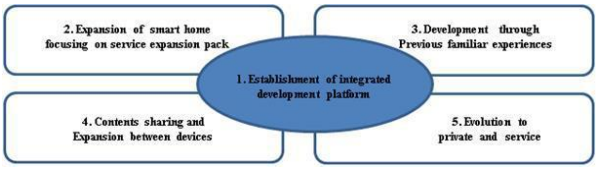
(4) Convenient house with All-In-One connection by integration of home appliances and contents
(5) Saving house with less time and labor for house chores
(6) Information house with information available at every corner of interior
Controllable house easily by anyone

Source : Home Needs by Consumers, Hyeon myeong Pyo,2013

2.2. Hindrances to Satisfaction of Consumer Needs

· High price of devices and low utilization
· Unavailability of integrated control of home appliances
· Different OS between home appliance manufacturers
· Lack of killer contents
· Limit on contents sharing
· Late commercialization due to new investment

2.3. Plan to Meet Consumer Needs



Source : Smart home five Grand Strategy, Hyeon myeong Pyo, 2013

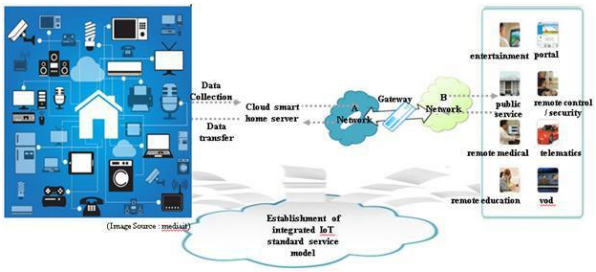
3. Smart Home Business Model

3.1. Smart Home Mirror

[9] Recommendation of clothes to wear according to result of weather analysis with previously provided clothes data
[10] Try-on simulation service function
Try-on simulation at home before order without visiting stores
[11] Provision of information for new products through preference analysis s
Comparison with existing clothes before purchase

3.2 Establishment of integrated IoT standard service model

· Provision of standard platform through open-type smart IoT Home model to meet needs of consumers and support different products through liaison of smart home IoT service model with home and service support (medical service, tourism, public service, education, etc.) industry



4. Conclusion

4.1 Conclusion and Suggestion

[7] Smart home business is in growing trend but new solutions should be considered to leap beyond it

[8] Recently emerging smart mirror product market is in growing trend but it is showing its limits due to lack of unwavering direction in its value creation

[9] Value revelation through smart home mirror takes its meaning as a solution to overcome obstacles in the smart home business.

[10] Preparation of technologies for products and services of smart home business in the future through connection and integration between mirrors

[11] Plan to prepare business model for sharing and integration of smart home mirror and smart home services (medical service, tourism, public service, education, etc.)

[12] Expected effect of creating various business models by meeting needs of customers based on connection between smart home IoT service model with service supporting industry

[13] Provision of standard platform to different smart home IoT products through development smart IoT home model

5. References

[1] Hyeon myeong Pyo, “Creating smart home strategy economy”, Korea Association of Smart Home,2013.

[2] KT Economic Research, “Smart home and abroad Trends and Implications”, Korea Association of Smart Home, 25.06,2013

[3] Teja Patankar, “2014 State of the Internet of Things Study from Accenture Interactive Predicts 69 Percent of Consumers Will Own an In-Home IoT Device by 2019”, Accenture, 19.08,2014.

Screening of Allosteric inhibitors for p21-activated kinases

Duk-Joong Kim[1], Chang-Ki Choi[1], Chan-Soo Lee[1], Kyung-Ah Kim[2], Eun-Young Shin[1], and Eung-Gook Kim[1]

[1]Department of Biochemistry and [2]Biomedical Engineering, College of Medicine, Chungbuk National University, Cheongju 28644, Korea

eyshin@chungbuk.ac.kr

Abstract

P21-activated kinases (PAKs) are key regulators of actin dynamics, cell proliferation and survival. Deregulation of PAK activity contributes to pathogenesis of various human diseases including cancer and neurological disorders. Using an ELISA-based screening protocol, we identified naphtho(hydro)quinone-based small molecules that allosterically inhibit PAK activity. These molecules interfere with the interactions between the p21 -binding domain (PBD) of PAK1 and Rho GTPases by binding to the PBD. Importantly, they inhibit the activity of full-length PAKs and are selective to PAK1 and PAK3 in vitro and in cells. These compounds are potentially useful for dissecting the PAK signaling pathway and can also be used as lead molecules for the development of more selective and potent PAK inhibitors.

Keywords: *p21-activated kinase, Cdc42, allosteric inhibitor*

1. Introduction

PAKs are Ser/Thr kinases classified into two groups on the basis of their structural and functional features: group I (PAK1–3) and group II (PAK4–6) [1]. Group I PAKs have an auto-inhibitory domain and a kinase domain and are activated by the binding of the active forms of Rho GTPases, such as Cdc42 and Rac1. Group II PAKs have no auto-inhibitory domains and are not activated by active Rho GTPases. PAK1, the best-characterized member of group I PAKs, forms auto-inhibited homodimers, in which the active site of the kinase domain in one monomer is blocked by the inhibitory switch domain (residues 87–136) of the other; the inhibitory switch domain overlaps partially with the p21-binding domain (PBD, residues 67–150) (Figure 1). When Cdc42•GTP or Rac1•GTP interacts with the PBD of PAK1, PAK1 is converted to a monomeric form, leading to a conformational change of its catalytic domain to restore its kinase activity [2-3]. This event induces autophosphorylation of Thr423 followed by autophosphorylation of multiple residues in PAK1. Because deregulation of PAKs is closely associated with various human diseases [4-5], small-molecule inhibitors of these kinases have great potential as therapeutic agents [6]. In addition, these compounds can also be used as powerful tools in studies aimed at understanding the PAK signaling pathway. Herein, we describe naphtho(hydro)quinone (N(H)Q)-based small molecules that allosterically inhibit PAK activity by binding to the regulatory domains (PBD) rather than to the ATP-binding sites. The developed compounds selectively inhibited the activities of the group I PAK kinases PAK1 and PAK3.

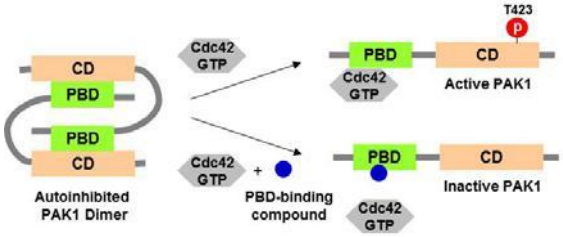


Fig. 1. Schematic representation of Cdc42-dependent PAK1 activation and its inhibition by small molecules.

2. Results

2.1. Screening for compounds that disrupt the Cdc42–PBD interaction

To search for small-molecule inhibitors that inhibit PAK1, we developed a screening protocol that involves an ELISA based on a strong Cdc42–PBD interaction. For this purpose, Cdc42 was expressed as a

GST fusion protein (GST-Cdc42) and PBD as a His-tagged protein (PBD-His). GST-Cdc42•GTP was added to a 96-well plate coated with PBD-His in the presence of small molecules (1,280 compounds) with molecular weights of less than 300 Da. The disruption of the Cdc42–PBD interaction caused by the small molecules was assessed using HRP-conjugated anti-GST antibody. Initial screening identified two compounds, 1,4-naphthohydroquinone (compound 1, 1,4-NHQ) and 2-methoxycarbonyl-1,4-naphthohydroquinone (compou-nd 2, 2-Mc-1,4-NHQ), which effectively inhibited the Cdc42–PBD interaction in vitro. Compounds 1 and 2 blocked the association of Cdc42 with PBD with respective IC50 values of 7.25 M and 6.59 M (Figure 2). We call these compounds the PBD-binding compounds hereafter.

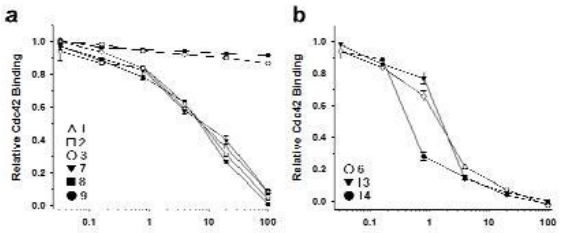


Fig. 2. Suppression of the Cdc42–PBD interaction by compounds.

2.2. Analysis of binding between compounds and PBD.

To determine whether compounds directly bound to the PBD, their binding affinities were measured using SPR spectroscopy. For these studies, purified GST (control) and GST-PBD were immobilized on a modified gold surface, and various concentrations of each compound were subsequently applied. The dissociation constants (Kd values) for the interactions of compound 2 with the PBD interactions were determined to be less than 7 M (Figures 3a and b). The Cdc42/Rac1-interactive binding (CRIB) motif (residues 75–90; PAK1 numbering) of the PBD is important for its interaction with Cdc42 and Rac1[7-8]. To test whether this motif was important for PBD interaction with the PBD-binding compound, we constructed a mutated PBD in which His83 and His86 were replaced with Leu (thus named PBD-LL) and whose Rho GTPase-binding ability was abolished.16 If the CRIB motif is involved in PBD interaction with the PBD binders, their binding affinities to the mutant PBD-LL should be attenuated. We therefore used SPR analysis to compare the binding abilities of the PBD binder (compound 2) to the wild-type PBD (Wt-PBD) and PBD-LL. As shown in Figures 2c, the binding affinities of the PBD-binding compound to PBD- LL were significantly reduced (to approximately half of those for Wt-PBD binding).

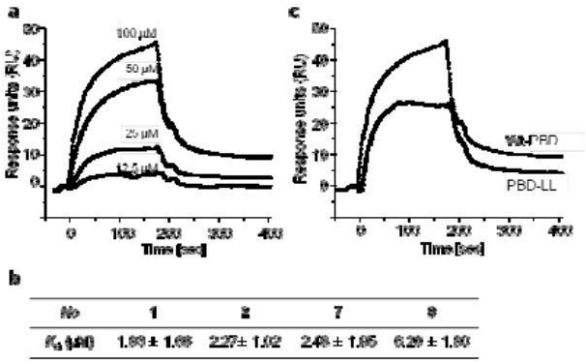
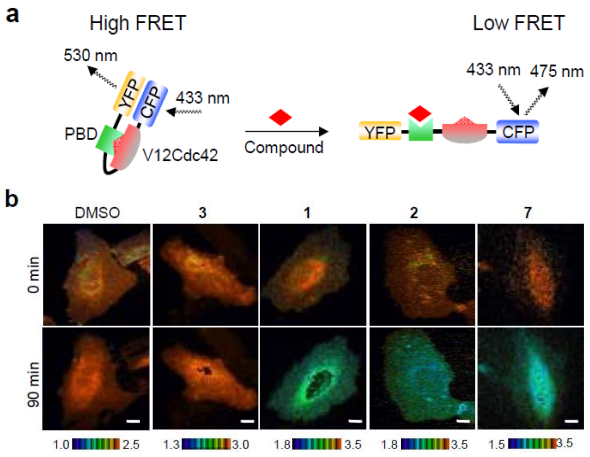


Fig. 3. Binding of compound 2 to the PBD.

2.3. Effects of compounds on the activity of full-length PAK in cells

We evaluated the inhibitory potency of the PBD binders in living cells. First, to determine whether the compounds suppress the Cdc42–PBD interaction, we performed FRET analysis with cells expressing the YFP-PBD and CFP-V12Cdc42 fusion proteins (Figure. 4a–c) [9]. V12Cdc42 is constitutively active because of its defective GTPase activity. Thus, V12Cdc42 should constitutively interact with PBD in cells. In the absence of PBD-binding compounds, high FRET should be observed due to the close proximity of YFP and CFP caused by the V12Cdc42–PBD interaction (Figure 4a). However, if PBD-binding compounds prevent the V12Cdc42–PBD interaction by binding to the PBD, FRET efficiency should be reduced as a result of a longer distance between YFP and CFP.



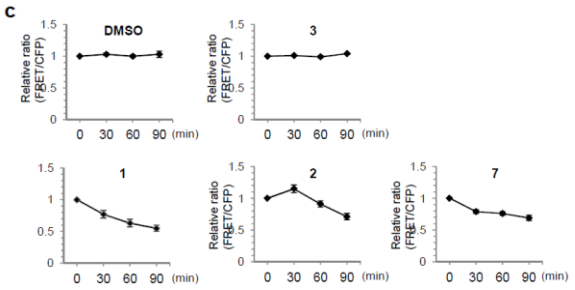


Fig. 4. Suppression of PAK1 activity by PBD-binding compounds in cells.

Changes in FRET efficiency were determined in the absence and presence of the compounds. When cells were exposed to 20 M compound 1, 2 or 7, no significant cytotoxicity was observed (data not shown), but the FRET efficiency decreased (Figure 4b; quantified in Figure 4c). In contrast, cells treated with inactive compound 3 did not exhibit any detectable change in FRET efficiency (Figure 4b and 4c). These results indicate that PBD-binding compounds can block the Cdc42–PBD interaction in cultured cells in a manner similar to that observed in vitro.

3. Discussion

In the present study, we developed a new ELISA-based high-throughput screening (HTS) system to search for potential allosteric inhibitors of group I PAKs. Our screening is based on the ability of the compounds to disrupt the interactions between Cdc42 or Rac1 and PBD of PAK1. We further showed that the identified compounds directly bound to the PBD and inhibited the activities of group I PAKs except PAK2 but not those of group II PAKs. Thus, we demonstrated the validity of our HTS for identification of allosteric inhibitors of PAKs. Modification of this HTS for screening of compounds that directly bind to the PBD using a chip-based chemical library would also be feasible. However, allosteric inhibitors including IPA-3 have lower potency than ATP-competitive inhibitors, an issue that must be resolved in the future.

The Cdc42/Rac1–PBD interaction involves the CRIB motif that contains the critical residues His83 and His86 [7-8]. In accordance with the importance of these histidines, a single substitution of any of these residues with Leu appeared to be sufficient to disrupt the interaction. Consistent with this result, allosteric inhibitors bound the double-mutant form of PBD, PBD-LL, less efficiently than Wt-PBD, and did not inhibit the activity of full-length PAK1-LL.

1, 4 -NQ derivatives are found in a large number of natural products, and some of them have been used as herbal medicines and show cytotoxicity against cancer cells and microbes [10-11]. Thus, 1, 4-NQ-based small molecules are attractive compounds for clinical use because of their potential application as anticancer and antimicrobial agents, although the scientific rationale at the molecular level has not yet been clearly defined. The present study provides evidence for group I PAKs as novel targets of naphthoquinone compounds. In recent years, group I PAKs, especially PAK1, have emerged as therapeutic targets in diverse types of cancer [12-13]. Given that NQs and their derivatives may target group I PAKs, they may be useful as lead compounds. Cytotoxicity of NQ-based compounds is mostly caused by their thiol reactivity; thus, chemical modifications to reduce this reactivity could generate non-toxic inhibitors [14] and may also reduce unwanted side effects arising from interactions with non-target proteins.

In summary, we have identified small molecules that allosterically inhibit PAK activity by interacting with the regulatory domains (PBD) rather than with the ATP-binding sites in vitro and in cells. The identified inhibitors selectively suppressed the activity of group I PAKs, particularly PAK1 and PAK3. Comparison of the inhibitors identified in the present work with those reported previously represents an important step toward the development of selective PAK inhibitors and pharmacological intervention in patients with PAK-associated diseases. In addition, the inhibitors characterized here could be valuable tools in studies aimed at understanding PAK signaling pathways.

4. References

[1] Jaffer ZM, Chernoff J. p21-activated kinases: three more join the Pak. Int J Biochem Cell Biol 2002; 34: 713-717.

[2] Morreale A, Venkatesan M, Mott HR, Owen D, Nietlispach D, Lowe PN et al. Structure of Cdc42 bound to the GTPase binding domain of PAK. Nat Struct Biol 2000; 7: 384-388.

[3] Lei M, Lu W, Meng W, Parrini MC, Eck MJ, Mayer BJ et al. Structure of PAK1 in an autoinhibited conformation reveals a multistage activation switch. Cell 2000; 102: 387-397.

[4] Kumar R, Gururaj AE, Barnes CJ. p21-activated kinases in cancer. Nat Rev Cancer 2006; 6: 459-471.

[5] Kreis P, Barnier JV. PAK signalling in neuronal physiology. Cell Signal 2009; 21: 384-393.

[6] Yi C, Maksimoska J, Marmorstein R, Kissil JL. Development of small-molecule inhibitors of the group I p21-activated kinases, emerging therapeutic targets in cancer. Biochem Pharmacol 2010; 80: 683-689.

[7] Manser E, Leung T, Salihuddin H, Zhao ZS, Lim L. A brain serine/threonine protein kinase activated by Cdc42 and Rac1. Nature 1994; 367: 40-46.

[8] Burbelo PD, Drechsel D, Hall A. A conserved binding motif defines numerous candidate target proteins for both Cdc42 and Rac GTPases. J Biol Chem 1995; 270: 29071-29074.

[9] Itoh RE, Kurokawa K, Ohba Y, Yoshizaki H, Mochizuki N, Matsuda M. Activation of rac and cdc42 video imaged by fluorescent resonance energy
transfer-based single-molecule probes in the membrane of living cells. Mol Cell Biol 2002; 22: 6582-6591.

[10] Tandon VK, Singh RV, Yadav DB. Synthesis and evaluation of novel 1,4-naphthoquinone derivatives as antiviral, antifungal and anticancer agents. Bioorg Med Chem Lett 2004; 14: 2901-2904.

[11] Tandon VK, Kumar S. Recent development on naphthoquinone derivatives and their therapeutic applications as anticancer agents. Expert Opin Ther Pat 2013; 23: 1087-1108.

[12] Kichina JV, Goc A, Al-Husein B, Somanath PR, Kandel ES. PAK1 as a therapeutic target. Expert Opin Ther Targets 2010; 14: 703-725.

[13] Eswaran J, Li DQ, Shah A, Kumar R. Molecular pathways: targeting p21-activated kinase 1 signaling in cancer--opportunities, challenges, and limitations. Clin Cancer Res 2012; 18: 3743-3749.

[14] Vasudevarao MD, Mizar P, Kumari S, Mandal S, Siddhanta S, Swamy MM et al. Naphthoquinone-mediated inhibition of lysine acetyltransferase KAT3B/p300, basis for non-toxic inhibitor synthesis. J Biol Chem 2014; 289: 7702-7717.

Flow generator system for calibration and comparison of air flow modules

Eun-Jong Cha, Mi-Jung Park, Ji-Sun Lim, Eun-Young Shin[1], Yang-Mi Kim[2], Ho-Sun Shon[3], Kyoung-Ok Kim[4], Kyung-Ah Kim
Department of Biomedical Engineering, [1]Department of Biochemistry, [2]Department of Physiology, [3]Medical Research Institute, School of Medicine, Chungbuk National University, Cheongju, [4]Department of Nursing, Woosong College, Daejeon, Korea
kimka@chungbuk.ac.kr

Abstract

An air flow generator system was developed to generate air flows of various levels simultaneously applied to two different air flow transducer modules. Axes of two identical standard 3l syringes were connected in parallel and driven by a servo-motor. Linear displacement transducer was also connected to the syringe axis to accurately acquire the volume change signal. The user can select either sinusoidal or square waveform of volume change and manually input any volume as well as maximal flow rate levels ranging 0~3 l and 0~15 l/s, respectively. Various volume and flow levels were input to operate the system, then the volume signal was acquired followed by numerical differentiation to obtain the air flow signal. The measured volumes and maximal air flow rates were compared with the user input data. The relative errors between the user-input and the measured stroke volumes were all within 1%, demonstrating very accurate driving of the system. In case of the maximal flow rate, most measured flow rates revealed relative errors ≤2%. These results demonstrate that the servo-motor controls the syringes with good enough accuracy to generate standard air flows. Therefore, the present system would be very much practical for calibration process as well as performance evaluation and comparison of two different air flow measurement modules.

Keywords: Calibration technique, Spirometer, Air flow generator

1. Introduction

Spirometry is a clinical test to evaluate how much air is breathed in/out how fast by the patient, which requires the measurements of air flow rates and volumes. Spirometers should satisfy various technical standards proposed by the American Thoracic Society (ATS) and the European Respiratory Society (ERS) for accurate diagnosis [1-6]. One of the standard volume calibration devices is a 3 *l* syringe suggested to operate manually several times followed by the evaluation of volume measurement accuracy of the spirometer at least once everyday. The present study developed a standard air flow generator system by connecting two identical 3 *l* syringes in parallel driven by a servo-motor for calibration as well as performance comparison of two different air flow measurement modules. Experiments to generate various tidal volumes and maximal flow rates were performed.

2. Methods

2.1. System configuration

The present flow generator system was configured as depicted in Fig. 1. Two identical 3 *l* syringes with diameters of 10 cm (Syr3.0, CKInt., Co., Korea) were connected in parallel, which were the standard calibration devices for spirometers [2-4]. These two syringes moved air in (inspiration) and out (expiration) in exactly same pattern by a servo-motor at the same time. A linear displacement transducer connected to the axis provided the volume signal (v(t)) during syringe movement.

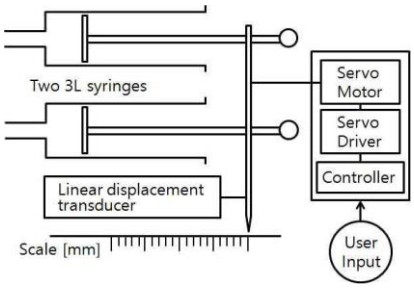


Fig. 1. System configuration

Syringes were driven in either sinusoidal or squared flow pattern (f(t)) with single or ten stroke mode. The

tidal volume (V_S) per stroke ranged 0-3 *l* and the maximal air flow rate (F_S), 0-15 *l*/s, set at any level by the operator. The ten stroke mode started with 2 strokes having the user defined F_S then another successive two strokes were followed with F_S decreased at 5 steps each, forming a total of 10 strokes.

2.2. Experimental procedure

Two series of experiments were performed to evaluate how accurately the system accomplished V_S or F_S, given by the user, respectively. v(t) signals were accumulated while V_S were stepwisely increased (1.0, 1.5, 2.0, 2.5, 3.0 l) at a constant F_S=6.0 *l*/s, in single sinusoidal mode operation. Tidal volume (V_M) were calculated by integrating v(t) for each stroke. Then, V_S were set constant to 3.0 *l* and ten stroke sinusoidal mode operation was performed with Fs=0-14 *l*/s. v(t) signal was numerically differentiated to obtain the maximal flow rate (F_M) accomplished by the system for each stroke. V_M and F_M were compared with V_S and F_S, respectively, to evaluate the driving performance of the present air flow generator system.

3. Results

3.1. Tidal volume accuracy

Fig. 2 shows v(t) with F_S=6 *l*/s and V_S=2 *l*. 2 *l* of air was completely expired in 0.2-0.3 s in a sinusoidal fashion, and maintained for 1 s representing V_M level followed by a sinusoidal decrease (inspiration) as such designed. V_M average values were compared with the corresponding V_S in Table 1, demonstrating accurate volume generation performance with a mean relative error of 0.751%.

3.2. Maximal flow rate accuracy

Fig. 3 shows v(t) with F_S=10 *l*/s and V_S=3 *l* in 10 stroke mode. v(t) was numerically differentiated to obtain the flow rate signal (f(t)) as shown in Fig. 4. As designed in 10 stroke mode, f(t) decreased stepwisely every two strokes. The largest values of f(t) were read to get F_M for each stroke. Relative errors of F_M compared with F_S are summarized in Table 2, showing accurate flow generation capability with mean relative error of 1.413%.

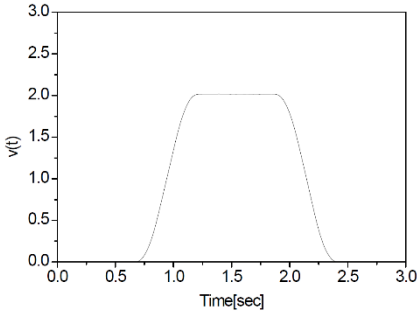


Fig. 2. v(t) example at F_S=6 *l*/s and V_S=2 *l*

Table 1. Vs and V_M data.

V_S	V_M	Relative error (%)
0.997	1.005	0.715
1.500	1.511	0.723
2.003	2.017	0.722
2.498	2.516	0.756
3.000	3.025	0.837
	Mean	0.751

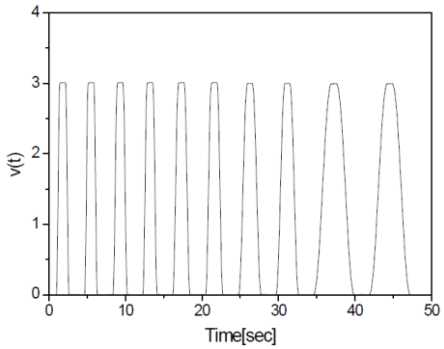


Fig. 3. Volume signal example at Fs=10 *l*/s for 10strokes

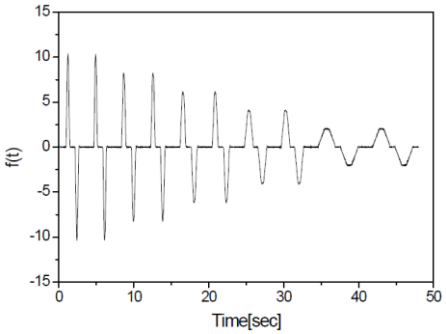


Fig. 4. Flow signal example at Fs=10 *l*/s for 10 strokes

Table 2. F_s and F_M data for 10 stroke mode

Fs	Relative error (%)
2	1.035

4	0.890
6	0.978
8	1.840
10	1.846
12	2.042
14	1.262
Mean	1.413

4. Discussion

Spirometer calibration is of great importance for accurate diagnosis. International standards suggest to apply 3 *l* syringe manually operated at least once a day [1-4]. The manual operation, however, introduces inconsistency due to different user habits as well as sometimes not enough flow range for calibration inevitable due to inconsistent manual operation. The present flow generator system was developed to provide wide enough flow range with operation conveniences for user. As a result of experiment, V_M was demonstrated to be almost the same with V_s defined by user, showing mean relative error<1%. F_M was also satisfactory with a mean relative error $\leq$ 2% well below the international standard limit of 5%.

While the present system generates accurate flow pattern as defined by user, it also enables comparison of two different air flow measurement modules by connecting each modules to two identical syringes in parallel, respectively. The servo-motor control guarantees accuracy as well as convenience for automatic operation. Either squared or sinusoidal flow wave can be selected for any preferred applications. The linear displacement transducer provides the continuous volume signal during the whole calibration procedure enabling various comparisons of volume/flow depending upon the user purposes in addition to a few parameter comparisons suggested by the international standards. Therefore, the present system could also be applied by the manufacturer for quality assurance of spirometer production.

5. References

[1] M.R. Miller, J. Hankinson, V. Brusasco, F. Burgos, R. Casaburi, A. Coates, R. Crapo, P. Enright, C.P.M. van der Grinten, P. Gustafsson, R. Jensen, D.C. Johnson, N. MacIntyre, R. McKay, D. Navajas, O.F. Pedersen, R. Pellegrino, G. Viegi and J. Wanger, “Standardization of spirometry", Eur Respir J, vol. 26, pp.319-338, 2005.

[2] A.D. Jr. Renzetti, “Standardization of spirometry”, Am Rev Respir Dis, vol. 119, pp.831-838, 1979.

[3] American Thoracic Society, “Standardization of spirometry: 1987 update”, Am Rev Respir Dis, vol. 136, pp.1285-1298, 1987.

[4] American Thoracic Society, “Standardization of spirometry, 1994 update”, Am J Respir Crit Care Med, vol. 152, pp.1107-1136, 1995.

[5] P.H. Quanjer, ed. "Standardized lung function testing. Report Working Party Standardization of Lung Function Tests. European Community for Coal and Steel". Bull Eur Physiopathol Respir, vol. 19:Suppl.5, pp.1-95, 1983.

[6] Official Statement of the European Respiratory Society, “Lung volume and forced ventilatory flows. Report Working Party Standardization of Lung Function Tests, European Community for Steel and Coal”, Eur Respir J, vol. 6:Suppl.16, pp.5-40, 1993.

Risk Factor of Non ST-segment Elevation Myocardial Infarction (NSTEMI) Patients with Diabetes

Ho Sun Shon[1], Kyung Ah Kim[2]
Medical Research Institute, Chungbuk National University, Cheongju, Korea[1]
Dept. of Biomedical Engineering, School of Medicine, Chungbuk National University, Korea[2]
E-mail: shon0621@gmail.com, kimka@chungbuk.ac.kr

Abstract

Background: *Diabetes occurs when pancreas produces less insulin or the produced insulin cannot be used well in our body. Aging of the population, the westernization of the culture including food, intake of high-calorie food, reduction of exercise, and obesity have conspicuously increased prevalence rate of diabetes. Also, diabetes is becoming one of the major causes for death. The most important cause of death by diabetics is coronary heart disease. The risk of death of diabetes patient with cardiac infarction is twice higher than those without diabetes in males and 4 times higher than those without diabetes in females.*

Methods: *This study used the data provided by KAMIR and included only NSTEMI patients among all the patients. Especially among patients with the history of diabetes, it targeted patients who had follow-up major adverse cardiac event (MACE) cases for 12 months and analyzed risk factors. Variables such as gender, age, preTIMIflow, Killip class, and systolic blood pressure were used for analysis. Through blood test that was carried out during patients' visit to the hospital, significance test was conducted through Glucose, Creatinine, CK, CK-MB, HsCRP, troponin-I, troponin-T, total cholesterol, triglyceride, HDL cholesterol, LDL cholesterol and NT-proBNP.*

Results: *The experiment analyzed patients with diabetes who arrived within 12 hours after the onset of chest pain. As a result, through NSTEMI patients who were followed up for 12 months, it was found that symbolic blood pressure (p=0.001), diastolic blood pressure (p=0.007), glucose (p=0.001) and triglyceride (p=0.013) were the independent risk factors that cause major heart problems.*

Conclusion: Among risk factors that cause MACE in the data of NSTEMI patients with the history of diabetes who were followed up for 12 months, we founded that the risk factors were vital sign, glucose, and triglyceride. Therefore, these factors can be used to diagnose and estimate prognosis of NSTEMI patients with the history of diabetes.

Acknowledgement
This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No-2013R1A1A206518).

The electrophysiological role of epigallocatechin-3-gallate and quercetin as TREK2 antagonists

Kyung-Ah Kim[1], Yangmi Kim[2]
Department of Biomedical Engineering[1] and Physiology[2], School of Medicine, Chungbuk National University, Cheongju
yangmik@chungbuk.ac.kr

Abstract

Two-pore domain potassium (K2P) channels are targets of physiological stimuli such as intracellular pH, fatty acid, mechanical stretch, neurotransmitter, and Ca^{2+} and have been known to set resting membrane potential. In present study, using single channel patch clamp methods, we observed that blocking effect on TREK2 channel by flavonoids such as epigallocatechin-3-gallate (EGCG) and quercetin in TREK2 stably expressing HEK293 cells (HEKT2). EGCG analogues, epicatechin(EC) had no significant inhibitory effects on TREK2 single channel activity. Also we confirmed that ECGG reduced cell proliferation in HEKT2 cells. We concluded that EGCG and quercetin represents the first known TREK2 channel inhibitor. It suggests that the flavonoids may work primarily by inhibiting TREK2 channel, leading to change of resting membrane potential and trigger the initiation of change in intracellular signaling for cell proliferation TREK2 channel is may, at least in part, contribute to cell growth.

Keywords*: *TREK2, Quercetin, proliferation, EGCG, Quercetin

1. Introduction

Two pore potassium channels that regulate ionic concentration in intracellular environment were classified type of 15 members including TWIK, TREKs, TRESKs, TASKs and THIK [1]. TREK-1 and TREK2 are activated by mechanical stretch, acidic pH, riluzole and temperature and inhibited by antidepressant. Quercetin and EGCG is an attractive therapeutic flavonoid for cancer treatment because of its beneficial properties including apoptotic, antioxidant, and antiproliferative effects on cancer cells [2]. However the K2P channel-related cell proliferation cellular mechanism of the flavonoids is still not thoroughly understood, especially in TREK2. Here we investigated the effect of activating or inhibiting drug on TREK2 using flavonoids. Also we examined whether EGCG and quercetin has anti-proliferation effects on HEKT2.

2. Methods

TREK2 expressing stable HEK293 cells (HEKT2) and no channel expressing HEK293 cells (HEK) were maintained in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 100 units per ml penicillin, and 100 µg per ml streptomycin in a humidified incubator at 37°C with 5% CO_2.

Electrophysiological recording was performed in the single channel recording using a patch clamp amplifier (Axopatch 200B, Axon Instruments, Inc., Foster city, CA, USA). Pipette and bath solution contained (mM) 150 KCl, 1 $MgCl_2$, 1 $MgCl_2$ and 5 EGTA, and was titrated to pH 7.2 with KOH. Relative current were plotted as a function of [flavonoid]. With the reasonable assumption that 100 µM EGCG produces maximal inhibition of TREK-2, averaged data from five patches were fitted to a Hill equation of the following form : $y = 1/(1+(K_{1/2}/[\text{flavonoid}])^n)$, where $K_{1/2}$ is the apparent concentration of flavonoid that produces half-maximal inhibition and n is the Hill coefficient. Origin software (Origin Corp., Northampton, USA) was used to fit the plot to the Hill equation. All values were presented as mean ± S.E. The differences between the means of control and treatment data were determined using the paired Student's t-test.

To evaluate the cell proliferation of HEKT2 treated with flavonoids and K channel blocker, cells are grown in microtiter plates in a final volume of 100 µl culture medium per well. Cells were incubated 3 days with flavonoid and K channel blocker. The absorbance of the samples was measured with a spectrophotometer

(ELISA reader) at a wavelength of 450 ~ 550 nanometer.

3. Results

Among the K2P channels, we tested the effects of flavonoid such as EGCG and quercetin on TREK2 using single channel recording. EGCG and quercetin inhibited TREK2 current was voltage-clamped under inside out patch configuration at -60 mV. TREK2 channel activity was reduced to 93% (n=5) and 83% (n=5) by flavonoids such as epigallocatechin-3-gallate (EGCG) and quercetin in HEKT2 cells, respectively. Whereas, EGCG analogues**,** epicatechin (EC) had no significant inhibitory effects on TREK2 single channel activity.

To analyzing of dose dependency of EGCG and quercetin on TREK-2 channel activity, we applied graded increase of concentration in EGCG and quercetin. The result showed inhibitory effect on TREK2 with a dose dependency. Half-maximal inhibition of concentration (IC_{50}) for EGCG was 19 μM. Quercetin also decreased the TREK-2 channel activity with dose dependent manner. Half-maximal inhibition of concentration (IC_{50}) for quercetin was 4 uM. The dose response curve was recorded at holding potential of 0 mV and ramp pulse from -100mV to 100 mV for 200 ms using inside out patch. Taken together, EGCG and quercetin blocked the TREK2 channel with dose dependently.

We tested another flavonoids such as epicatechin(EC), an analoge of EGCG, and apigenin, [5,7-Dihydroxy-2-(4-hydroxyphenyl)-4*H*-1-benzopyran-4-one] source from parsley, on TREK2. Green tea catechins include (-)-epigallocatechin gallate (EGCG), (-)-epigallocatechin (EGC), (-)-epicatechin gallate (ECG) and (-)-epicatechin (EC). The test was performed inside-out patch mode at holding potential of -60mV. Unlike EGCG or quercetin, apigenin (50 μM) was not affected the TREK-2 channel activity. Epicatechin 50 μM also was not inhibited the channel activity. These results suggested that EGCG and quercetin could inhibit the channel activity among the tested flavonoids.

TREK-2 has been known to be activated with negative membrane stretch. We tested the mechanosensitivity of TREK2 in the absence of EGCG and in the presence of EGCG (25 μM). The negative pressure (-10 ~30 mmHg) was activated the TREK-2 channel activity at holding potential of -60mV and the release of pressure was recovered to basal level. And the channel activity was still remained after pretreatement of EGCG.

We tested whether TREK- 2 affects the cell proliferation since the flavonoid has been known to influence on cell proliferation and flavonoid inhibited the TREK2 channel in above results. We analyzed the cell viability in the presence of flavonoid using XTT assay. The non-transfected HEK 293 cell viability was not affected by quercetin (50 μM), EGCG (50 μM), and TEA(10mM). Contrary, TREK2 stable expressing HEK 293 cells viability was affected to 70% (n=4) by EGCG. These results suggested that the inhibition of TREK2 channel by EGCG may regulate the cell viability.

4. Conclusion

In the present study, we demonstrated that TREK2 channel inhibited by intracellular application of EGCG and quercetin and also the EGCG decreased cell viability in TREK2 stable expressing HEK 293 cells. From our results, we concluded that EGCG and quercetin represent the first known TREK2 channel inhibitor and only EGCG reduced the HEKT2 cell proliferation. It suggests that the flavonoids may work primarily by inhibiting TREK2 channel, leading to change of resting membrane potential and trigger the initiation of change in intracellular signaling for cell proliferation TREK2 channel is may, at least in part, contribute to cell proliferation.

5. References

[1] F. V. Sepulveda, L. Pablo Cid, J. Teulon, M. I. Niemeyer, "Molecular aspects of structure, gating, and physiology of pH-sensitive background K2P and Kir K^+-transport channels". *Physiol Rev*, Vol. 95, No. 1, pp.179-217, Jan, 2015.

[2] K. V. Hirpara, P. Aggarwal, A. J. Mukherjee, N. Joshi, A. C. Burman, "Quercetin and its derivatives: synthesis, pharmacological uses with special emphasis on anti-tumor properties and prodrug with enhanced bio-availability". *Anticancer Agents Med Chem*, Vol. 9, No. 2, pp.138-61, Feb, 2009.

Energy Balance of Smart Grid

Sanghyuk Lee, Kyeongsoo Kim
Dept. of Electrical and Electronic Engineering, XJTLU, Suzhou, China
{Sanghyuk.Lee, Kyeongsoo.Kim}@xjtlu.edu.cn

Abstract

Realization of entropy on fuzzy set for multiple facts has been carried out. Fuzzy entropy was realized with the help of valid distance measure, that is, commonly used Hamming distance. With the knowledge of fuzzy entropy realization, entropy design was extended to the multiple fact data. As a results, we obtained that information uncertainty was limited by the total fact (n) minus one, that is, $n-1$*. Proposed results were clarified by the clear proof derivation.*

Keywords: multiple facts; fuzzy entropy; decision making; similarity measure

Acknowledgment

This work was also supported by Centre for Smart Grid and Information Convergence of XJTLU.

Analysis The Risk factor of Death in Stomach Adenocarcinoma Patients

Jeong Ho Lee[1], Kwang Ho Park[2], Keun Ho Ryu[*]
[1,2,*)]*Database and Bioinformatics Laboratory, School of Electrical and Computer Engineering, Chungbuk National University, South Korea*
{[1]jhlee,[2]khblack,[*]khyu}@dblab.chungbuk.ac.kr

Abstract

Despite Stomach Adenocarcinoma being common, its prognosis has not been improved significantly in recent years. There are a lot of argument about the causes leading to death in Stomach Adenocarcinoma. There are many risk factors in stomach adenocarcinoma which leading death to patients. In this papers, Apriori algorithm of Association rules technique is used to analysis the factors of Stomach Adenocarcinoma patients and propose risk factor leading to death to these patients. The Experimental results showed the factors have association rules on survival and death of patients. Using these results expected to contribute to the prognosis management

Keywords: Gastric Cancer, Stomach Adenocarcinoma, Apriori, Association Analysis

1. Introduction

Stomach adenocarcinoma is a cancer that affects the stomach. The stomach is an organ of the gastrointestinal tract responsible for the digestion of food which enters it from the oesophagus and over 90% of the cancers that occur in the stomach are Stomach adenocarcinomas. This name implies that the cancer is located in the stomach where affects cells that would normally make up glands and has malignant potential [1].

Almost one million new cases of stomach cancer were estimated to have occurred in 2012 (952,000 cases, 6.8% of the total), making it the fifth most common malignancy in the world. Stomach Adenocarcinoma is the third leading cause of cancer death (723,000 deaths, 8.8% of the total). Mortality rates of Stomach Adenocarcinoma have declined since 1975. It is due to the improvement in medical skills and equipment [2]. However, most people are still diagnosed with cancer which lead to death, is crucial to identify the cause of death rather than to manage outcomes to evaluate the effectiveness of cancer treatment and survival cure rate.

Stomach Adenocarcinoma has many risk factors, including Gender, Age, Ethnicity, Geography, Helicobacter pylori infection, Stomach lymphoma, Tobacco use, Overweight, Previous stomach surgery, A family history of stomach cancer, etc. [3] However, There are a lot of controversy to identify the prognostic factors [4,5,6]. In this papers proposed if any of these factors affect the cause of death by applying Apriori algorithm.

2. Related Work

2.1. Apriori algorithm

The Apriori algorithm is a state of the art algorithm most of the association rule algorithms are somewhat variations of this algorithm [7]. It first finds the set of large 1-item sets, and then set of 2- itemsets, and so on. The number of scan over the transaction database is as many as the length of the maximal item set. Apriori is based on the following fact: The simple but powerful observation leads to the generation of a smaller candidate set using the set of large item sets found in the previous iteration. The Apriori algorithm [8] is given as follows:

```
Apriori()
L1 = {large 1-itemsets}
k = 2
while Lk−1 ≠ ∅ do
begin
        Ck = apriori_gen(Lk−1)
        for all transactions t in D do
        begin
                Ct = subset(Ck, t)
                for all candidate c ∈ Ct do
                        c.count = c.count + 1
        end
        Lk = { c ∈ Ck | c.count ≥ minsup}
        k = k + 1
end
```

2.2. Related Work

The effect of old age on the development of Postoperative Complication in surgery for Gastric Carcinoma [4] and treatment strategy of Gastric Cancer in patients older than 80 years of age [5] examined the effect of old age on the prognosis after surgery in patients who were to undergo surgery after being diagnosed with gastric cancer. This is the problem for analysis based only on the age factor.

Analysis of prognostic factor and Gastric Cancer Specific Survival Rate in Early Gastric Cancer Patients and Its Clinical Implication [6] found that lymph node metastasis was the risk factor for gastric cancer-specific survival. However, this study is analyzed only using early cancer patients' data.

In this paper, Consider the impact of complex factors about factor of death, not a single connection factors by applying association rules.

2.3. Dataset

The dataset is obtained from The Cancer Genome Atlas (TCGA). TCGA is a platform for researchers to search, download, and analyze data sets generated by TCGA. It contains clinical information, genomic characterization data, and high level sequence analysis of the tumor genomes. Most studies that use TCGA dataset prefer to use genomic information, such as methylation data or miRNA data. [9]

3. Proposed Method

3.1. Data Preprocessing

The dataset from The Cancer Genome Atlas (TCGA) is raw data. It is already normalized and composed of 62 attributes and 443 patients. However, some properties had duplicated values, absolutely no correlation attributes as data storages areas and there had many null value. Thus, data preprocessing is required to clean up the data. In order to apply Apriori Algorithm, the attribute values should be Unary attributes, Binary attributes, Missing values, Nominal attributes, Empty nominal attributes. The dataset has discretized attribute of age by 70-year-old [4] and attribute of node count by average [10]. Finally, the dataset was being preprocessed that composed of 30 attributes and 443 patients. The dataset is composed of numerous variable attributes, as shown in Table 1

Name of attribute	abbr	description
brc patient	BP	patient identification code
histologic diagnosis	HD	histological type
tumor grade	TG	neoplasm histologic grade
prospective collection	PC	tissue prospective collection indicator
retrospective collection	RC	tissue retrospective collection indicator
gender	G	gender
race	R	race
history other malignancy	HOM	other diagnosis
history nodes examined he count	HNEHC	number of lymph nodes positive by he
residual tumor	RT	residual tumor
ajcc tumor pathologic pt	ATPP	pathologic T
ajcc nodes pathologic pn	ANPP	pathologic N
ajcc metastasis pathologic pm	AMP	pathologic M
ajcc pathologic tumor stage	APTS	pathologic stage
vital status	VS	vital status
tumor status	TS	person neoplasm cancer status
history reflux disease indicator	HRDI	reflux history
antireflux treatment	AT	antireflux treatment
family history of stomach cancer	FHOSC	family history of stomach cancer
radiation treatment adjuvant	RTA	radiation therapy
treatment outcome first course	TOFC	primary therapy outcome success
new tumor event dx indicator	NTEDI	new tumor event after initial treatment
age	A	age
anatomic neoplasm subdivision	ANS	anatomic neoplasm subdivision
antireflux treatment type	ATT	antireflux treatment type
barretts esophagus	BE	barretts esophagus
h pylori infection	HPI	helicobacter pylori
targeted molecular therapy	TMT	targeted molecular therapy

3.2. Analysis

To analyze the data, WEKA [11] (Waikato Environment for Knowledge Analysis) is used. Weka is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.
This paper appointed attribute of vital status to class attribute for the result of the factor of death and analyzed dataset using Apriori algorithm in Association rules to find the complex data contained attribute of vital status.

4. Experimental results

An association rule is an implication expression of the form X → Y, where X and Y are disjoint itemsets, i.e., X ∩ Y = Ø. The strength of an association rule can be measured in terms of its support and confidence. Support determines how often a rule is applicable to a data set, while confidence determines how frequently items in Y appear in transactions that contain X.

$$\text{Support, } s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N} \quad (1)$$

$$\text{Confidence, } c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \quad (2)$$

These are the formal definitions of metrics. Confidence measures the reliability of the inference made by a rule.

Table2. Experimental Results

Rules	Confidence
PC=Y, RC=N, TS=TF, FHOSC=N → VS= A	0.99
PC=Y, RC=N, TS=TF → VS=A	0.98
PC=Y, RC=N, HOM=N → VS= A	0.98
PC=Y, RC=N, HOM=N, AMPP=M0, TS=TF → VS=A	0.98
PC=Y, RC=N, RT=R0, TS= TF → VS=A	0.98
TS=WT, HPI=N → VS=D	0.79
PC=N,TS=WT, HPI=N → VS=D	0.79
PC=N, TS=WT, BE=N → VS=D	0.79
PC=N, RC=Y, TS=WT, BE=N → VS=D	0.77
RC=Y,TS=WT → VS=D	0.72

From the results show in Table2. When attribute values of vital status are Alive, Generally, it illustrates a rule that contains the PC=Y, RC=N, HOM=N, TS=TF (Tumor Free). It means patients have treatment with a prospective rather than retrospective data, and no history of other cancer and the free status from tumor. When attribute values of vital status are Dead, it illustrates a rule that contains the PC=N, RC=Y, TS=WT (With Tumor). It means patients have treatment with a retrospective data and remaining tumor. Interestingly Both results of vital status are included in almost the same attributes. And It has an opposite data values.

5. Conclusion.

In this paper, the risk factor affect the death has been analyzed by applying association analysis. The calculation shows the result on the complex relationship factor(attribute), rather than a single factor(attribute) about the mortality. It based on the result of the analysis of this study should be expected to contribute more actively research for future prognosis management. For the future work, we will apply other association rules to the dataset. and we will analyze what factors are associated and to compare the results.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No.2013R1A2A2A01068923).

10. References

[1] Dicken BJ, Bigam DL, Cass C, et al, "Gastric adenocarcinoma: Review and considerations for future directions", *Ann Surg*, 2005, pp. 27-39.

[2] FERLAY, Jacques, et al, "Cancer incidence and mortality worldwide: sources, methods and major patterns", GLOBOCAN 2012, *International journal of cancer*, 2015, pp. E359-E386.

[3] American Cancer Society, Inc. All rights reserved. The American Cancer Society is a qualified 501(c)(3) tax-exempt organization: http://www.Cancer.org, 2016.

[4] SHIN, Dae Geun, et al, "The effect of old age on the development of postoperative complication in surgery for

gastric carcinoma", *Journal of the Korean Surgical Society*, 2005, pp. 455-458.

[5] KIM, Yong Jin, et al. "Treatment Strategy of Gastric Cancer in Patients Older than 80 Years of Age", *Journal of the Korean Surgical Society*, 2005, pp. 30-34.

[6] HYUNG, Woo Jin, et al, "Analyses of prognostic factors and gastric cancer specific survival rate in early gastric cancer patients and its clinical implication", *Journal of the Korean Surgical Society*, 2003, pp. 309-315.

[7] AGRAWAL, Rakesh, et al, "Fast algorithms for mining association rules", *Proc. 20th int. conf. very large data bases, VLDB*, 1994, pp. 487-499.

[8] AGRAWAL, Rakesh, IMIELIŃSKI, Tomasz, SWAMI, Arun, "Mining association rules between sets of items in large databases", *ACM SIGMOD Record*, 1993, pp. 207-216.

[9] Park, Kwang Ho, et al. "Classification of HighRisk Patients with Clear Cell Renal Cell Carcinoma using C4.5", FITAT, 2015, pp 115-119.

[10] KOTSIANTIS, Sotiris, KANELLOPOULOS, Dimitris. "Discretization techniques: A recent survey.", *GESTS International Transactions on Computer Science and Engineering*, 2006, pp. 47-58.

[11] WEKA projects: Data Mining Software in Java, Available: http://www.cs.waikato.ac.nz/ml/weka/, 2013.

Design of a Security Framework for Big Data

Razan Abualgasim[1], Anwar F.A. Dafa-Alla[2],
[1]*Independent researcher, Khartoum, Sudan*
[2]*Global Applied Programs (GAP), College of Applied Studies & Community Service, King Faisal University, Alahsa, KSA*
[1]*razan_gasim@hotmail.com*, [2]*Adafaalla@kfu.edu.sa*

Abstract

Today big data is generated from many sources and there is a huge demand of storing, managing, processing and querying on big data. So we have some of the technical and scientific challenges.

In big data security and privacy are managed by three Vee's of big data; Volume, Velocity and Variety. Different data source and format with the nature of data and high volume create a security challenge. Also with the increasing popularity of the big data, the security issues introduced through adaptation of this technology are also increasing.

The traditional security mechanisms which are used are reconsidered because of these big data deployments. Ability to visualize, control and inspect the network links and ports is required to ensure security also big data can be used for predictive analysis.

This paper discusses the challenge of security in big data, it design for a security framework to address security risk.

Keywords: Big data, security, privacy and framework.

1. Introduction

Many companies are using the technology to store and analyze petabytes of data about their company, business and their customers. As a result, information classification becomes even more critical. For making big data secure, techniques such as encryption, logging, honey pot detection must be necessary.

The amounts of data that are traded on a daily basis are very large and diverse. The companies, institutions and health organizations using such data for reports, manufacturing and services improved.

The diversity of data sources, formats, and data flows, combined with the streaming nature of data acquisition and high volume create unique security risks and the move of data between applications and different tiers open the door to privacy violation. Thus due to type and volume of data used in big data environment we see that, it needs different use for security tools to prevent our data. This work provides security control framework for an enterprise big data environment.

2. Definition

Big Data is the word used to describe massive volumes of structured, semi-structured and unstructured data that are so large; that means it is very difficult to process this data using traditional databases and software technologies.

Big data refers to technologies that involve data that is too divers, fast changing or massive for conventional technologies, skill and infrastructure to address efficiently. Differently the volume, velocity, and variety of data interrelation are too great. Big Data enable any organization to data creation, collection, retrieval, manage, analyze and making decision that is remarkable in terms of volume, velocity, and variety [1]. The three main terms that signify Big Data have the following properties:

i. Volume: refers to the quantity of data that is being manipulated and analyzed in order to obtain the desired results. The social media, financial institution, medical institution, sensors and logs producing data in order of terabytes every day and this amount of data is definitely difficult to be handled using the existing traditional systems.
ii. Variety: Big data comes in all types of formats emails, videos, audios, transactions etc., unstructured data type is difficult to handle with traditional tools and techniques that are not capable enough in performing the analysis on the data which is constantly in motion.

iii. Velocity: This means how fast the data is being produced and how fast the data needs to be processed to meet the demand or mine large amount of data within a pre-defined period of time.

3. Security in big data

According to [2], big data analysis is the process of applying advanced analytics and visualization techniques to large data sets to uncover hidden patterns and unknown correlations for effective decision making.

In many organizations, the deployment of big data for fraud detection is very attractive and useful.

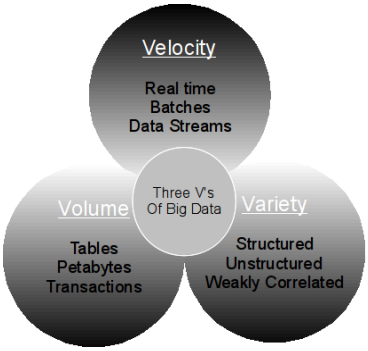


Figure 1. Three V's of big data [3]

Data security involves the encryption of the data, policies for data sharing, resource allocation and memory management algorithm.

According to [4], the top ten big data security and privacy challenges are:

1. Secure computations in distributed programming frameworks: Distributed programming frameworks utilize parallelism in computation and storage to process massive amount of data, like a MapReduce. There are two types of attack prevention measures, securing the mapper and securing the data.
2. Security best practices for non-relational data stores: Non-relational data stores popularized by NoSQL databases are still evolving with respect to security infrastructure. Developers using NoSQL database usually embed security in the middleware.
3. Secure data storage and transactions logs: Data and transaction logs are stored in multi-tiered storage media. Auto-tiering solutions do not keep track of where the data is stored.
4. End-point input validation/filtering: Many big data use cases in enterprise settings require data collection from many sources. The process of data collection represents a key challenge.
5. Real-time security monitoring: Real-time security monitoring has always been a challenge, given the number of alerts generated by devices. This problem might even increase with big data, given the volume and velocity of data streams.
6. Scalable privacy-preserving data mining and analytics: Anonymizing data for analytics is not enough to maintain user privacy.
7. Cryptographically enforced access control and secure communication: To ensure that the most sensitive private data is end-to end secure, data has to be encrypted based on access control policies. To ensure authentication a cryptographically secure communication framework has to be implemented.
8. Granular access control: Data that could otherwise be shared is often swept into a more restrictive category to guarantee sound security. Granular access control gives data mangers accessing to more granular data to be shared as much as possible without compromising secrecy.
9. Granular audits: To get to the bottom of a missed attack, we need audit information. This is because of compliance, regulation and forensics reasons. Auditing is not something new, but the scope and granularity might be different.
10. Data provenance: Provenance metadata will grow in complexity due to large provenance graph generated from provenance-enabled programming environments in big data applications.

4. Proposed framework

The following section provides the target security architecture framework for Big Data security environment. The suggested framework consists of four major areas which are:

- Access tier.
- Processing tier.
- Data management and protection tier.
- Network tier.

The above 'four tiers' of Big Data Security Framework are further decomposed into thirteen sub-components, to relieve the security risk and threat vectors to the Big Data. The overall security framework is shown below.

4.1. Access tier

To get the value of analysis big data it is important to restrict access to origin data and to gain

value creation. The three components of this tier namely Authentication, Authorization and Password policy enforcement.

i. Authentication: To solve the problem of authentication Hadoop use Kerberos. Kerberos is an ideal way to apply the concept of authentication in a big data despite some flaws, which include Single point of failure, strict time requirements, requires user accounts to all trusted relationship between clients and servers; but the bright side it gives a high percentage of authentication. This feature can be enabled by mapping the UNIX level Kerberos IDs to that of Hadoop. Kerberos is highly recommended as it supports authentication mechanisms throughout the cluster; manage user groups. Hadoop supports Kerberos as a third party.

ii. Authorization: Authorization can be enforced at varying levels of granularity and in compliance with existing enterprise security standards such as File and Directory Permissions and Role Based Access Controls (RBAC). Mapping is done in the Authentication level is leveraged by the Authorization and the users can be authorized to access data at the HDFS folder level. Authorization control with Apache Knox for column/row access restrictions for users and optionally configures Apache Accumulo if cell level restrictions are required for HBase/Hive

iii. Password policy enforcement: The password policy is a part of security awareness- the knowledge and attitude members of an organization regarding the protection of the assets of that organization.

4.2. Processing Tier

Working with new data sources brings number of analytical challenges. The relevance and severity of those challenges will vary depending on the type of analysis being conducted, and on the type of decisions that the data might eventually inform. After data collected, analysis process is come to extract information and knowledge from data. In this phase, data mining methods such as clustering, classification and association rule mining are used.

The components of this tier include Visualization, Secure computation and Logging/Audit.

i. Visualization: Due to the volume of data in big data it is extremely impossible to find anomalies using traditional methods. Dimension reduction and data projection that used in data visualization gives an abstract view to data that mean it does not get valid geometric representations to data. But the unit circle algorithm used in data visualization can map large number of data points to a unit circle that help user to make appropriate data storage and transmission decisions.

ii. Secure computation: Secure computations in distributed programming frameworks and a powerful cryptographic primitive that allows multiple parties to perform rich data analytics over their private data, while preserving each individual or organization's privacy [5]. This mechanism can implement in big data by using the MapReduce.

iii. Logging / Audit: Tiering storage means assigning different data types to different storage media, but in the growth of data capacity auto-tiering do not keep track of where the data is store. Transaction logs are sequential records for all modification happened in database, while the actual data is contained in a separate file. Audit log help to understand and monitoring big data cluster, Auditing is necessary for managing security compliance and other requirements such as Audit Data and Audit Reporting.

Scribe and LogStash are open source tools that integrate into most big data environments, as do a number of commercial products. Without actually looking at the data and developing policies to detect fraud, logging is not useful.

4.3. Data management and protection tier

Protecting data tier start by protecting data collected from different source, we mean by different source structured, unstructured and semi structured data.

Organizations, such as governmental agencies, often need to collaborate on security tasks, data sets are exchanged across different organizations, resulting in these data sets being available to many different parties. Apart from the use of data for analytics; security tasks may require detailed information about users. As a result, detailed user mobility information may be collected over time by the access control system. This information if misused can lead to privacy breaches.

Data management and protection tier components are data classification, data discovery, data tagging and input validation.

i. Data Classification: The purpose of this policy is to establish a framework for classifying and handling university data based on its level of sensitivity, value is required by the University's Information Security Plan. Classification of data will aid in determining baseline security controls for the protection of data. Effective data classification is important activities that can lead to effective security control implementation in a big data platform.

When organizations deal with an extremely large amount of data, by clearly being able to identify what data matters, what needs cryptographic protection among other and what fields need to be prioritized first for protection. The priorities of protection can be determined by classification.

ii. Data Discovery: Data discovery platform creates a foundation for data security and protocols.

Every enterprise has a set of enterprise security standards that are very rigid in terms of their data usage, data access by employees, types of data that should be accessible. Defining data within an enterprise (both internal and external) is crucial. It is important to know what kind of data an enterprise has, where it is stored, how and why is it stored there.

Data discovery tools and software for visualization, integration, data migration, etc. help enterprises identify and locate sensitive structured and unstructured information and classify them. The entire process is automated, thus preventing anomalies.

iii. Data Tagging: Data tagging make you understand the tend-to-end data flows in your Big Data environment. Data tagging are used in the era of big data to assign information to each object.

Another benefit of data tagging, it create hierarchies of access control.

iv. Input Validation: Input validation is the process of assigning semantic meaning to unstructured and un-trusted inputs to an application, and ensuring that those inputs respect a set of constraints describing a well-formed input [6].

Input validation is important for big data because it collected from multiple sources.

4.4. Network Tier

As Big Data efforts grow in scope and importance, the network will play a critical role in enabling quick, sustainable expansion while also ensuring systems are linked to existing transaction and content environments.

Packet level encryption, access control and Tokenization are the components of this tier.

i. Packet level Encryption: Encryption protects data copied from the cluster. One or two NoSql variants provide encryption for data at rest but most do not. In another words packet level encryption don't work with NoSql clusters.

Worse, most available encryption products lack sufficient horizontal scalability and transparency to work with big data. This is a critical issue. Data must be encrypted to ensure the sensitive data is end to end secure and accessible to authorized entities.

ii. Access Control: Big data access control systems require collaboration among processing domains as protected computing environments, which consist of computing units under distributed Access Control management. To enable organization monitoring roles and authorities for user's access control must implement in infrastructure layer, simplify complexity in the application space and adopt authentication and mandatory access control.

ii. Tokenization: Tokenization provides a very high level of data protection also reduce the data type and length of the original data can be preserved. In big data environment tokenization can built in Hadoop for specific function such as Personally Identifiable Information or Protected Health Information.

The final frame work also sketch in figure (2).

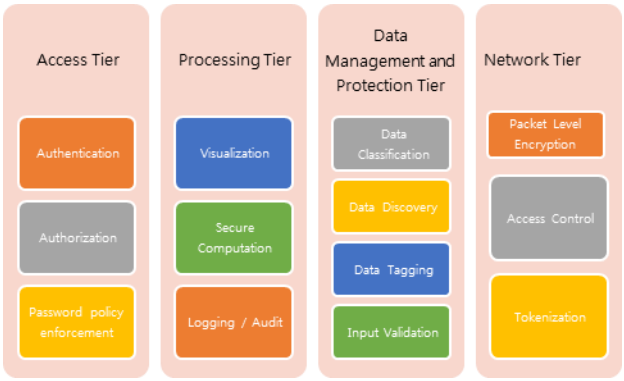


Figure 2. Proposed Framework

5. Conclusion

The paper starts by an introduction about the goal of research then provide main concepts on big data, it also provides the big data framework which can further expand and customize to the organization environment and target reference architecture around big data security.

Finally the research recommended to:

- Select the techniques and products according to organization size.
- Access control and network traffic are critical points in the big data environment.

6. References

[1] V. Sharma, N. Joshi " The Evolution of Big Data Security through Hadoop Incremental Security Model" (May, 2015) ." International Journal of Innovative Research in Science, Engineering and Technology. [online]. Vol 4. (5), pp. 3489-3493.

[2] K.Jaseena and J. David "Issues, Chalenges. And solutions: Big data mining", College Marampally, 2015, India [online] available at http://airccj.org

[3] Washington University [online]

[4] http://www.cs.wustl.edu

[5] BEAKTA, Rahul. “Big Data And Hadoop: A Review Paper”.

[6] C. Liu, X. Shaun, K. Nayak, Y. Huang† and E. Shi., Univ of Maryland and Indiana Univ. "A Programming Framework for Secure Computation",2015.

[7] SCHOLTE, Theodoor, et al. An empirical analysis of input validation mechanisms in web applications and languages. In: Proceedings of the 27th Annual ACM Symposium on Applied Computing. ACM, 2012. p. 1419-1426.

Using jointly constrained optimization to identify both recurrent and individual copy number variations (CNVs) from multisample aCGH

Peihua Chen[1], Hongmin Cai[1,*], Xi Yang[1] and Guorong Wu[2]
School of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong, China. 510006[1]
Department of Radiology and BRIC, University of North Carolina at Chapel Hill, USA[2]
E-mail: hmcai@scut.edu.cn

Abstract

Recent studies have illustrated association between copy number variations (CNVs) and particular tumor types. By the help of different high-throughput sequencing technologies, one can obtain high volume sequenced data for multiple samples with affordable price, facilitating reliable CNV identification for discrimination the difference between tumor cells and normal ones. However, copy number is highly dynamic in cancer cells due to its heterogeneousity and thus individually specific. Current advances in precision medicines advocates of identifying recurrent CNVs in samples as an indication that the tumors share the same origin and thus possibly also have common oncogene drivers and tumor insurgence. Accurate difference between the recurrent CNVs from individual one is key to explain phenotype differences as well as tumor subgrouping.

This paper present a general framework to identify and discriminate two types of CNVs, namely sample-wised individual and group-wised recurrent CNVs, from multi-sample sequencing profiles. Based on several general assumptions on the sample-wised and group-wised CNVs, the proposed model reconstructed the copy number by a convex optimization with multi-constraints. Efficient numerical algorithm to deal with huge dimensional data was designed and analyzed. Extensive experiments on both simulated and empirical datasets were conducted to demonstrate the performance of the proposed method by comparing with popular alternative methods. The nice experimental results demonstrated the superiority of the proposed method in detecting and discriminating of the two CNV patterns.

New Method to Determine Viewing Angle Analysis of Point Light Source Display

Densmaa Batbayr[1], Enkhmunkh Tumurbaatar[2], Ganbat Baasantseren[3]

[1]*Ulaanbaatar State University, Ulaanbaatar, Mongolia*
[2,3)]*National University of Mongolia, Ulaanbaatar, Mongolia*
[1]*densmaa2012@gmail.com,* [2]*enkhmunkht@gmail.com,* [3]*ganbatb@gmail.com*

Abstract

In this paper, we present a new method to determine of the viewing angle of Point Light Source (PLS) display. A three-dimensional (3-D) point appears on a cross section of collected elemental points so viewing angle is equal to an angle between two extreme rays. According to result of simulation, the viewing angle of PLS displays are depending on the position of integrated point, size and focal length of elmental lens.

Keywords: Integral image 3-D display, Point Light Source display (PLSD)

1. Introduction

An Integral imaging [1] is a 3-D display [2]. The Integral image technology has some advantages [3] that it does not require any special glasses and has continuous viewpoints [4] within the viewing angle. It also provides full parallax, full color, and real time [5]. However, conventional Integral image display has drawbacks such as small viewing angle [6–8], limited viewing resolution [9, 10] and short depth range[11, 12].

S. Jung et al. viewed that the of integral image display possible to increase the lens matrix methodology depending on the shift [6]. There are many researches who calculate the viewing angle. However, there are no researches that consider that why the viewing angle is small and depends on the position of integrated point. Thus, we proposed a method to calculate the viewing angle from the position of integrated point.

2. Point light source display

The PLS display consists of a light source, a collimating lens, a lens array, and spatial light modulator (SLM). The light source is in the focal point of the collimating lens, as shown in Figure 1. A collimated light are collected by the lens array on a focal plane of the lens array. This collected light is like a many light sources. It are named a point light source (PLS). When the SLM displays the elemental images, the lights from the PLS are modulated to create integrated 3-D image.

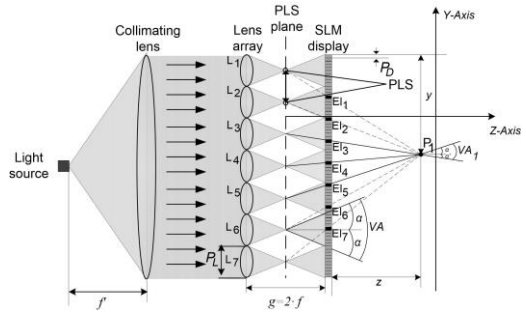


Figure 1. Structure of PLS display

The light rays from the PLS passing through the corresponding elemental image points then converge at the 3-D image point P_1, and an observer can see the point P_1. These integrated point P_1 is appearing at cross section of seven rays that are crated seven elemental points of SLM such as EI_1, EI_2, EI_3, EI_4, EI_5, EI_6, and EI_7. Note that four elemental images from EI_1, EI_2, EI_6, and EI_7 are not located just behind corresponding lenses L_1, L_2, L_6, and L_7, respectively, that they are not available. However, three elemental images EI_3, EI_4, and EI_6, are located just behind corresponding L_3, L_4, and L_6, respectively, that they are available. The integrated 3-D point of SLM from the elemental point on the following equation.

$$y = \left(i - \frac{1}{2}\right) + \left(P_L * \left(i - \frac{1}{2}\right) - P_D * i\right) * \frac{f+z}{f}, \qquad (1)$$

where y is a distance of display, i is index of elemental lens, P_L is a length of elemental lens, f is a focal of elemental lens, z is a distance of 3-D image from the lens array, P_D is pixel pitch of display.

3. To define viewing angle of PLS display

In conventional calculation, a viewing angle of PLS display [13] is given by

$$VA = 2 * \alpha = 2 * arctan\left(\frac{P_L}{2*f}\right) \quad (2)$$

From Equation (2), the viewing angle is constant and depends on the focal length and size of elemental lens. However, the viewing angles of each integrated points are different because the viewing angle depends on the position of 3-D point.

From Figure 1, 3-D integrated point appears in the cross section of diverged rays from the PLSs so a viewing angle of PLS display is determined by diverging angle of the PLS. The 3-D point P_1 is forming at the cross section of rays. The viewing angle of integrated point that is determined an angle between two extreme rays. Since α and α' are angles of two extreme rays to reconstruct P_1, the viewing angle of P_1 is equal to

$$VA = \arctan\left(\frac{y-\left(i-\frac{1}{2}\right)*P_L}{f+z}\right) + \arctan\left(\frac{\left(j-\frac{1}{2}\right)*P_L-y}{f+z}\right) \quad (2)$$

where i is index of elemental lens that districts opening extreme rays of P_1, j is index of elemental lens hat districts closing extreme rays of P_1. The viewing angle is depending on the position of integrated point, size and focal length of elemental lens as show in Equation (3).

4. Result of simulation

Table 1 shows specification of parameters for the simulation. In the simulation, we used 1 mm lens array because this lens array used in experiment.

Table 1. The calculation of the parameter.

№	Specification	Characteristic
1	Number of elemental lens	20 (W) x 20(H)
2	Size of elemental lens	1мм (W) x1 мм (H)
3	Focal length	3.3 мм
4	A distance of SLM from the lens array	2*3.3 мм

In the first, we created the elemental images on five different positions, as shown in Figure 2. The elemental point is located just behind corresponding lenses respectively, and they are available in the elemental point. From the result of simulation, when the integral image is close to lens array, the number of elemental image is fewer than farther ingenerated image from lens array.

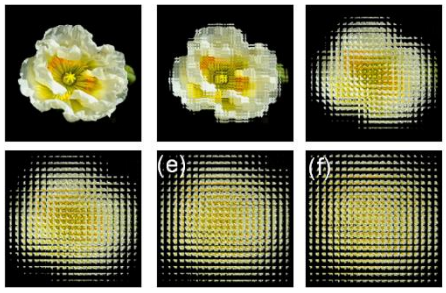


Figure 2. Elemental image by the reconstruction of PLS display. (a) 3-D image and integral image. (b) Elemental image at *z*=2.2 mm. (c) elemental image at *z*=9.9 mm. (d) elemental image at *z*=17.6 mm. (e) elemental image at *z*=25.3 mm. (f) elemental image at *z*=33 mm.

Figure 3 shows calculation of the viewing angle of 3-D point that is determined an angle of between two extreme rays. The viewing angles of 3-D point have calculated from Equation (2) in accordance with parameters of Table 1, the distance of lens array is from 0-60 mm.

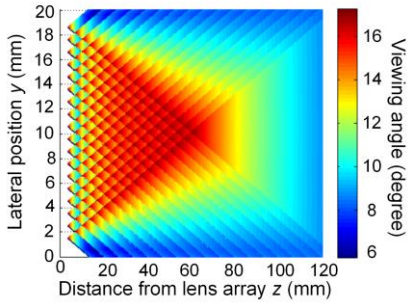


Figure 3. Calculation of the viewing angle of integrated points.

The lateral positions in the centers of viewing angles are depending on the distance from lens array, as shown in Figure 4. It is varying that viewing angle of integrated two points P_1 and P'_1. The viewing angle of integrated point that determined an angle of between two extreme rays, so that the angles depend on the distance to lessen.

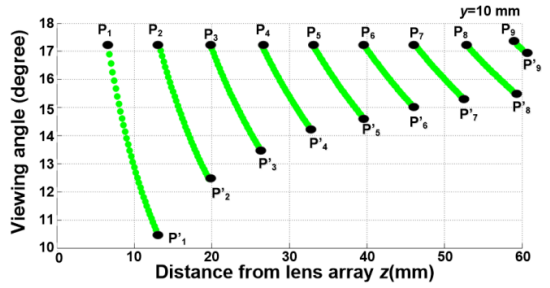


Figure 4. The viewing angle of integrated points on the plan *y*=10 mm

The viewing angles of points P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8, and P_9 are depending on position of integrated point. The new method defined are calculated in the maximum viewing angle is 17.23°. This maximum viewing angle is equal to conventional PLS display.

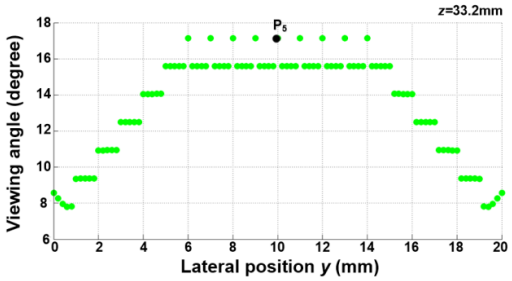


Figure 5. The viewing angle of integrated points on the plan *z*=33.2 mm from the SLM

For example, viewing angles of pixels at z=33.2 mm from SLM and lateral positions such as 6 mm, 7 mm, 8 mm, 9 mm, 10 mm, 11 mm, 12 mm, 13 mm, and 14 mm are 17.14° in Figure 5.

5. Summary

We proposed a method to the viewing angle of PLS displays are depending on the position of integrated point, the size and the focal length of the elemental lens. The viewing angle of the integrated point is determined an angle of between two extreme rays. From the results of the simulation, the maximum viewing angle is 17.23°. It is equal to conventional method.

6. References

[1] A. Stern and B. Javidi, "Three-Dimensional Image Sensing, Visualization, and Processing Using Integral Imaging," *Proc. IEEE*, vol. 94(3), 2006.

[2] N. A. Dodgson, "Autostereoscopic 3-D displays," *Computer (Long. Beach. Calif)*, vol. 38(8), 2005, pp. 31–36.

[3] D. Ting-Chung and V. Tech, Eds., *Digital Holography And Three-Dimensional Display*. New York, USA, 2006.

[4] J.-H. Park, G. Baasantseren, N. Kim, G. Park, J.-M. Kang, and B. Lee, "View image generation in perspective and orthographic projection geometry based on integral imaging," in *Optics express*, 2008, vol. 16(12), pp. 8800–8813.

[5] F. Okano, H. Hoshino, J. Arai, and I. Yuyama, "Real-time pickup method for a three-dimensional image based on integral photography," *Appl. Opt.*, vol. 36(7), pp. 1598–1603, 1997.

[6] J. H. Park, S. W. Min, S. Jung, and B. Lee, "Analysis of viewing parameters for two display methods based on integral photography," *Appl. Opt.*, vol. 40(29), pp. 5217–5232, 2001.

[7] R. Martínez-Cuenca, H. Navarro, G. Saavedra, B. Javidi, and M. Martinez-Corral, "Enhanced viewing-angle integral imaging by multiple-axis telecentric relay system," *Opt. Express*, vol. 15(24), pp. 16255–16260, 2007.

[8] G. Baasantseren, J.-H. Park, K.-C. Kwon, and N. Kim, "Viewing angle enhanced integral imaging display using two elemental image masks," *Opt. Express*, vol. 17(16), pp. 14405–14417, 2009.

[9] D.-H. Shin, B. Lee, and E.-S. Kim, "Effect of illumination in an integral imaging system with large depth of focus," *Appl. Opt.*, vol. 44(36), pp. 7749–7753, 2005.

[10] J.-S. Jang, F. Jin, and B. Javidi, "Three-dimensional integral imaging with large depth of focus by use of real and virtual image fields," *Opt. Lett.*, vol. 28(16), pp. 1421–1423, 2003.

[11] M. Kawakita, H. Sasaki, J. Arai, F. Okano, K. Suehiro, Y. Haino, M. Yoshimura, and M. Sato, "Geometric analysis of spatial distortion in projection-type integral imaging," *Opt. Lett.*, vol. 33(7), pp. 684–686, 2008.

[12] R. Martínez-Cuenca, G. Saavedra, A. Pons, B. Javidi, and M. Martínez-Corral, "Facet braiding: a fundamental problem in integral imaging," *Opt. Lett.*, vol. 32(9), pp. 1078–1080, 2007.

[13] J. Park, J. Kim, Y. Kim, and B. Lee, "Resolution-enhanced three-dimension / two-dimension convertible display based on integral imaging," *Opt. Express*, vol. 13(6), pp. 1875–1884, 2005.

Comparison of classification algorithms for the fruit yields

Jong Seon Woo[1], Youngjun Piao[2], Hyunwoo Park[3], Keun Ho Ryu[*]
[1,2,3,*)]*Database and Bioinformatics Laboratory, School of Electrical and Computer Engineering, Chungbuk National University, South Korea*
{[1]jswoo, [2]pyz, [3]hwpark, [*]khryu}@dblab.chungbuk.ac.kr

Abstract

Owing to the importance of productivity in agricultural industry, many researches have been conducted about factors that have influence on crops yields. On the Bare ground, fruits are affected by climate factors. Nowadays some of researches focus on analyzing factors by using association rule mining. In this paper, we proposed the optimal classification method to predict crop yields and this proposed method is expected that can predict fruit yields. Climate data set from Korea Meteorological Administration (KMA) and Agricultural yields data set from Korea National Statistical Office (KOSTAT) were used in this experiment. Before implement the experiment, we used feature selection to reduce unnecessary features and compared 3 algorithms that K-nearest neighbor, artificial neural network and recurrent neural network. The experimental results have shown that KNN classifier is the optimal classification method for this data set.

Keywords: optimal, predict, classification, feature selection, fruit yields

1. Introduction

Recently, researchers have shown that climate factors have influence on various industries. Research from National Oceanic and Atmospheric Administration(NOAA) announced that agriculture, plant, leisure and construction industry were closely interrelated with climate factors. [1]

Agricultural industry has been influenced by natural factors and artificial factors. Natural factors are the climate factors such as precipitation, temperature and humidity etc. Artificial factors are soil quality and fertilizer etc. Especially, meteorological factors are the key factors of agricultural productivity.

In the greenhouse human can directly managing the growing environment of crops. [2] However, many crops are yielded from the bare ground and agricultural productivity is affected by climate changes. [3]

Owing to the fact that fruits are yielded from the bare ground, previous researches have reported the key climate factors of fruit yields by association rule mining. [4]

Based on the key factors, we compared classifiers and propose the optimal classification method to predict the representative bare ground fruit apple yields.

In this paper, climate data set from Korea Meteorological Administration (KMA) and agricultural yields data set from Korea National Statistical Office (KOSTAT) were combined and used for the experiment.

The overall structure of the study takes the form of five chapters, including this introductory chapter.

Related works were discussed in chapter two. The third chapter is concerned with the methodology used for this study. The fourth section presents the data set information and the results of the experiment. Finally, the conclusion gives a brief summary in chapter 5.

2. Related work

Recently, some researches have conducted about relations between climate factors and crops yields. Especially, some of research analyze correlation between productivity and climate factors by association rule mining. [4] In this related work, they propose the key climate factors in agricultural yields. However, this previous study has not dealt with the prediction of yields.

In other research, researchers have proposed a specific formula to predict crops productivity. [5] In this research, the climate data set has only two attributes. Also this proposed formula is not applicable to other countries or regions.

In china, some researches have conducted about how climate factors effects to winter wheat and summer corns yields. [6] In this research they

proposed the important period for wheat and corns yields in specific region. However, the proposed period is not able to predict the crops yields.

3. Classifiers

3.1. Artificial Neural Network

In computer science and related fields, artificial neural networks are computational models inspired by animal central nervous systems (in particular the brain) that are capable of machine learning and pattern recognition. [7] They are usually presented as systems of interconnected "neurons" that can compute values from inputs by feeding information through the network. Like other machine learning methods, neural networks have been used to solve a wide variety of tasks that are hard to solve using ordinary rule-based programming, including computer vision and speech recognition. The word network in the term 'artificial neural network' refers to the inter–connections between the neurons in the different layers of each system. An example system has three layers. The first layer has input neurons, which send data via synapses to the second layer of neurons, and then via more synapses to the third layer of output neurons. More complex systems will have more layers of neurons with some having increased layers of input neurons and output neurons. The synapses store parameters called "weights" that manipulate the data in the calculations. [8]

An ANN is typically defined by three types of parameters: 1. The interconnection pattern between different layers of neurons 2. The learning process for updating the weights of the interconnections 3. The activation function that converts a neuron's weighted input to its output activation.

One type of network sees the nodes as “artificial neurons”. These are called artificial neural networks (ANNs).

3.2. K-Nearest Neighbor

K- Nearest Neighbor is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The K-NN algorithm is among the simplest of all machine learning algorithms. [9]

The neighbors are taken from a set of objects for which the class or the object property value is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

In the classification phase, k is a user – defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point.

3.3. Support Vector Machine

SVM is a non-parametric classification approach that can separate multimodal class distributions in high-dimensional feature spaces by using nonlinear kernel functions U, which meet Mercers conditions (Vapnik, 1998). Based on this so-called kernel-trick the n-dimensional input space is mapped into a higher dimensional Hilbert feature space. The optimization problem being solved is based on structural risk minimization (Vapnik, 1998). The strategy of SVM is to discriminate the classes by fitting an optimal separating hyper-plane (OSH) to the training data of two classes within the feature space, and to maximize the margins between the OSH and the closest training samples (the support vectors). [10] SVM only focus on the training samples that are closest to the edge of the class distributions (Mathur and Foody, 2008). Although SVM was originally designed for binary classification problems it can be extended for solving multi-class problems. Detailed description of the concept of SVM is given in Burges (1998).

4. Experimental result

4.1. Dataset

Climate data set from Korea Meteorological Administration(KMA) and Agricultural yields data set from Korea National Statistical Office (KOSTAT) were used in this experiment.

In the original climate data set, the attributes are temperature, daily temperature range, humidity, precipitation, amount of snow cover, solar radiation, sunshine, cloud cover, wind speed and air pressure. Climate data was combined with apple yields data set.

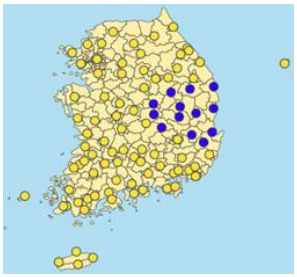


Figure 1. map of selected fruits plantation

Based on the related work [4], used related attributes were temperature, daily temperature range, humidity, precipitation and solar radiation. In the data preprocessing, yields data was labeled 1 and 0 based on the average yields value.

4.2. Performance evaluation and results

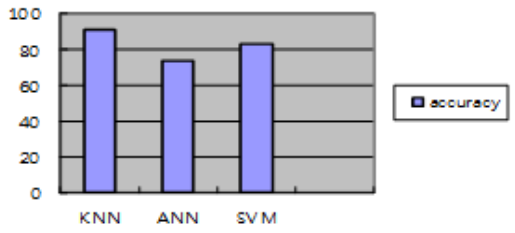


Figure 2. Accuracy of classifiers

Accuracy of classifiers using preprocessed data is presented as shown in figure 2. The results were achieved by average value of 10-fold cross validation for each classification algorithms. The KNN achieved accuracy of 91.2%. The ANN achieved accuracy of 73.5%. The SVM achieved accuracy of 82.7%.

5. Conclusion and Future work

In this paper, we suggest the optimal classification method for the fruit yields in Korea. We employ three population algorithms called KNN, ANN and RNN. We evaluated these algorithms using classification accuracy.

In our experiment, we came to conclusion that, KNN is the most suitable algorithm to predict fruit yields.

In the future work, we consider the tool of monitor and predict fruit yields can be developed based on this result.

Acknowledgement

This research was supported by Export Promotion Technology Development Program, Ministry of Agriculture, Food and Rural Affairs(No.114083-3). This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No.2013R1A2A2A01068923).

6. Reference

[1] United States Department of Commerce National Oceanic and Atmosphere Administration (NOAA), “Economic statistics for NOAA (2008).”, 2008.

[2] Jeong, W. J., Myoung, D. J., & Lee, J. H. (2009). “Comparison of climatic conditions of sweet pepper’s greenhouse between Korea and the Netherlands.”, Journal of Bio-Environment Control.

[3] Jones, P. G., and Thornton, P. K., “The potential impacts of climate change on maize production in Africa and Latin America in 2055.”, Global environmental change, 13(1), 2003, pp. 51-59.

[4] Jong, S. W., Erdenbileg Batbaatar, Keun, Ho, Ryu., (2016)., “Association rule Mining between Climate factors and Fruits yields.” Korea Computer Information Conference., 24(1), 23-25.

[5] CHANG, Ching-Cheng., “The potential impact of climate change on Taiwan's agriculture.”, Agricultural Economics, 2002, 27.1: 51-64.

[6] SHAO-E, Yang; BING-FANG, Wu., “Research on the Relationship between Meteorological Factors and Yields of Winter Wheat and Summer Maize in North China Plain.”, In: Bioinformatics and Biomedical Engineering(iCBBE), 2010 4th International Conference on. IEEE, 2010. p. 1-4.

[7] Abhishek, K., Kumar, A., Ranjan, R., & Kumar, S. (2012, July)., “A rainfall prediction model using artificial neural network.”, In Control and System Graduate Research Colloquium (ICSGRC), 2012 IEEE (pp. 82-87). IEEE.

[8] Cakir, Y., Kirci, M., & Gunes, E. O. (2014, August). “Yield prediction of wheat in south-east region of Turkey by using artificial neural networks.”, In Agro-geoinformatics (Agro-geoinformatics 2014), Third International Conference on (pp. 1-4). IEEE.

[9] Yesilbudak, M., Sagiroglu, S., & Colak, I. (2013). “A new approach to very short term wind speed prediction using k-nearest neighbor classification.”, Energy Conversion and Management, 69, 77-86.

[10] Li, N., & Liu, C. (2011, July)., “Application of SVM to the prediction of water content in crude oil. In Control, Automation and Systems Engineering (CASE).”, 2011 International Conference on (pp. 1-4). IEEE.

Development of Robotics Teaching

Yanyan Ji[1], Hui Zhang[2], Chunyan Ji[3]
[1,2,3)]*Computer Science and Technology Programme*
United International College
P.R.China
{[1]yyji, [2]amyzhang, [3]chunyanji}@uic.edu.hk

Abstract

Robot is getting more and more popular, numerous applications having been used in different aspects. The traditional robot course focus on lots chips, sensors and limited tasks which based on pure technology area. Since robotics is a multi-disciplinary subject at graduate level, in order to improve the teaching quality and make the course diversified, this paper discuss several issue related to the recent development of an undergraduate robotics course at the BNU-HKBU United International College. The course includes the principle knowledge of automatic control such as platform, power, and robot manipulators and latest technology implementation. The course culminates in a free project to combine the sensors implemented as a creative project which related to the real life problem solving and robotic competition in which only one champion team can beat the rest of teams by building an ideal robot both on the construction and coding.

Keywords: Robotics, Teaching, Microcontroller, Hardware, Software, 3D printer.

1. Introduction

Robotics course is a multidisciplinary science that contains lots knowledge from many areas like computer science, mathematics, physics, mechanical engineering, material science, electrical engineering, computer engineering, industrial engineering and manufacturing engineering. Students are forced to learn a lot to apply the engineering concepts to practical situation, not even robot tasks are sensitive to the environment, which is invaluable teaching tool for our students since they had not done lots hands on work during the high school .Some robotics subjects require a mathematical background higher than the undergraduate level, this make the main challenge in the development of robotics teaching to design and review the course organization which the complexity fit the requirement and make it acceptable and enjoyable to junior and senior students.

Robotics course has its multidisciplinary nature which provides an opportunity to remove the barriers among different areas and integrate all the knowledge in a single course, but the difficulty not only in the building technique but also the software implementation keep it out of the undergraduate level for many years. Recently advances in hardware and software tools for education greatly release the computation and programming burden on students. Notable examples of software packages for Compute Aided Robotics Education (CARE) include the Robotics Toolbox for Matlab [1], Labview [2] package provides graphical approach allows non-programmers to build programs by dragging and dropping virtual representations of lab equipment with which they are already familiar, there are different blocks such as Labview for Arduino and Labview for Mindstrom. As the development inexpensive microprocessor based control boards that also make the robotics teaching easier to teach students control small robotics devices. Examples of such controllers are the Basic stamp [3] available at low cost from Americal Parallax company, Raspberry Pi [4] single-board computers developed in England, and Open-source electronic prototyping platform Arduino [5] from Italy.

In this paper we will present a robotics course for undergraduate students developed at the Bei Jing Normal University-Hong Kong Baptist University United International College (UIC). Unlike the robotics offered in various academic institute or other universities, this one combines the principle knowledge of automatic control such as platform, power, robot manipulators and latest technology implementation. It will covers from the robotics theory design to real building techniques where students use Legos, 3D printed material and various electronic devices to make useful robot application which related to the real life.

2. Course Organization

The course title is "Introduction to Robotics" which is offered as major elective course for CST students and free Elective course for non-major students. It is 3 credits and 40 contact hours in one semester. It aims to introduce students to the concepts involved with autonomous robotic systems. The objective of this course is to use a hands-on approach to introduce the basic concepts in robotics, focusing on mobile robots. This class consists of one hours of lecture and three hours laboratory session weekly or two hours of lecture and 2 hours in the lab. The lecture part is used to deliver the fundamentals of sensors, motors and micro-controllers while the laboratory focus on hands on work to finish the tasks by implemented with the hardware which covered in the Lecture.

3. Course Content

This course is divided into two different major parts. One is the principle of the robot, which covers geometric models of robot movement, path planning, robot vision and motor control. Textbooks in support of robot path planning and robot vision include those by Farbod Fahimi [6] ,Junichi Takeno [7], Paul, R.P [8] and Foley, J.D [9].

The other part is more laboratories oriented and covers robot building techniques, microcontroller programming, sensors, circuits and 3D design. There are lots on-line source, and textbooks in support of robot building techniques include those by Gordon McComb [10],Lydia Cline [11] ,Hugh F. Durrant [12] and Marco Schwartz [13].

3.1. Part1. Principle of Robot

This part of the course is the preparation for the part2 and also covers topics which allow students for the deep study in robotics. The mainly content are listed as following:

Mathematical Concepts: Robot motion need to use the mathematical models to visualize and verify, this part is review and an introduction to trigonometric functions, vectors and matrices, and geometric transform. Matlab will used in solving numerical problems which happens in the lecture discussion and homework assignment.

Robot Path Planning: Students are required to design the robot go through the maze with the obstacles within limited time, path planning is one of the key technology in this tasks. The aim of it is to find the shortest safe path in the maze, the optimization of the path planning is acquired.

Robot Vision: The Students ae taught to design the recognizing engagement in human-robot interaction, then they are required to add the human face recognize model in the 3D printed robot.

Motor Control: Motors and Actuators make Robot move. The students are taught to the concepts of different motors like DC motors, Standard Motors, and Continuous motors. Combining those servo motors to make the movement of the 3D printed robot move smoothly.

3.2. Part2. Building Robot

This part is more focus on the hands on work, students need to implemented the knowledge which cover in the part1 and make a robot which meet the requirement of the specify tasks. Several topics covered are list here.

Mobile Robots: Different robot kits for instance like the Boe-Bot Robot [3], Robot Shield with Arduino, Activity Bot Robot Kit and Mindstorms EV3, a set of Legos with DC motors are taught and students are asked to experiment in building various shapes robots.

Microcontrollers: The Basic Stamp [3], a micro controller, PIC16C57 developed by Parallax, Inc. The board, small size and battery powered, is used in technology education and as an easy-to-program, quick to implement solution in many industries including manufacturing, process control and robotics. The Parallax PBASIC language interpreter in its microcontroller, it has easy to use commands for basic I/O, more advanced commands let the BASIC Stamp module interface with other integrated circuits, communicate with each other, and operate in networks. Students are also provided the Propeller Activity Board, which is 8-core propeller microcontroller, designed especially for STEM education, and it is well-supported with free C language programming software.

- Sensors: Students are taught the principles of the communication, usage and operation of a variety of sensors for detect the light, the distance, the gravity, the heat and Infrared Light, and then implemented those sensors to the robot tasks.
- Circuit: Electricity is almost everywhere, and the electric circuit are important for Robotics, students need to learn sufficient techniques for analyzing and designing circuits according to the Robotic activities.
- 3D Design: The 3D modelling and 3D printing happen to be greatly developed from being theoretical to a reality, which are available to produce models as well as designs for products in

recent years. Students need model and design 3D robot parts, applying architecture in build robot hardware, programming design implemented on Arduino board combined with smart sensors to make a humanoid 3D printed robot which can grab hold of things, tilt its head and move its arms around in various ways. (see the Figure 1)

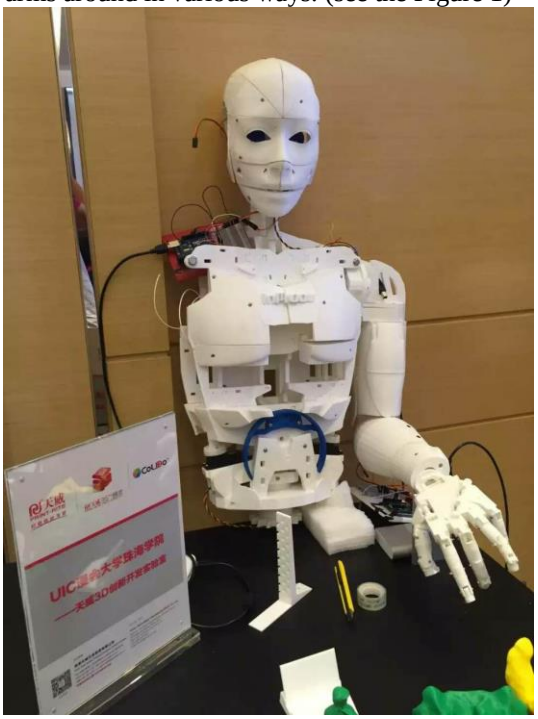


Figure 30. 3D printed Robot

- Free project and Competition: Student are required to make a free project combine the sensors which they had learnt, which motive students searching information from web to find a creative idea which is related to the real life. At the final, students are divided into teams of 3 or 4 students to do the Ping-Pang Competition. The special table and rules are released before they start design, build, and program their own robots. The class contest improves their teamwork ability and problem solving. All the learning material and research reference are also required presented in the presentation.

4. Laboratory Su pport

This course is offered by Computer Science Program Science and Technology Division, a fully equipped laboratory is provided with computers, tools, cabinet or container for the tiny parts of robotic course. Students are instructed by many different Robot kit sets for instance like the Mindstorms Lego Robots, Boe-Bot Robot, ActivityBot, and Robot Shield with Arduino. All the software like the PBASIC, Arduino, and Propeller C have installed in the Laboratory. Students are also allowed to use the 3D printer to DIY the Robot which make it more fit and stable during the Robot demonstration.

5. Student Response

The robotics class was first offered at the UIC in 2009 only for Computer Science major students. Their response to the course was overwhelmingly favorable (see the Figure 2), you can find out the interesting is most evident, coding will be most difficulty part for our students. During the robot contest, many other major students are attracted to join and see the competition. So much good comments received from many non-major students. Then the request of offering another similar robotics course for non-major students is proved by the college. Robotic course for CST students has a level prerequisite of a course in programming and mathematical which guarantees that all students enrolled are at least in their year fifth semester. Students also consider this course to be a good preparation for the FYP topic and field of study for the apply the high-degree study .Many non-major students choose this course experiment the real connection between the hardware and software to make their college study life more interesting and practicable.

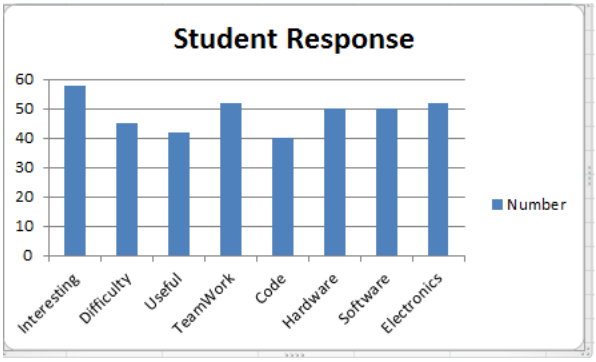


Figure 2. Student Response

6. The Ping-Pong Robotics Contest

The final Ping-Pong Robotics Contest invited all the students and staff in UIC, many friends and relatives those non-UICers also join this event, newspaper will report the contest, guests are excited to see the performance of each robots especial during the contest. It is also being a particularly enjoyable part of the Robotics class.

This Robotics contest was organized with several teams of three to four students each.15 balls on the special playing surface with 6 balls evenly spaced in the player's flat area and 3 balls at the top plateau. The round will be started by the judge turning on the starts light which located underneath the table at the center of each robot's start circle. For the beginning 5 seconds of the round. Each round has only 60 seconds and the contest will be a double elimination competition. The winner will be the side which has fewer balls. Since the contest has the limited time and robots size is designed within the requirement, it is interesting to note that the students' strategy which are not only efficient but also the flexible robot size by their talent construction design. The final competition team had the sturdiest mechanical design, the most down-to-earth approach to perceived changes in different side since the color is not the same and the simple program which implemented the efficient strategy. It is always that the worst-performing robot in the contest had the most complicated programming planned by its designers.

7. Conclusion

This paper had discussed the development of Robotics Teaching by presented the robotics course which offered for undergraduate students at UIC. Unlike the tradition robotic teachings which focus on the pure technology area; this course combines lectures on the principle of robotic with robot building techniques. In the principle of robot part, the course covers the geometric models of robot movement, path planning, robot vision and motor control. This part is an excellent introduction to general robotics and great preparation for student to pursue further study and also the Final Year Project. The robot building techniques contains series of lectures and laboratories that covers different microcontroller, circuits and programming, building lego robot combined with other robot kits like the Arduino Robot Shield, using 3D printer to design humanoid robot and integrated with Servo motors and sensors to design smoothly movement of robot. The course culminates in a class competition where teams of students compete with each other by stable robot design and free project which based on the sensors and building technology which covered to design useful and creative robot application related to the real life. The completion and the free project which offer an opportunity to apply the knowledge gained in this course and to make this course more attractable, in that case the Robotics Teaching will become more easily and efficient.

8. Acknowledgement

The authors wish to acknowledge the final support of the UIC DST-HKBU CSD Joint Research Center for Active Media Computing. Special thanks to Rachid Manseur [14] contributes to the structure of the course modelling.

9. References

[1] CORKE, Peter, "A robotics toolbox for MATLAB", *Robotics & Automation Magazine*, IEEE, 1996, 3.1: 24-32.

[2] ESSICK, John, "Hands-on introduction to LabVIEW for scientists and engineers", Oxford University Press, 2012.

[3] KUHNEL, Claus, "ZAHNERT, Klaus. Basic stamp: an introduction to microcontrollers", Newnes, 2000.

[4] DONAT, Wolfram, "Make a Raspberry Pi-Controlled Robot: Building a Rover with Python, Linux, Motors, and Sensors", Maker Media, Inc., 2014.

[5] MONK, "Simon. Programming Arduino.", United States of America: McGraw-Hill Companies, 2012.

[6] FAHIMI, Farbod, "Autonomous robots: modeling, path planning, and control", Springer Science & Business Media, 2008.

[7] TAKENO, Junichi, "Creation of a Conscious Robot: Mirror Image Cognition and Self-Awareness" CRC Press, 2012.

[8] PAUL, Richard P, "Robot manipulators: mathematics, programming, and control: the computer control of robot manipulators" Richard Paul, 1981.

[9] FOLEY, James D., et al, "Fundamentals of interactive computer graphics," Reading, MA: Addison-Wesley, 1982.

[10] MCCOMB, Gordon, "Robot builder's bonanza", McGraw-Hill, Inc., 2003.

[11] CLINE, Lydia, "SketchUp for Interior Design: 3D Visualizing, Designing, and Space Planning", John Wiley & Sons, 2014.

[12] DURRANT-WHYTE, Hugh, "Integration, coordination and control of multi-sensor robot systems (sensors, robotics)", 1986.

[13] SCHWARTZ, Marco; MANICKUM, Oliver, "Programming Arduino with LabVIEW", Packt Publishing Ltd, 2015.

[14] MANSEUR, Rachid, "Development of an undergraduate robotics course", In: Frontiers in Education Conference, 1997.

27th Annual Conference. *Teaching and Learning in an Era of Change*. Proceedings. IEEE, 1997. p. 610-612.

The System Design based on the Real-time Electricity Pricing

Zhi Yuan Chen[1], Hai Jing Jiang[2], Ding Wei[3], Tie Hua Zhou[4], Ling Wang[*]
[1,2,3,4,*)]*Department of computer science and technology, school of information engineering, Northeast Dianli University, Jilin, China*
[1]*952614992@qq.com,* [2]*haijing_103702@126.com,* [3]*649993290@qq.com,* [4]*thzhou55@163.com,* [*]*smile2867ling@163.com*

Abstract

In recent years, a type of processing mode that aims to solve the information exchange of residential and electricity market regulation based on real-time electricity price information interaction system mode is achieved as Smart Grid developed. The system which can display the exchange of online real-time electricity price not only provides the reasonable supervision for user electricity consumption but also conveniently manage electricity information for electricity manufacturer or market control terminal. This paper proposes a model that aims to explore and design the real-time electricity pricing system

Keywords: Real-time electricity, Smart Grid, System model

1. Introduction

The energy demand in most countries is increasingly growing with the development of economy and society. Energy is the important material basis which the human depends on for survival and development, but the energy will be exhausted gradually and the environment has been worsened increasingly with the process and development of human society. These issues trouble the further development of human civilization, and they remind human to improve the efficiency of energy development and utilization and to strength environmental protection at the same time. In recent years, smart grid was proposed to meet the increasing demand. However, the plot of real-time power pricing in China is still in the primary stage and has not been promoted, meanwhile there is no mature mechanism that can contribute to design the real-time electricity pricing. The price mechanism of our country is the way that government takes some money to electricity as the subsidies for residents, in which people can easily see the features that the electricity price does not fluctuate at different times. This traditional and unreasonable electricity pricing mechanism is too inflexible to handicap the reasonable utilization of power resources. The real-time electricity pricing, as an economical measure of demand-side management that can reduce energy consumption and saving sources is becoming an increasingly significant role. Theoretically, real-time power pricing can effectively enhance economic benefits, optimize allocation of resources, and save resources. Many famous foreign economists explicitly analyze and discuss the real-time electricity pricing with different aspects. And they have great hopes for economic benefits created by implementing real-time electricity pricing. Some major electricity companies get ideal effect in the process of implementing.

As for consumer, users can schedule their production and living through the real-time power pricing according to fluctuating of price which can reduce their consumption and cost up to minimum or shift their load from peak price to off-peak price. With the development of society and popularization of intellectual apparatus, such as air-conditioning, washing machine, when the price is below a certain value made in previous, these smart apparatus will charge itself or operate automatically. However, as for service provider and electricity plants, the real-time electricity can used to adjust the load in electricity use which can take full advantage of the power energy and increase the utilization ratio of equipment. More than that, the real-time power pricing can help to improve load curve of electricity consumption under the adjustment of electricity market and reduce the emission of noxious gas such as carbon dioxide and sulfur dioxide. This will be of great significance for the development of smart grid in China.

2. Related works

There are a huge number of literature about research of real-time electricity pricing at home and abroad,

Schibuola et al.[1] presented that the photovoltaic is relevantly modifying the market of electricity supply because its generation is intense and concentrated in a few hours in the daytime. And He et al. [2] pointed out that real-time electricity pricing will have great significance on the promotion of energy conservation. Consumption-shifting measures, such as thermal energy storage system, may reduce electricity costs for customers under uniform rate pricing with demand charges if consumption is shifted away from their peak-usage time [3]. Nilsson et al.[4] suggest that real-time price-based demand response(namely consumers) programs affect residential electricity consumption, and the economic and environmental consequences of the change in consumption behavior and also mention that the increasing focus on the environmental consequences of electricity production and consumption has led to discussion on how to develop new technologies, strategies, and business solutions that promote more effective and sustainable production, distribution, and consumption of electricity. Faria et al.[5] designed a novel device that it presents DemSi, a demand response simulator that allows studying demand response actions and schemes in distribution networks. It undertakes the technical validation of the solution using realistic network simulation based on PSCAD. Jie et al. [6] proposes a novel real time pricing approach to match supply with demand for smart grid. Warren [7] and Rodolfo [8] are both mention the storage system of electricity. Lujano et al. [9] studied an optimal load management strategy for residential consumer that utilizes the communication infrastructure of the future smart grid which considers predictions of electricity prices, energy demand, renewable power production. Acharjee [10] suggests a strategy for steps to implement smart grids in India.

In China, Smart Grid has been the national significant development strategy and its construction under the unified supervision and support of government is driven with combining the nation with related enterprises' power. Real-time electricity pricing system includes variety of impact factors, such as user management, peak time, climate, user satisfaction. The entire model is split into three parts, pertaining the simulation procedure, the considered real-time system and the applied control strategies respectively.

3. The system architecture of the real-time electricity pricing

Price response is a critical factor which the real-time pricing could be successful whether or not. There are some aspects that it is must be considered when the real-time electricity pricing system is designed. To start with, the users are respond to electricity price made by electricity manufacturer and can implement two-way communication between the users and the server providers, we can accurately formulate the comfort pricing system. Secondly, the user data must be fully measured so that the electricity company can be convenient to analyze the related data. The important thing is that the measured data need to be encrypted in order to ensure its security and prevent lawbreaker from stealing significant commercial data or users themselves from modifying the important data. In the last, the total framework is displayed using five small parts that directly reflect the real-time power pricing. The basic model framework is shown in Fig.1.

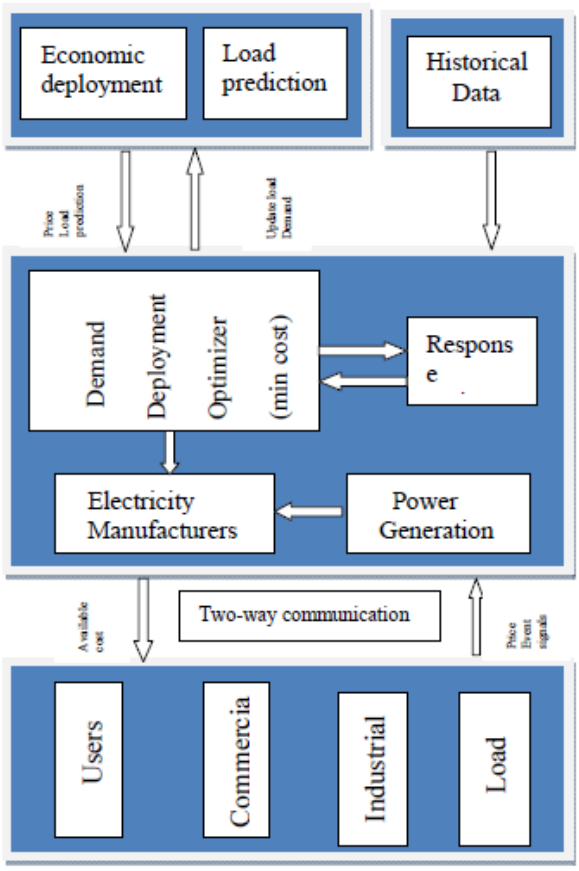


Figure.1. The total framework of real-time electricity pricing system

4. The implement of real-time electricity pricing system

In this part, we adopt B/S mode to design a virtual visual interactive interface in order to help user fast query the electricity price exchange, successfully

implementing the bidirectional communication between the consumers and the server providers.

At the end of this paper, a real operating interface to everyone will be shown in the front of the one after users log on the operating system of electricity successfully. From this operating interface, users can clearly see the electricity management, real-time display, modification of electricity, price query and changing curve of real-time pricing, electricity news and so on, Meanwhile users can be convenient to respond to the displayed information and then can take some effective measures to adjust the electricity consumption plan avoiding the peak time. This realtime power system is an original research that is not fully completed. In theory, real-time power pricing can well use the price to effectively affect the allocation of electricity power and have the most possibility of implementing the total social surplus. The operating interface of system is shown in Fig.2.

Figure.2. The operating interface of real-time electricity pricing system

5. Conclusion

The main work of this paper is to design a information interaction system of real-time electricity pricing based on real-time interaction between the user and electricity provider. Before starting this research, I summarize some experience of implementation power pricing mechanism home and abroad, read a lot of theoretical research and references of experts. Besides, I analyze some factors that can affect real-time power pricing such as user management, peak time and climate, though there are some other impact factors. So user should adjust the consumption pattern according to the actual demand electricity use, for example, user should consume the power when the load shift from peak hours to off-peak hours. In addition, another result may be that user can use the virtual interface to implement two-way communication with power provider in order to control the balance between price and supply of electricity.

Acknowledgement

This work was supported by the Education Department Foundation of Jilin Province (No.201698), by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, and by the Science and Technology Plan Projects of Jilin city (No.201464059).

6.. References

[1] L. Schibuola, M. Scarpa , and C Tambani, “Demand response management by means of heat pumps controlled via real time pricing”, Energy and Buildings, 2015, 90: 15-28.

[2] Y. He, J. Zhang, “Real-time electricity pricing mechanism in China based on system dynamics.”, Energy Conversion and Management, 2015, 94: 394-405.

[3] M. Yalcintas , W.T. Hagen, and A. Kaya, “Time-based electricity pricing for large-volume customers: A comparison of two buildings under tariff alternatives.”, Utilities Policy, 2015, 37: 58-68

[4] A.Nilsson , P.Stoll , and N. Brandt, “Assessing the impact of real-time price visualization on residential electricity consumption, costs, and carbon emissions.” Resources, Conservation and Recycling, 2015

[5] P. Faria, Vale Z, “Demand response in electrical energy supply: An optimal real time pricing approach”, Energy, 2011, 36(8): 5374-5384.

[6] J. Yang, G. Zhang, K.Ma, “Matching supply with demand: A power control and real time pricing approach” International Journal of Electrical Power & Energy Systems, 2014, 61: 111-117.

[7] P.Warren , “A review of demand-side management policy in the UK.”, Renewable and Sustainable Energy Reviews, 2014, 29: 941-951.

[8] R.Dufo-López, “Optimization of size and control of gridconnected storage under real time electricity pricing conditions”, Applied Energy, 2015, 140: 395-408.

[9] J.M. Lujano-Rojas, C. Monteiro, and R. Dufo-Lopez, et al, “Optimum residential load management strategy for real time pricing (RTP) demand response programs.”, Energy Policy, 2012, 45: 671-679

[10] P. Acharjee, “Strategy and implementation of Smart Grids in India”, Energy Strategy Reviews, 2013, 1(3): 193-204

A Data Mining Approach for Bearing Failure Prediction Using Multiple Non-linear Features

Heon Gyu Lee and Hoon Jung
Electronic and Telecommunications Research Institute Republic of Korea
{hg_lee, hoonjung}@etri.re.kr

Abstract

The main objective of this paper is to suggest a novel method for fault diagnosis of bearing using data mining technique. We also develop and then propose a novel methodology useful in developing the various non-linear features helpful in diagnosing bearing condition. Various function-based prediction models are applied in order to detect and extract those which provide the better differentiation between normal and abnormal data. In our experiments, all non-linear features are used for constructing the diagnosis model. As a result, MDA method outperformed the other models.

Keywords: Inspection module, Bearing Failure, PHM, CBM, Diagnostics, Prognostics, Fault diagnosis.

1. Introduction

Rail transport is one of the most important vehicles in the world. Train's security and comfort fall into the more and more significant matter people concerned.

Because the speed of the KTX (Korea train express) is continually increased, the bearings of trains must be more accurate and credible. The reports of Korea Railroad corp. show that the overhaul of bearing still depends on the experience of maintainers, since existing maintenance system is not enough for application and usually make mistakes. To increase the maintain work efficiency, a new maintenance system such as CBM for train bearings which should be more credible and easy use is urgently desired [1].

Condition-Based Maintenance (CBM) is a maintenance philosophy used by industry to actively manage the health condition of assets in order to perform maintenance only when it is needed and at the most opportune times. CBM can drastically reduce operating costs and increase the safety of assets requiring maintenance. Corrective/reactive maintenance can have severe performance costs, and preventive/scheduled maintenance replaces parts before the end of their useful life. CBM optimizes the tradeoff between maintenance costs and performance costs by increasing availability and reliability while eliminating unnecessary maintenance activities [2].

Electronic and Telecommunications Research Institute (ETRI) develops and implements technologies that enable CBM, including data acquisition systems, management and tracking software, and condition monitoring algorithms based on prognostic health management (PHM) and data mining.

CBM components are an optimized mix of: *i*) maintenance technologies (diagnostics, prognostics), *ii*) reliability-centered maintenance (RCM) - based processes, and *iii*) enablers (total asset visibility. The CBM process can be applied to maintain activities in all industries, including weapons systems, jet engines, Wind turbine generators, Marine diesel engines, Circuit card manufacturing, and train bearing.

In this paper, our goal is to propose quantitative measures for bearing fault along with a suitable prediction method to enhance the reliability of examination and treatment for maintenance of trains units. To achieve this goal, the proposed method, shown in Figure 1, works in two parts:

We, first, defined and extracted non-linear feature vectors from bearing vibration signal. Then, we applied prediction methods to predict faulty bearing having abnormal conditions such as outer race fault. For the prediction step, we evaluated several supervised learning methods to select a suitable prediction method. We tested the function-based prediction models to validate these methods' accuracy for diagnosing bearing conditions. Experiments also indicated that with proper classification methods, the results of diagnosis could be improved.

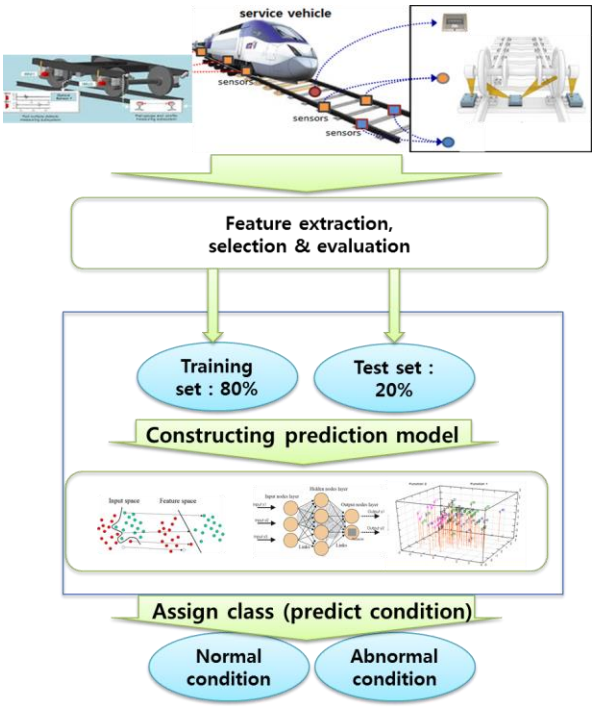


Figure 31. A flowchart summarizing individual steps for diagnosing bearing fault.

Moreover, this paper presents the development of a fault diagnostic system based on CBM that detects the operating condition of train units, such as the bearing temperature, vibration, and gear oil deterioration, and notifies the operator of potential problems or abnormal conditions. In order to CBM in railway maintenance field, we also have developed algorithms for CBM solving the complex problems of process optimization with complex input/output relationships, pattern recognition with incomplete data and anomaly detection for earliest indications of adverse performance shifts.

2. Data Preprocessing

Every vibration signal sampled is a one dimension discrete dataset $(y_1, y_2, \cdots\cdots, y_n)$ where n is the length of the vector.

Wavelet transform is an important signal processing method to extract signal features because of its ability in multi resolution analysis. After pretreatments, non-stationary signal analysis tool wavelet packet transform is taken and reconstruct wavelet packet coefficients in typical frequency bands which contain useful information to eliminate background noise. For time domain analysis is intuitive and accurate, ten parameters, including average value, effective value, maximum peak, square root amplitude, variance, skewness, kurtosis, peak factor, waveform factor and pulse factor, which are more sensitive than other time domain parameters, are selected (see reference [3]).

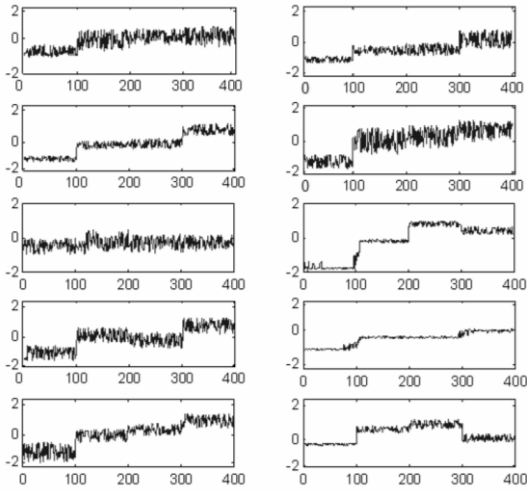


Figure 32. An example of time domain feature.

At that time, different signals should have different structure. It is realistic to presume that signal also contains non-linear features due to the complex regulation mechanisms controlling it. The various non-linear features can be evaluated by extracting regularity indexes. We explain these non-linear features in Section 3.

3. Feature Vector Extraction

3.1. Non-linear Features

The various nonlinear characteristics of acoustics and vibration of bearing can be evaluated by extracting regularity indexes and analyzing fractal scaling [3].

Approximate Entropy (ApEn): Defined as the rate of information production, entropy quantifies the chaos of motion. *ApEn* quantifies the regularity of time series, so is also called a "regularity statistic". It is represented as a simple index for the overall complexity and predictability of each time series. In our study, *ApEn* quantifies the regularity of the time interval. The more regular and predictable the time interval series, the lower will be the value of *ApEn* [4]. First of all, we reconstructed the time interval time series in the *n*-dimensional phase space using Takens theorem. Takens suggested the following time delay method for the reconstruction of the state space as follow:

$$D_t = [RR(t),\ RR(t+\tau),\ ...,\ RR(t+(n-1)\tau)],$$

where *n* is the embedding dimension and is the time delay. In this study, the optimal value of was 10. The mean of the fraction of patterns with length m that resemble the pattern with the same length beginnings at interval *i* is defined by

$$\Phi^m(r) = \frac{1}{N-m+1}\sum_{i=1}^{N-m+1} \ln\left[\frac{number\ \ of\ \ |D_m(j)-D_m(i)|<r}{N-m-1}\right]$$

In the above equation, $D_m(i)$ and $D_m(j)$ are state vectors in the embedding dimension, *m*. Given *N* data points, we can define *ApEn* as

$$\mathrm{ApEn}(m,r,N) = \Phi^m(r) - \Phi^{m+1}(r),$$

where *ApEn* estimates the logarithmic likelihood that the next intervals after each of the patterns will differ. In general, the embedding dimension, *m*, and the tolerance, *r* are fixed at *m* = *2* and $r = 0.2 \times SD$ in time series data.

The Hurst Exponent (H): Hurst Exponent *H* is the measure of the smoothness of a fractal time series based on the asymptotic behavior of the rescaled range of the process. For a time series data of length *N* {*u(n), n=1, ..., N*}, where $u(n) \equiv R(n) = t(R_{n+1}) - t(R_n)$ is the n^{th} *RR*(time) intervals defined by difference in time position for R-peaks. Running means $\bar{u}(n)$ for given *n*, and accumulated deviations from the mean $X(l, n), l = 1, ..., n$: are calculated as follows.

$$\bar{u}(n) = \frac{1}{n}\sum_{k=1}^{n} u(k),$$

$$X(l, n) = \sum_{k=1}^{l} [u(k) - \bar{u}(n)]$$

The range *R(n)* is the distance between the minimum and the maximum value of *X*, and is rescaled by dividing it by the standard devotion *S(n)*.

$$R(n) = \max_l X(l, n) - \min_l X(l, n),$$

$$S(n) = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(u(k) - \bar{u}(n))^2}$$

The rescaled range $\left(R/S = \frac{R(n)}{S(n)}\right)$ is a dimensionless quantity. The Hurst Exponent *H* is defined as,

$$\frac{\log(R/S)}{\log(T)},$$

where *T* is the duration of the sample of data. If *H* = 0.5, the behavior of the time series is similar to a random walk. If *H* < 0.5, the time series covers less distance than a random walk. But if *H* > 0.5, the time series covers more distance than a random walk [5].

Exponent α of the $1/f$ Spectrum (f_α): Self-similarity is the most distinctive property of fractal signals. Fractal signals usually have a power spectrum of the inverse power law form, $1/f^{\alpha}$, where *f* is frequency, since the amplitude of the fluctuations is small at high frequencies and large at low frequencies. The exponent α is calculated by a first least-squares fit in a log-log spectrum, after finding the power spectrum from *RRIs* (RR time intervals). The exponent α is significant because it has different values for normal and abnormal units [6].

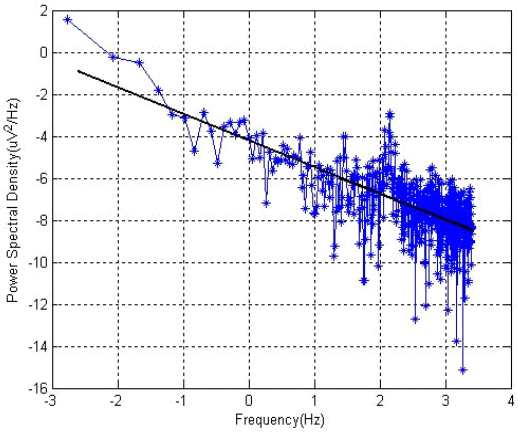


Figure 33. Frequency versus power for estimate 1/f scaling coefficient.

All features used in our study are summarized in Table 1 and individual steps of recording and processing the signals in order to obtain the various non-linear features for diagnosing bearing failure are shown in Figure 3.

Table 6. Description of non-linear features

Feature	Description
ApEn	Approximate Entropy
H	Hurst Exponent
f_α	Exponent α of the *1/f* Spectrum

3.2. Feature Selection and Evaluation

Because of the negative effect of irrelevant features on most classification, it is common to precede learning with a feature selection to eliminate all but the most irrelevant features. Feature selection

[7] consists of feature ranking and selecting steps. In the feature ranking step, all features are sorted and assigned rank. The selecting algorithm assesses the predictive ability of each feature individually and the degree of redundancy among them, preferring sets of features that are highly correlated with the class but have low inter-correlation. After all features of signal were extracted, we performed feature selection and evaluation using ANOVA *F*-test [8]. Feature ranking considers on feature at a time to see how well each feature alone predicts the target class. The features are ranked according to a user-defined criterion. Available criteria depend on the measurement levels of the target class and feature. In the feature selection problem, a ranking criterion is used to find features that discriminate between healthy and disease patients. The ranking value of for each feature is calculated as $(1-p)$, where p is the p-value of appropriate statistical test of association between the candidate feature and the target class. All features are continuous-valued, we use p-values based on F-statistics. This method is to perform a one-way ANOVA F-test for each continuous feature.

Let C be a target class with J categories, N be a total number of cases and X is the feature under consideration with I categories. The p-value based on F-statistics is calculated by

$$\text{Prob}\left(F(J-1, N-J) > F\right), \left(F = \frac{\sum_{j=1}^{J} N_j (\bar{x}_j - \bar{\bar{x}})^2 / (J-1)}{\sum_{j=1}^{J} (N_j - 1) s_j^2 / (N-1)} \right)$$

where N_j is the number of cases with $C=j$, $\bar{x}_j$ is the mean of feature X for target class $C=j$, s_j^2 is the sample variance of feature X for class $C=j$, $\bar{\bar{x}}$ is the grand mean of feature X and $F(J-1, N-J)$ is a random variable follows a F-distribution with degrees of freedom $J-1$ and $N-J$. If the denominator for a feature is zero, set the p-value=0 for the feature. Features are ranked by p-value in ascending order. In this study, any p-value less than 0.05, significant test threshold, was accepted as significant. A feature relevance score $(1-p)$ is calculated. The features having values less than 0.95 mean that they have low score and therefore they are removed. Afterwards, this subset of features is presented as input to the classification methods. We perform feature selection once for each dataset and then different classification methods are evaluated. The results of feature selection and evaluation are described in Table 2.

Table 2. Result of feature selection and evaluation.

Rank	Feature	Relevance sore (1-p)
1	f_α	0.985
2	H	0.965
3	$ApEn$	0.955

4. Function-based Bearing Fault Detection Model

SVM (support vector machine) is an useful technique for classification and the topic on the entire family of kernel based learning methods has developed into a very active field of machine learning research. *SVM* is essentially a two-class classifier, although the classifier can be extended to multiclass classification. The goal of *SVM* is to construct a model which predicts target value of data instances in the testing set. It has been shown that *SVM* is consistently superior to other supervised learning methods [9]. The *SVM* generates input-output mapping functions from a set of labeled training data. The mapping function can be either a classification function or a regression function. For classification, nonlinear kernel functions are often used to transform input data to a high-dimensional feature space in which the input data becomes more separable compared to the original input space. Maximum-margin hyper-planes are then created. The produced model only depends on a subset of the training data near the class boundaries.

Given a training set of class label pairs $(x_i, y_i),\ i=1,...,l$ where $x_i \in R^n$ and $y \in \{1,-1\}^l$, the support vector machine requires the solution of the following optimization problem .

$$\min_{w\,b\,\xi} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{l}\xi_i$$

$$\textit{subject to}\quad y_i(\omega^T\phi(x_i)+b) \geq 1-\xi_i,\ \xi_i \geq 0$$

where vectors x_i are mapped into a higher dimensional space by the function ϕ. Then *SVM* finds a linear separating hyper-plane with the maximal margin in this higher dimensional space. $C>0$ is the penalty parameter of the error term. Moreover, $K(x_i, x_j) \equiv \phi(x_i)^T\phi(x_j)$ is called the kernel function. For our experiments, we applied the sequential minimal optimization (*SMO*) algorithm by using the *RBF* (radial basis function) kernel for training a support vector classifier.

An artificial neural network (ANN) is useful to consider complicated nonlinearity, while a multilayer perceptron (MLP) NN is currently utilized for time series prediction. An ANN model consists of learning, parameter coordination, verification, and forecasting steps. At the learning step, the structure of the NN is determined by learning the nonlinear relationship between input and output variables using the backpropagation algorithm. The verification stage attempts to predict using the structure determined by learning and minimizes the error with ANN model learning. The accuracy of forecasted wind power patterns is verified by analyzing the performance error with mean absolute error (MAE). In the study, an MLP model provided by Java WEKA [9] was used. The nodes in this network are all sigmoid (except for when the class is numeric, in which case the output nodes become unthreshold linear units).

MDA (Multiple Discriminant Analysis) is an analysis of dependence method that is a special case of canonical correlation [10]. With more than two groups, there will potentially be more than one discriminant function that can be used to explain the differences among groups. For example, if we want to discriminate among three groups, two canonical discriminant functions will be derived. The first discriminant function separates group 1 from groups 2 and 3, and the second discriminant function separates group 2 from group 3. In addition, we can obtain classification function for prediction through the discriminant analysis. Classification function is generated for each group. If new case we have to classify comes into existence, this subject will belong to a group that has the highest value of classification function.

5. Experimental Results

In this section, we describe our experiments in building bearing failure detection model on the dataset from the MFPT bearing data [11]. This dataset has been provided by MFPT and the original data source can be found at [12]. A bearing fault dataset has been provided to facilitate research into bearing analysis. The dataset comprises 7 sets of data. The first 4 sets of data come from a bearing test rig with: baseline (good condition bearing), an outer race fault, outer race fault with various loads and inner race fault with various loads. The next 3 sets of data are from real-world faults, being from: an oil pump bearing, Intermediate speed bearing and a planet bearing.

The present study used evaluation measures such as the precision, recall, F_1-value, and accuracy to evaluate the three classification algorithms. Formal definitions of these measures are given [13]. To evaluate classification performance w.r.t. the number of instances and class labels, we used a confusion matrix. In this experiment, we consider 3 classes, baseline, outer race fault and inner race fault because outer race fault with various loads data have too many data tuples with imbalanced class distribution. Table 3 records the accuracy of classifiers used in a confusion matrix.

Table 3. Confusion matrix for classification models.

Actual Class		Predicted Class		
		baseline	*outer race fault*	*inner race fanult*
SVM	*baseline*	**75%**	25%	0%
	outer race fault	4%	**93%**	3%
	inner race fanult	58%	12%	**30%**
ANN	*baseline*	**75%**	20%	5%
	outer race fault	7%	**87%**	6%
	inner race fanult	50%	8%	**42%**
MDA	*baseline*	**92%**	8%	0%
	outer race fault	4%	**86%**	10%
	inner race fanult	35%	15%	**50%**

Also, four performance measures of classifiers are presented in Table 4 and Figure 4.

Table 4. A description of summary results.

Classifier	Precision	Recall	F_1-value	Class
SVM	0.748	0.873	0.805	*baseline*
	0.688	0.550	0.611	*outer race fault*
	0.647	0.423	0.512	*inner race fanult*
ANN	0.809	0.873	0.881	*baseline*
	0.769	0.750	0.937	*outer race fault*
	0.579	0.423	0.900	*inner race fanult*
MDA	0.88	0.863	0.871	*baseline*
	0.822	0.925	0.871	*outer race fault*
	0.565	0.500	0.531	*inner race fanult*

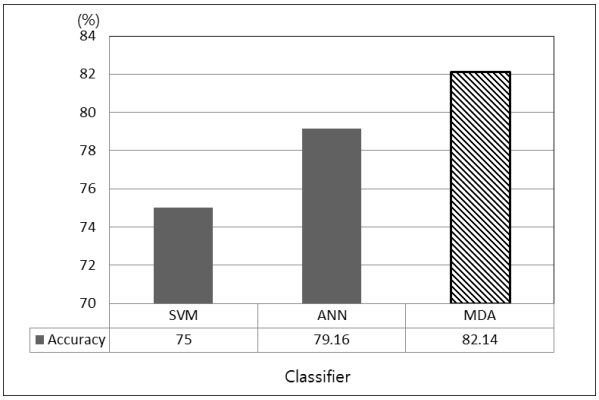


Figure 4. Accuracy comparison.

6. Conclusion

The aim of this paper was to develop an accurate and efficient method to automate the process of predicting bearing failure. Direct application of existing classification methods provide satisfactory accuracy but takes prohibitively long training time to be used in practice. Since MDA classifier were shown to be more accurate than other existing methods.

Future research will use the suggested data mining techniques on real big data related to maintaining collected over years, analyze the results,

7. References

[1] P. Li, F. Kong, and L. Dang, "A New Fault Diagnosis System for Train Bearing Based on PCA and ACO," *Int'l Conf. on Logistics Systems and Intelligent Management*, pp. 526-530, 2010.

[2] T.M. In, et al, Strategy Plan for the Efficiency of Railway Operation and Maintenance, KORAIL, Korea, Nov. 2013.

[3] H.G Lee, et al, "Coronary artery disease prediction method using linear and nonlinear feature of heart rate variability in three recumbent postures," Journal of Information Systems Frontiers, Vol. 11, pp.419-431, Sep., 2009.

[4] S. Pincus and W. Huang, "Approximate entropy: Statistical properties and applications," *Communication in Statistics-Theory and Methods*, 1992.

[5] S. Katsev, I. L'Heureux, "Are Hurst exponents estimated from short or irregular time series meaningful?" *Computers & Geosciences* 29, pp.1085-1089, 2009.

[6] E. Milotti, "1/f noise: a pedagogical review," 2002.

[7] A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research 3*, pp.1157-1182, 2003.

Making Virtual Tour Suitable for Oculus Rift

Javkhlan Rentsendorj[1], Baatarbileg Altangerel[2], Oyun-Erdene Namsrai[*]
[1,2,*)]*National University of Mongolia*
{[1]javkhlan, [2]baatarbileg, [*]oyunerdene}@seas.num.edu.mn

Abstract

Virtual tour (VT) is a prominent problem in virtual reality (VR) research. VR is a technology for experiencing three dimensional computer graphics, since late 2012 the Oculus Rift gave VR a new boost.

Through open source development and an enthusiastic active community many new opportunities have arisen, for example graphic interfaces aimed for browser based immersive experiences.

We introduce Virtual Tour (VT) for the web using the Oculus Rift. It can bring the feeling of reality and immersion.

Keywords: virtual reality, virtual tour, oculus rift, ocurus, panorama, immersion

1. Purpose, Context and History

1.1 Purpose

Virtual tour has been defined as a "simulation of an existing location, usually composed of a sequence of videos or images" [1] and provides and natural, intuitive way of human-computer interaction. Most web pages already have two of these three properties but miss the immersive quality or sense of presence in a virtual space. To accomplish this, a total field of view in the VT environment is needed, and sensor data on head behavior and body movements are desirable (see Figure 1).

Right now we browse the websites today in two dimensions, we click through links, pages and create bookmarks and taps.

The Oculus Rift is a head mounted device that enables a user to interact with 3D virtual environments in a natural way, and is for this reason suitable for experiencing virtual tour content in a web browser.

Oculus Bridge is a WebSocket based plugin to access tracking data provided by the Oculus Rift and manage the display configuration to view virtual reality online content. It is easy to install and implement in your own JavaScript code. [2]

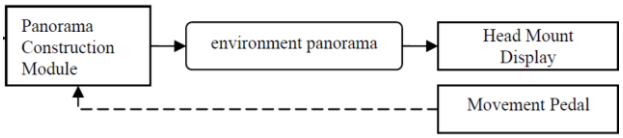


Figure 1.

1.2 Context

The Oculus Rift is a head mounted display (HMD) (see Figure 2) for experiencing virtual reality environments. It is founded by Palmer Luckey. He was moderator of Meant to Be Seen 3D forum and developed, with the help of this forum the first inexpensive Virtual Reality headset.

Figure 2.

The first Oculus Rift developers kit was financed by Kickstarter, backers of 300 dollar or more received this developers kit. The campaign raised 2.4 million dollar, and was bought by Facebook earlier 2014, this gave the large VR content developers the confidence to invest in the upcoming medium according to Oculus' CEO Brendan Iribe.

Mark Zuckerberg posted on Facebook in March when he announced the acquisition: "Imagine enjoying a courtside seat at a game, studying in a classroom of students and teachers all over the world, or consulting with a doctor face-to-face by putting on goggles in your home." [3]

Making virtual tours are signs of the technology finally having reached a level that allows comfortable, exciting VR experiences at home. Next to new immersive games, VR applications beyond our imagination for many different industries ranging from art to healthcare and military to education could come into being in the next few years.

Because of the level of our inseparability with the internet will only increase in years to come, it is important that online content can handle these new VR technology devices or are even coded intended for this medium. We think browsers in the future will be programmed to support VR devices, but for now we can use software like Oculus Bridge for experiencing virtual tour and applications in our browsers.

1.3 History of Virtual Tour for the Web

Development of VT for the web actually coincides with the development of three dimensional graphics for the web.

One of the first attempts was VRML (Virtual Reality Modeling Language) [5] which was a file format for three dimensional graphics designed specifically for the web.

The origin of the term 'virtual tour' dates to 1994. The first example of a virtual tour was a museum visitor interpretive tour, consisting of 'walk-through' of a 3D reconstruction of Dudley Castle in England as it was in 1550. This consisted of a computer controlled laserdisc based system designed by British-based engineer Colin Johnson.

2. Strengths and Weaknesses

2.1 Browser-based VT applications versus Client-based VT Applications

This section assesses the strengths and weaknesses of browser-based VT applications compared to client-based VT applications.

2.1.1. Strenghts

- Browser-based applications run in the browser and, in case of the Oculus Bridge, with the use of WebSockets. WebSockets are supported by all major browsers; Google Chrome, Firefox, Safari, Internet Explorer and Opera [6]. Additionally WebGL is supported by all major browsers.
- Applications running in the browser do not need extra extensions or custom drivers.
- Browser or web-based applications are easily updated; updates are pushed from the server side, thus no patches or expansions are required and all updates happen unnoticed.
- A browser-based application requires only a single
- code base for all operating systems e.g. HTML CSS and JavaScript.
- Applications running in the browser are not fixed to one machine, this way they can be run in any machine or device and accessed from any location.
- The same browser functions as platform for many
- different applications (ecommerce, booking, hotel, restaurant, games, social networking etc.).

2.1.2. Weaknesses

- The Oculus Bridge SDK only works if other applications using the Oculus Rift are turned off. This seems to be a problem of the Oculus Rift SDK [7].
- In general applications tend to run faster on the clientside. Although JavaScript, with its possibility to run on the client-side, has some benefits, a client-based application will perform better.
- Browsers will still need updates and browsers will not provide support

2.2 Natural browsing versus VT browsing

This section assesses the strengths and weaknesses of natural browsing (the way we are used to 'browse' today) compared to VT browsing. VT browsing being the use of a HMD and VT for internet browsing.

2.2.1. Strengths

- 'Browsing' the web is one of the main activities of the internet today. This activity however is completely two dimensional and screen based. The advent of VT for the web enables this activity to become three dimensional; it enables two dimensional information to be communicated in three dimensions
- Websites become three dimensional environments instead of two dimensional pages.
- VT browsing enables the viewer to see connections between information sources in 3D which could make the connections more clear [20].

2.2.2. Weaknesses

- Naturally, VT for the web will require getting used to. We are not used to three dimensional web pages. Information is always displayed on a two dimensional display.
- Particularly the positioning of text will need some consideration since we are used to reading from two dimensional surfaces.

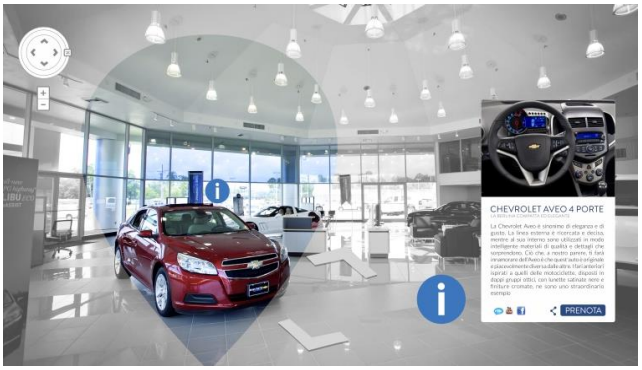
- We are not used to wearing a HMD while interacting with computers. Will we have to wear this all the time?

4. Typical Categories of VT

Head Mounted Devices offer an almost infinite number of possibility in a wide variety of fields. This section discusses some of the typical virtual tours and possible future development.

4.1. Point of Interest or E-Commerce

By interacting with the point of interest a descriptive profile will be opened, where you can include texts, photos and customized content, such as the price of the element that you're looking at. In a car showroom, each car can represent a point of interest, allowing you to find more detailed information and the reseller's contact. Same for furniture shops, clothing shops, etc.

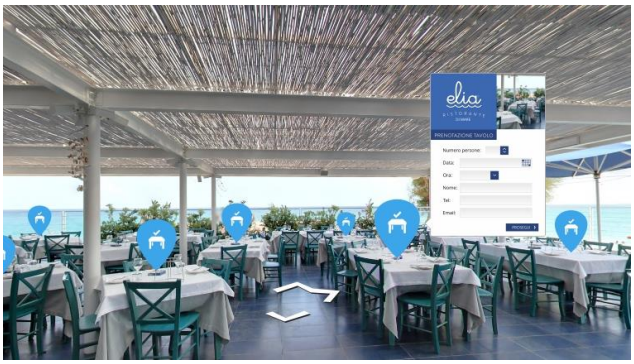


4.2. Areas of Interest

You can move through the rooms, going from an area of interest to another. The author of the tour will be able to include in any area of interest photos and customized contents. According to the various situations new elements can be inserted. The area of interest elements are shareable through social media. Passing from an area of interest to another the information changes automatically, for a fluid and continous path.

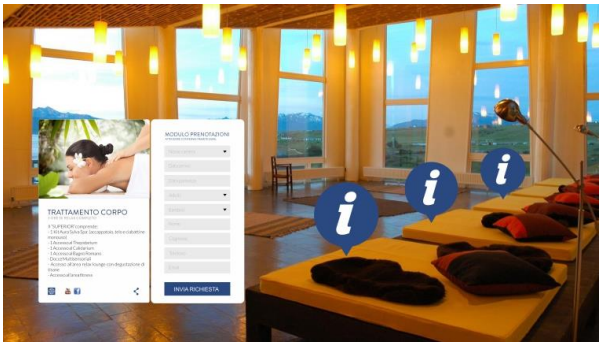
4.3. Table Booking

VT can handle also accomodating structures like restaurants, cafes, pubs, sandwich shops but also spa and beauty shops. It allows the user to book exactly the table he prefers or the kind of service that he'd like to get.



4.4. Hotel / Spa

The possibility to choose a booking service is offered as well. The default model contains informations about the room or the service that will be booked and can be customized by choosing various modules. Hotel room reservation is increased more and more nowadays.[10]



5. Development

We're creating virtual tour suitable for Oculus Rift with following application. Pixtra Panostitcher v1.5, Ocurus 1.0, Adobe Photoshop and Sony Sound Forge 8.0.

Pixtra Panostitcher was used to assist photos stitching process. It helped developer to stitch the images and also allowed developer to manually adjust the overlap between photos during stitching process.

Adobe Photoshop was used to adjust the color balance and the brightness of the panorama.

Finally, the Ocurus system was used to develop the application by importing the panoramas, integrating the sounds and adding navigational features with HMD.

5.1. Steps

The prototype was developed by adapting the five steps defined by Chen [18].

5.1.1. Node Selection. Node has been selected to maintain visual consistency when moving from one point to another. The camera was mounted on a tripod and centered at its nodal point. Then the camera was rotated in 360 degree of x-axis from the start point to the end point. Each image must have overlap with the start point.

5.1.2. Photo Stitching. The purpose of stitching is to create an ideal panoramic image from a set of overlapping pictures. In producing a good panoramic image, 50 percent overlap pictures is needed because the adjoining pictures may have a very different brightness level. But it also may vary depending on the image features in the overlapping regions. For a normal stitching session, the pictures were stitched automatically and some pictures were stitched manually by adjusting some points as remarks that were used to adjoin the pictures.

5.1.3. Image Compression. After stitching and editing process were completed, the panoramic image was resized to the optimal size that can be rendered by the application

5.1.4. Interface Design. Before developing the application, the template has been chosen. Some hand drawn sketches were made to guide the design process. There were some adjustments made to the template based on the sketches.

5.1.5. Hotspots Marking. Hotspot (hot area) identifies regions of a panoramic image for interactions, such as navigation or activating actions. The hotspot image does not need to have the same resolution as the panoramic image. By clicking on the hot spot mark, this application will bring user to the hotspot region. Arrow symbols were elected as indicator where this application will bring user to the hotspot areas once it is clicked.

5.1.6. Panoramic Linking. The linking process connects view orientation between adjacent panoramic nodes. The links were attached to a hot spot so that the user may activate the link by clicking on the hotspot. In linking process, besides creating a link between panoramic images, transition effect from the current scene to the hot spot scene can also be selected. In addition, transition duration was set up in this phase.

5.1.7. Add an OBJ importer. Three.js offers many ways of generating and importing three dimensional geometry. One way is by importing .obj files. OBJ is a for the most part universally accepted 3D geometry file format; you can export .obj files out of most 3D modeling software. In order for the OBJ import to work we will need to add the OBJLoader library[5] to the HTML file. Save OBJLoader.js to the oculus-bridge-master/examples/lib/ folder and add this line to first_person.html:

```
<script src="lib/OBJLoader.js"></script>
```

Next we will need to add a 'models' folder to oculusbridge-master/examples/ and add a .obj file (in this case name it: "helloOBJLoader.obj"). Finally we will need to add a couple of lines to the first_person.js JavaScript file found at oculus-bridge-master/examples/js/first_person.js. Inside the initGeometry() function add these lines:

[5] This library can be found here : http://threejs.org/examples/js/loaders/OBJLoader.js.

```
var loader = new THREE.OBJLoader();
loader.load('models/helloOBJLoader.obj', function (object) {
    object.traverse( function (child) {
        if (child instanceof THREE.Mesh) {
            child.material.color.setRGB (1, 0, 0);
        }
    });
    scene.add( object );
});
```

This imports a red colored version of your OBJ model! Now you can import any object file into your browser based three dimensional world.

6. Final Thoughts

Oculus Rift will give us many new application opportunities, among which the development of three dimensional graphics for the browser and special applications for Virtual Tour.

According to Mark Zuckerberg, founder and CEO of Facebook after the acquisition in March 2014: "Oculus has the chance to create the most social platform ever, and change the way we work, play and communicate." [22]

We implemented a virtual tour system based on stitched picture. This system is robust, adaptive and easy operating. By using panoramic photos and utilizing webGL library, we generate a Virtual Tour by the syste, and browse the virtual scene on the cylindrical panorama.

Of course, a full virtual tour system need other components. Future works may include: constructing a sound feedback system by picking the sound information at several points of the tour.

We are looking forward to experience applications using the Oculus Rift together with other intuitive input devices such as the Microsoft Kinect and the Motion Leap, we think this will make a big difference for the ease of use and enjoyability of VT systems. The future of VR and VT will be literally limited by our imagination.

7. References

[1] PIMENTEL, Ken; TEIXEIRA, Kevin, "Virtual reality through the new looking glass", 1993.

[2] Oculus Bridge, "A utility and javascript library to link the Oculus Rift with the web." https://github.com/Instrument/oculus-bridge

[3] Rubin, Peter. "The Inside Story of Oculus Rift and How Virtual Reality Became Reality", blog post on wired.com. http://www.wired.com/2014/05/oculus-rift-4/?mbid=social_fb

[4].Mazuryk, Tomasz, and M. Gervautz. "Virtual reality-history, applications, technology and future.", 1996.

[5].VRML (Virtual Reality Markup Language) http://whatis.techtarget.com/definition/VRML-Virtual-Reality-Modeling-Language

[6].Webrift. http://wwwtyro.github.io/webrift/

[7] VR.js. https://github.com/benvanik/vr.js/tree/master

[8] "Why VRML Failed and What That Means for OpenOffice.", http://cafe.elharo.com/ui/why-vrmlfailed-and-what-that-means-for-openoffice/

[9] Kesmai, Wikipedia. http://en.wikipedia.org/wiki/Kesmai

[10] "Radisson Hotel Virtual Tour Statistics.", http://www.panorama.nu/ROI.pdf

[11] Oculus VR, http://www.oculusvr.com/

[12] Whiting, Nick. "Integrating the Oculus Rift into Unreal Engine 4", blog post on gamasutra.com. June 11th 2013.

[13]http://gamasutra.com/blogs/NickWhiting/20130611/194007/Integrating_the_Oculus_Rift_into_Unreal_Engine_4.php

[14] Brunner, Grant. "Oculus Rift's time warping feature will make VR easier on your stomach", blog post on extremetech.com

[15]http://www.extremetech.com/gaming/181093-oculusrifts-time-warping-feature-will-make-vr-easier-onyour-stomach

[16] http://www.pcmag.com/article2/0,2817,2455445,00.asp

[17] P. Baudisch, D. Tan, D. Steedly, and E. Rudolph, "Panoramic Viewfinder: Providing A Real-Time Preview To Help Users Avoid Flaws In Panoramic Pictures.", *Proceedings of OZCHI.*", 2005.

[18] S. Chen, "QuickTime VR-An Image Based Approach to Virtual Environment Navigation.", *ACM SIGGRAPH International Conference on Computer Graphics and Interactive Techniques*, 2005, pp.29-38.

[19] Harvard Virtual Tour. Retrieved April 1, 2009, from www.hno.harvard.edu/tour/

[20] H.W. Kang, S.H. Pyo, K.I. Anjyo, and S.Y. Shin, "Tour into the picture using a vanishing line and its extension to panoramic images.", *Computer Graphics Forum*, Blackwell Publishers Ltd, Vol.20(3), 2001, pp. 132-141.

[21] J. Nielsen, "Why You Only Need to Test With 5 Users.", Retrieved March 19, 2009, from www.useit.com/alertbox/20000319.html

[22] Z. Pan, "Easy Tour: A New Image Based Virtual Tour System.", *ACM SIGGRAPH International Conference, Session 8-3*, pp. 467-471.

[23] M. Roussou, "Learning by Doing and Learning Through Play: An Exploration of Interactivity in Virtual Environments for Children.", *ACM Computers in Entertainment*, Vol.2(1), pp.1-23
.

[24] N. Schmidt and J. Krone, "Constructing An Efficient And Easily Distributable Virtual Tour.", 2005.

Anomaly Detection Based Performance Improvement of Existing Business Intelligence System

Tsatsral Amarbayasgalan[1], Iderbaatar Munkhuu[2], Otgonnaran Ochirbat[3], Oyun-Erdene Namsrai[*]
[1,2,3,*)]*New Mongol Institute of Technology, Ulaanbaatar, Mongolia*
School of Engineering and Applied sciences,
National University of Mongolia, Ulaanbaatar, Mongolia
{[1]a_tsatsral, [2]alex8_1015}@yahoo.com, {[3]otgonnaran, [*]oyunerdene}@seas.num.edu.mn

Abstract

In anomaly detection, the goal is to find objects that are different from most other objects. Data must be without anomaly so that information can be true. It is possible to improve the quality of result eliminating anomaly data that influence on false information. For instance, anomaly detection is usually used in credit card fraud detection, telecommunication fraud detection, network intrusion detection, fault detection applications.

Previous research work, we proposed solution of how to implement business intelligence system that is to generate decision support information from big data efficiently and apply it into supermarket's business intelligence system. Under this solution, we have studied Hadoop data storage system, Hive data warehouse software, Sqoop data transmission tool and etc., successfully implemented them.

In this paper, we have implemented anomaly detection stage to improve the result of supermarket's business intelligence system based on the previous research work. Users can eliminate anomaly on data visualization before using business intelligence system. In other words, related products information from supermarket's business intelligence system has been improved by eliminating anomaly in the right way.

Keywords: data mining, anomaly detection, association rules mining.

1. Introduction

Since 1990, registration software was focused through the world and as a result of this; sufficient amount of data was collected. Now amount of data is growing rapidly. Also the size of raw data is big, but it is not enough valuable *[HYPERLINK "" \l "Tsa15" 1]*. Thus, instead of storing collected data as inactively, it is to predict the future by making variety of processing on it as well as using it used for decision-making, yet wrong prediction can be made from raw data.

Data from special situation, which is different from others, influences to generate wrong result. For example, we use data all citizens' income in order to estimate average income of "A" district. If there is the man who get one billion paid, average income of particular district can be high. So, it is necessary to eliminate anomaly from data before processing it.

In this paper, it is to provide the solution of approach and technologies suit for various business intelligence systems in accordance with the general architecture of system shown Figure 1 and then apply it into supermarket's business intelligence system. This architecture has advantages from the previous architecture by adding new anomaly detection stage between the data storage and the prediction stages.

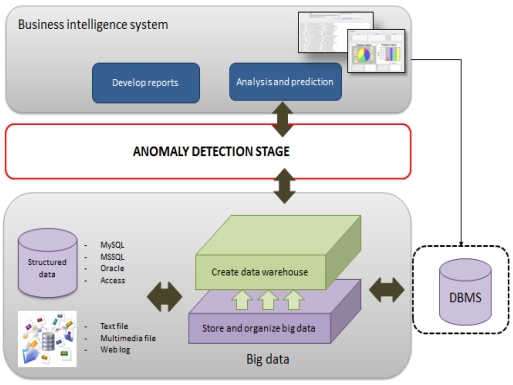


Figure 1. General design of improved architecture

By using a business intelligence system in supermarket, it is available to discover customer's purchasing pattern and relevant products. It is possible not to spend extra expenditure on the followings by discovering related products and goods:

- Shelf placement of products and goods,
- Promotion of related products,
- Product bundles by mixing related products,
- Particular product and suitable understanding of customer needs.

But, Anomaly data can be eliminated previously in order to discover related products. For example, wrong result can be calculated that fairytale books, which are not sold daily, but sold massively during the children day.

Novelty of the research work is to propose the approach, algorithm and technologies required for store big data, create data warehouse from it, eliminate outlier and extract knowledge from data warehouse and prepared data in a comprehensive manner.

2. Related work

We have studied what technology is used to store data for analysis in big organizations and anomaly detection approaches.

2.1. Facebook

Facebook used MySQL database for distribute data for analysis and used Python scripts that pinged stats back to a central MySQL database. The main problem with this setup was that historical analysis was difficult on data was spread over many machines and aggregating data to the analytical database was a slow, inefficient process. So Facebook built a 10 TB Oracle warehouse. But this solution would have been convenient for small and medium-sized organizations. Because impression logging was turned on which generated over 400 GB of data on the first day. Thus they got a Hadoop cluster to replace the data collection and processing tiers [HYPERLINK "" \l "Bro10" 2]. The principal is that it inserts a number of TB daily logs into HDFS file system. Also data in HDFS file system and MySQL server is integrated into Hive data warehouse and results of analysis on the Hive warehouse are stored back to MySQL and Oracle servers 3]. The main processing of big data is made on the Hive data warehouse and its results are transmitted into a simple database system for accessing many times. End users will get preprocessing data from MySQL and Oracle servers.

2.2. EBay

EBay is an international organization and e-commerce company. It has 120 million active users and 350 million products, performed 300 million searches per day. Thus they use Hadoop technology for store and process data such as user click, products, transaction, customers, feedback and auction. Research team built a 4 node cluster in 2007, 28 node cluster in 2009 and 532 node cluster in 2010 [HYPERLINK "" \l "www" 4]. The principal is like Facebook but one feature is access data easily by using OLAP tool.

2.3. Anomaly detection

Network vibrant research area applications use artificial intelligence, machine learning, and state machine modeling. Nowadays anomaly detection techniques are used it [HYPERLINK "" \l "Mar03" 5]. Our research main goal is to improve result of business intelligence system using anomaly detection method.

3. Proposed solution

The proposed solution has been entirely designed to use Hadoop and open-source technologies based on Hadoop. The advantage of Hadoop is that it can store both of structured and unstructured data. And open-source technologies based on Hadoop are being invented and they are making more flexible usage of it.

Firstly, data for report and prediction is collected from various data sources to HDFS file system and is loaded required data from it into data warehouse. Hive data warehouse query the data using a SQL like language called HiveQL. But taking summarized report from Hive data warehouse directly is slow (it takes a long time to develop each report due to processing big data). Thus, summarized results on Hadoop and Hive are copied into MySQL, and then user application gets reports from MySQL that spends less time.

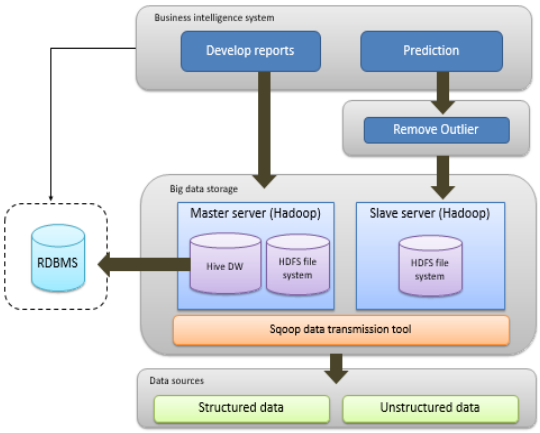


Figure 2. The solution of methods and technologies of implemented architecture

In addition, anomaly detection can be made on prediction data, but anomaly detection stage is not

needed for generating reports. As for our supermarket's business intelligence system anomaly data is risky to make prediction.

4. Implementation

We have implemented anomaly detection system based on previous supermarket's business intelligence system. In other words, we have performed implementation in according with proposed solution.

There are two types of processing have been performed on the supermarket's business intelligence system. First one is annual, quarterly, monthly reports depends on time are extracted by HiveQL from the Hive data warehouse. Second processing is to discover information that how products are sold together using data mining association analysis. The implementation steps of proposed solution:

1. Prepare server for store big data
2. Load supermarket's data into HDFS, build data warehouse and generate report
3. Eliminate anomaly /Anomaly detection system/
4. Discover relevant products from data in HDFS file system

We have considered step number three and four. In previous research work first, second, fourth steps have been implemented. So we have implemented anomaly detection system additionally.

4.1. Eliminate anomaly

We have detected outlier using standard deviation. The standard deviation is a measure of how widely values are dispersed from the average value.

$$\sqrt{\frac{\sum(x-\bar{x})^2}{(n-1)}} \qquad (1)$$

Anomaly detection steps:
k=1, i_k=k^{th} product

1. Calculate standard deviation(std)
2. Calculate average(avg)
3. Count i_k purchase each month(count)
4. If(std±avg)>count or (std±avg)<count
5. If result of step 4 is true, eliminate anomaly
6. Repeat step3

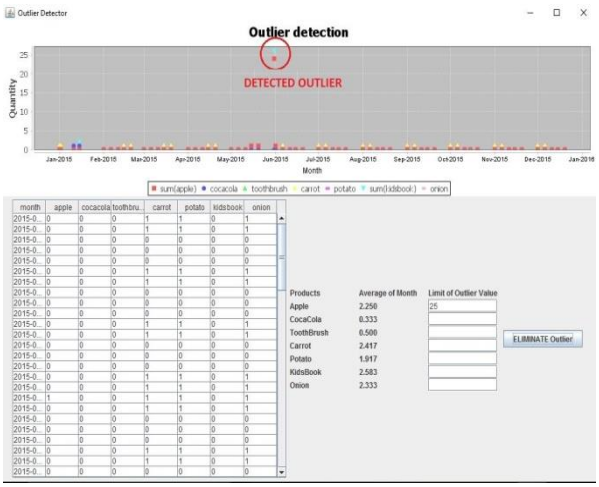


Figure 3. Anomaly detection system

5. Result

In this section, we present result of implemented anomaly detection system by proposed solution. Figure 5 shows general collaboration schema of anomaly detection system and previous business intelligence system.

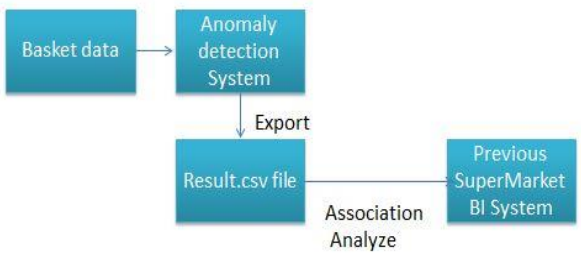


Figure 4. Collaboration schema of Anomaly detection system and Supermarket BI system

We use supermarket's data sources for testing this anomaly detection system. The data source includes 8 attributes (date, apple, coca cola, toothbrush, carrot, potato, kids book, onion).

month	apple	cocac...	toothbr...	carrot	potato	kidsbook	onion
2015-01-01	0	0	0	1	1	0	1
2015-01-02	0	0	0	1	1	0	1
2015-02-01	0	0	0	0	0	0	0
2015-02-06	0	0	0	0	0	0	0
2015-02-11	0	0	0	0	0	0	0
2015-02-15	0	0	0	1	1	0	1
2015-02-20	0	0	0	1	1	0	1

Figure 5. Data source

First we have discovered related products from two kinds of data sets which are anomaly removed and

anomaly not removed. Then these two results have been compared.

```
Best rules found:

 1. onion=T 28 ==> carrot=T 28    <conf:(1)> lift:(3.1)
 2. potato=T 23 ==> carrot=T 23    <conf:(1)> lift:(3.1)
 3. potato=T 23 ==> onion=T 23    <conf:(1)> lift:(3.21)
 4. potato=T onion=T 23 ==> carrot=T 23    <conf:(1)> li
 5. carrot=T potato=T 23 ==> onion=T 23    <conf:(1)> li
 6. potato=T 23 ==> carrot=T onion=T 23    <conf:(1)> li
 7. carrot=T 29 ==> onion=T 28    <conf:(0.97)> lift:(3.
 8. apple=T 27 ==> kidsbook=T 26    <conf:(0.96)> lift:(
```

Figure 6. Related products from anomaly not removed data set

Figure 6 shows related products from anomaly not removed data set. The confidence was configured as 90 percent and minSupport was configured as 0.01. But discovered relation of products is not compatible. Because according to result, rule of 96 percent confidence have been discovered that kid's book and apples are sold together. But this rule is valid only children's day and not valid daily.

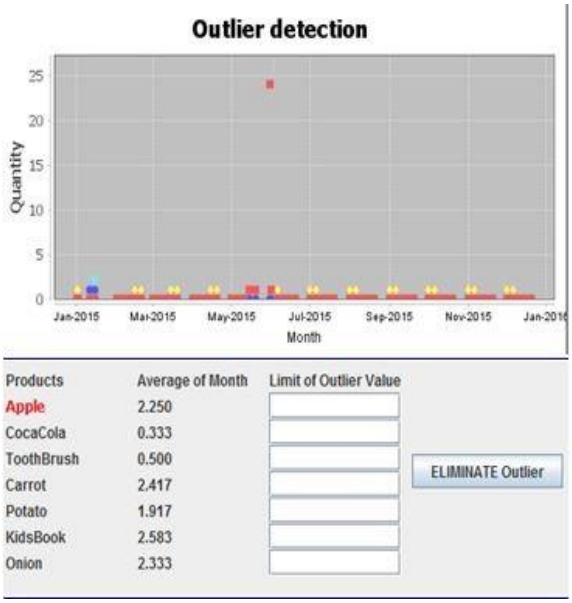


Figure 7. Result of anomaly detection

In Figure 7, the red box refers to apples purchase. Purchase of the apples was constant except May, but it increased rapidly in May due to Children's Day. So, data of the apples is anomaly in May.

```
Best rules found:

 1. onion=T 28 ==> carrot=T 28    <conf:(1)> lift:(3.1
 2. potato=T 23 ==> carrot=T 23    <conf:(1)> lift:(3.
 3. potato=T 23 ==> onion=T 23    <conf:(1)> lift:(3.2
 4. potato=T onion=T 23 ==> carrot=T 23    <conf:(1)>
 5. carrot=T potato=T 23 ==> onion=T 23    <conf:(1)>
 6. potato=T 23 ==> carrot=T onion=T 23    <conf:(1)>
 7. carrot=T 29 ==> onion=T 28    <conf:(0.97)> lift:(
```

Figure 8. Related products from anomaly removed data set

Figure 8 shows related products from anomaly removed data set. The confidence and MinSupport were configured same to previous processing. On this occasion, result is compatible. Because kid's book and apples are not related to each other. The discovered result has been improved by remove anomaly data.

6. Summary

In previous research work, we have proposed solution of how to implement business intelligence system that is to generate decision support information from big data efficiently and apply it into supermarket's business intelligence system. Under this solution, we have studied Hadoop data storage system, Hive data warehouse software, Sqoop data transmission tool and etc., successfully implemented them.

In this research work, we have improved architecture of previous research work and add anomaly detection system on supermarket business intelligence system.

We have tested two kinds of data set which are anomaly removed and not removed. Wrong result has been avoided by removing anomaly from data set.

One of the advantages of our implementation is cost saving, cause we have used open-source software and we proposed several thousand of simple computers for implementation, but its installation and configuration is more difficult than commercial software.

The one important thing in supermarket's business intelligence system is discovering relevant products and FPGrowth algorithm in Apache Mahout Library is used for discover relevant product. The advantage of this algorithm is available to works on distributed machines by parallel and processing data faster due to doing a few loop. Thus it is more compatible with Hadoop system.

7. References

[1] O. N. T. Amarbayasgalan, "The approach of implementing business intelligence system: possibility to analyze big data," in *FITAT 2015*, Jilin, China, 2015, p. 228.

[2] G. Brown. (2010, June) www.redfin.com. [Online]. HYPERLINK "https://www.redfin.com/blog/2010/06/evolving_a_new_analytical_platform_with_hadoop.html"

[3] N. Jain, Zheng Shao, Prasad Chakka and Ning Zhang, Facebook Data Infrastructure Team Ashish Thusoo, Joydeep Sen Sarma. Hive-A Petabyte scale data warehouse using Hadoop.

[4] www.slideshare.net. [Online]. HYPERLINK "http://www.slideshare.net/madananil/hadoop-at-ebay"

[5] M. T. C. Ji, "Anomaly Detection in IP Networks," *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, vol. 51, 2003, pp. 2191-2204.

Activity Recognition based on Clustering Methods for Senior Homecare Services

Thi Hong Nhan Vu[1], Yang Koo Lee[2], Oyun-Erdene Namsrai[3]
[1]*Human Machine Interaction Laboratory, Faculty of Information Technology, UET, Vietnam National University, Hanoi, Vietnam*
[2]*Positioning/Navigation Technology Research Section, Robot/Cognitive Convergence Research Division, IT Convergence Technology Research Laboratory, ETRI, Daejeon, Korea*
[3]*School of Engineering and Applied Sciences, National University of Mongolia*
[1]*vthnhan@vnu.edu.vn*, [2]yk_lee@etri.re.kr, [3]*oyunerdene@seas.num.edu.mn*

Abstract

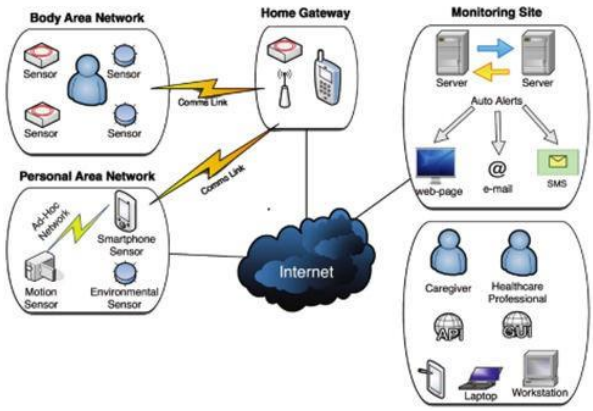
In modern society, most seniors prolong their independence. To guarantee the safety of them when living on their own, we need to monitor their activities all the time and react to critical situations. The rapid advances in wireless networks, wearable sensors, and communications technologies pave the way for the advent of homecare service systems. Activity recognition is a crucial task in building such systems. This paper investigates two clustering methods, k- means and Self-organizing map (SOM) for recognizing human daily activities. An experiment is performed on a real data set. The results show that k-means performs pretty well in classifying two activities; however the accuracy is pretty low when the data set is scaled up to five activities. SOM outperforms k-means in most cases of data sets. On average, the resulting accuracy of SOM is 87% and of k-means is 54% for five activities. As a result, SOM is most suitable to be integrated in systems for providing remote homecare services.

Keywords: activity recognition, clustering techniques, senior homecare services

1. Introduction

With high living standards the elderly populations of developed countries increase. Researchers estimate that the group of 65 year olds and over will account for 20% of the overall population [8]. Most of the seniors do not want to live at any eldercare facility but at their homes. To guarantee the safety of seniors with disability, homecare services should be available to assist them in emergency situations any time.

One of the solutions to building homecare service systems is to deploy Body Area Network (BAN) and Personal Area Network (PAN), which help monitor the seniors continuously. BAN consists of sensors worn by people and PAN consists of sensors embedded in the environment. Data collected by BAN and PAN transferred to monitoring sites via Home Gateway and Internet, which would later be analyzed by healthcare service providers (e.g., caregiver, healthcare professionals). Figure 1 illustrates a homecare service scenario.



.**Figure 1.** Homecare service scenario [8]

Activity recognition has been drawn a lot of attention lately in the field of remote care. The discovery of daily activities helps us find out the potential health problem. Thereby the hazardous conditions would be prevented.

An intelligent home care emergency system uses a neural network for fall detection. This system applies a tri-axial accelerometer and report the discovery of fall to an emergency center [1]. So far human behavior recognition has been received a lot of attentions especially in security and remote surveillance applications. Lately, an outdoor camera-based surveillance system was developed to estimate human postures [3]. Images captured by a camera are used to train a classifier for posture recognition. Manhattan distance is applied when training the clustering model. Labels of postures are done manually after the training phase. Even though the accuracy of the classifier is high and computation time is small, the deployment of camera in monitoring the seniors at home may raise privacy concern, so on-body sensors can be employed as an alternative for recognizing human activities. On-body sensors are multi-axis accelerometers.

In this paper, we investigate two clustering techniques, namely k-means in [6] and Self-organizing map (SOM) in [5], [7] to recognize human activities. An experiment is performed on a real data set to determine which is most suitable for the purpose of activity recognition. The results show that k-mean technique works well for distinguishing two activities (‘*walking’* and ‘*lying’*; *‘sitting’* and ‘*standing up from lying’*). However, when we scale up the data set to more than two different activities k-means gives low recognition accuracy. In contrast, SOM performs pretty well for recognizing even five activities (‘*walking’, ‘lying’, ‘sitting’, ‘standing up from lying’,* and *‘sitting on the ground’*) with average accuracy reaching up to 87%. This is because SOM is a type of neural networks, which has some advantages such as low sensitivity to noise. For recognizing five activities, k-means only obtains 54% on average for accuracy. As a result, SOM could be integrated in senior homecare systems for detecting falls or a loss of consciousness. Life quality of the seniors is thereby improved and life threatening is prevented as well.

2. Methods for activity recognition

The implementation of the health surveillance system in home-care settings involves the consideration of many variables that can be categorized into two types, namely sensing variables and classifier variables (see Figure 2)

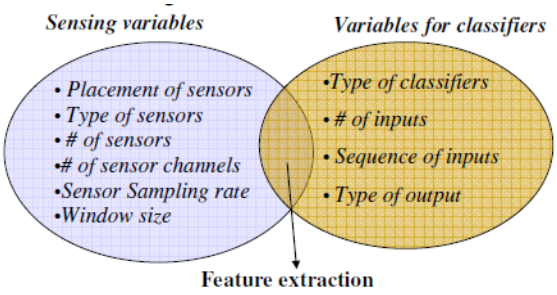


Figure 2. Different types of variables for classifiers .

2.1. K-means clustering method

K-means has been known one of the most easy toimplement clustering techniques so far [6]. Its algorithm considers flat representation of data as it bet suited when the data is known or partly known. A specific number *k* of clusters are determined by this algorithm according to the Euclidean value before clustering commences. By iterative batch training, the objective of its algorithm is to partition the total data set in to *k* clusters so that the following total intracluster variance is minimized:

$$I = \sum_{j}^{k} \sum_{i=1}^{N} a_{ji} \parallel X_i - c_j \parallel \qquad (1)$$

where N is the number of samples, and a_{ji}=1 when sample X_i belongs to the cluster j and a_{ji}=0 otherwise. As can be seen, it simply minimizes the sum of all quadratic distances of each sample to its associated cluster center, which is also called cluster prototype in some publications.

2.2. Self-organization map

One of the well-known and robust machine learning techniques for human activity recognition is neural networks.

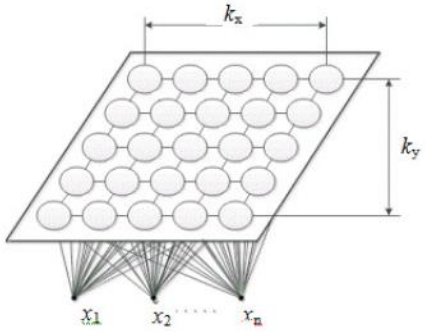


Figure 3. Structure of SOM

In this paper, we investigate SOM in comparison with k-means for analyzing the activities of the seniors at their home. SOM works by transforming the objects in a high dimensional space into a map. The advantage of SOM is that it preserves the topological properties of the input space. That means, if the objects are close in the input they will be close in the output space. This features makes SOM useful for visualizing lowdimensional views of high-dimensional data. SOM has two-layer structure. The input layer is represented by a vector for the input data. The output layer is represented by a map, which is consisted of a fixed number of neurons. The neurons are placed in a regular fashion in a grid. Associated with each neuron in the map is a weight vector W of the same dimension as the input vector. This vector W indicates the position of the neuron in the map space. Figure 3 illustrates the structure SOM with size k_x x k_y and the input vector $X=\{x_1,x_2\ldots,x_n\}$ for the input layer.

The algorithm of SOM works in two stages, namely learning and testing. The objective of learning in SOM is to cause different parts of SOM to respond similarly to certain input samples. SOM applies competitive learning.

In the learning stage, the input data samples are used to builds the maps. This is done by a competitive process. At each learning step, an input vector X from the data space is placed onto the map by finding the neuron whose weight vector is closest to the vector taken from the data space. The neuron whose weight vector is the most similar to the input vector is called the best matching neuron (BMN). The weight of BMN is then adapted towards the input vector. The magnitude of the change decreases with time and with distance from BMN. Given the input vector X and the weight vector w(t) of the BMN at time t is updated with the following formula:

$$W(t+\tilde{1}) = W(t) + \alpha(t)(X - W(t)) \qquad (2)$$

in which α =[0,1] is the monotonically decreasing learning rate. To achieve a topological mapping, the neighbors j of the winner neuron can adjust their prototype vector towards the input vector as well, but in a lesser degree, depending on how far away they are from the winner. Usually a radial symmetric Gaussian neighborhood function η is used for this purpose:

$$\forall j: W_j(t+1) = W_j(t-1) + \eta(t').\alpha(t).(X(t) - W_j(t)) \quad (3)$$

with the neighborhood function η decrease as the value of t'. When the training process ends, all the input vectors are mapped onto the weight vectors of neurons. In the testing set, the unknown data samples that are not used in the learning phase would be applied.

3. Experimental results

The data set used for evaluating the performance of k-means and SOM is published by [2]. Activities in the data set that need to be recognized include *walking, falling, lying down, lying, sitting down, sitting, standing up from lying, on all fours, sitting on the ground, standing up from sitting, standing up from sitting on the ground*.

The parameters with the default values used in the experiment are summarized in Table 1.

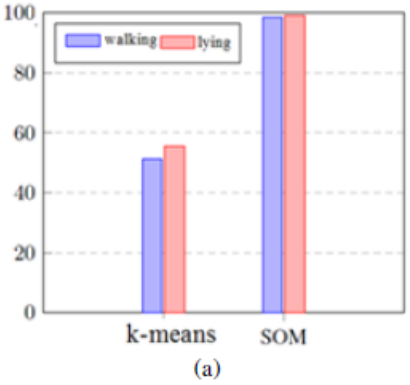
Table 1. Summary of parameters

Parameters	Meaning	Default value
k	Number of clusters	
$k_x \times k_y$	Size of SOM	30x30
α	Learning rate of SOM	0.3

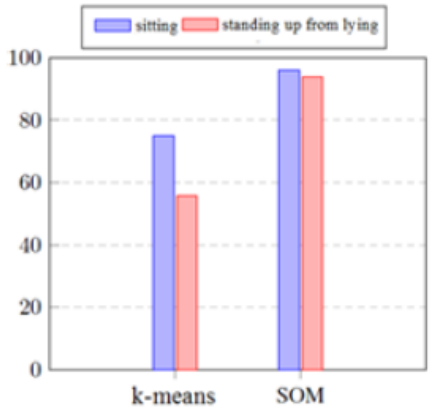
To evaluate the performance of the clustering techniques, we apply the recognition accuracy is calculated by the following formula:

$$accuracy = \frac{\#of\ correctly\ recognized\ activities}{total\ number\ of\ activities} \qquad (4)$$

The first test was carried out to recognized two pairs of activities, namely '*walking*' and '*lying'* in Figure 4(a) and '*sitting'* and '*standing up from lying*' in Figure 4(b). The results show that the accuracy of SOM can reach up to 99%. In contrast, that value of kmeans is not that high.



(a)



(b)

Figure 4. Recognizing a pair of activities

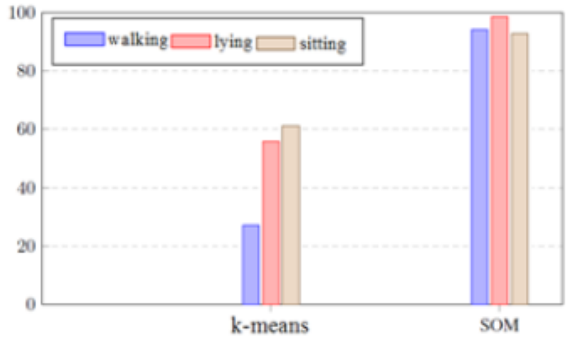


Figure 5. Recognizing three activities

The second test was performed on the data set with three activities (Figure 5). K-means got the lowest accuracy for '*walking"*. The same result as the first test was obtained with SOM. SOM can recognize three activities accurately and perform better than k-means.

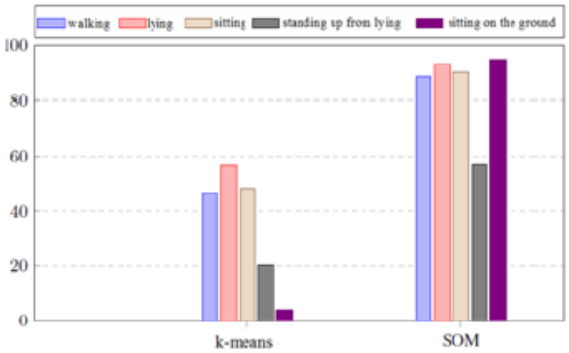


Figure 6. Recognizing five activities

In the third test, the data set was scaled up with five activities (Figure 6). The results show that SOM outperforms k-means, especially for recognizing '*sitting up from the ground'* SOM achieves 95% for its accuracy while k-means obtains only 3.9%. On average, the resulting accuracy of SOM is 87% and of k-means is 54% in this case. This is because data acquired from sensors always contain noise. As many other neural networks, SOM has low sensitivity with noise. On the contrary, k-means is sensitive to noisy data and outliers. In the next experiment we will apply F-measure in [4] to more different activities in consideration of spatial, temporal and context information to evaluate the performance of the clustering techniques.

4. Conclusions

Human activity recognition is an important task in remote surveillance applications, especially senior homecare services for seniors. The objective of activity is to determine the states of elder people via the analysis of body worn sensor data or camera.

In this paper, we carried out an experiment to evaluate the performance of two clustering techniques with different data sets collected from multi-axis accelerometers for different activities. The experimental results show that SOM achieved much higher resulting accuracy than k-means especially when the data set is scaled up with many different activities. In recognizing two activities, the accuracy of SOM could go up to 99% and in recognizing five activities, accuracy of SOM could achieve 87% on average while k-means only obtained 54%. This experiment promises that SOM can be applied to senior homecare service systems for detecting hazardous situation in short-term behavior such as falls or a loss of consciousness.

In the ongoing work, we would carry out the experiment for recognizing activities in real-time.

5. References

[1] J.I., Pan, C.Y., Yung, C.C., Liang, L.F., Lai, "An intelligent homecare emergency service system for elder falling", World Congress on Medical Physics and Biomedical Engineering,. Springer, 2006, pp. 424--428.

[2] B., Kaluza, V., Mirchevska, E., Dovgan, M., Lustrek, M., Gams, 'An Agent-based Approach to Care in Independent Living, International Joint Conference on Ambient Intelligence" (AmI-10), Malaga, Spain, In press.

[3] P., Spagnolo, M., Leo, A., Leone, G., Attolico, and A., Distante, "Posture Estimation in Visual Surveillance of Archaeological Sites", Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance, 2003.

[4] A. Dalli, "Adaptation of the F-measure to Cluster Based Lexicon Quality Evaluation", EACL 2003 Workshop

[5] R. O., Duda, P.E., Hart, D. G., Stork, "Pattern Classification", Wiley-Interscience; 2nd edn. 2000.

[6] L., Costa, and R. M., Cesar, "Shape Analysis and Classification: Theory and Practice", CRC Press, 2001

[7] Alpaydin, E., "Introduction to Machine learning",
References: Learning, The MIT Press, 2004

[8] D., Vouyioukas1 and A., Karagiannis, "Pervasive homecare monitoring technologies and applications", ISBN 978-953-307-354-5, June 20, 2011

Path Planning of Mobile Robot using Position System and Virtual Plane Approach in Dynamic Environment

Enkhtsogt.P[1] Khurelbaatar.Ts Ph.D[1], Zorig.B[2]
Department of Electronic,
[1]School of Information and Communications Technology,Mongolian University, Science and Technology, Mongolia
[2]Department of Electronics and Communication Engineering,School of Applied Science and Engineering,National University of Mongolia
[1]tsogtoopaul@gmail.com, hurlee77@yahoo.com, zorig@seas.num.edu.mn

Abstract

This paper presents fast measuring method using a sensor's fusion for path planning of mobile robot based on virtual plane approach in dynamic environment.

This sensor fusion approach is based on the position and orientation estimation using combination of web camera, encoder signals to improve and correct the measurement of moving object's position and velocity. We use the Kalman filter algorithm to combine sensors information. The performance of the proposed method is demonstrated by simulation results using experiment with moving obstacles.

Keywords: Path planning algorithm, service robot, robot designing and navigation, control law

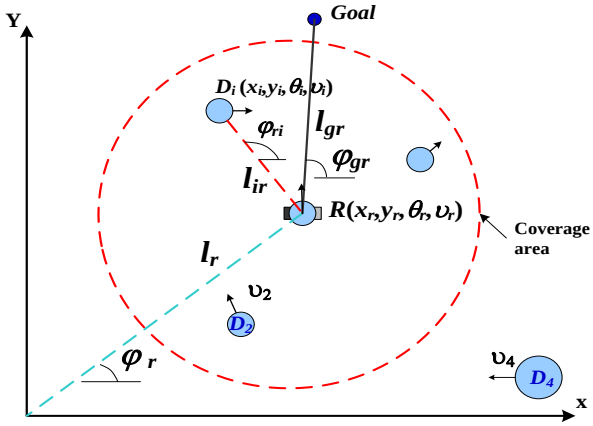
1. Introduction

Service robots which act in environments populated by humans have become very popular in the last few years. A variety of systems exists which act by using mobile robot in hospitals, office buildings, department stores, libraries and museums. Tasks of mobile robots include surface clearing, deliveries, helps and the exploration of unknown terrain [1]-[3]. This paper presents a path planning algorithm that is implemented for navigation of a wheeled mobile robot in a dynamic environment. Our used virtual plane approach is an invertible transformation equivalent to the workspace which is constructed by using a local observer [4]. Speed of the mobile robot and orientation angle are independently controlled using simple collision cones and collision windows constructed from the virtual plane. Therefore based on the virtual plane, it is possible to determine the intervals of the linear velocity and the paths that lead to collisions with moving obstacles.

To follow planned path which is established by above method we suggest controller design based on quaternion approach [5]. A mobile robot is defined as a moving, intelligent and autonomous vehicle. We made a control design for mobile robot with our proposed method.

2. Mobile Robot Kinematics

The Figure.1 shows the navigation for the mobile robot by the orientation and the speed. The line of sight of the robot l_r is the imaginary straight line that starts from *O* and it is directed toward the reference point of robot *R* and the line of sight angle ϕ_r which is the angle made by l_r.



Figureure. 1. Geometry of the navigation problem.

The line of sight of the robot l_r is the imaginary straight line that starts from the origin and is directed toward the reference center point of the robot R. The line-of-sight angle θ_r is the angle made by the sight

l_r. The distance l_{gr} between robot R and the goal G is calculated by

$$l_{gr}=\sqrt{(y_g - y_r)^2 + (x_g - x_r)^2} \quad (1)$$

Where (x_g, y_g) is the coordinates of the final goal point and $(y_r\ x_r)$ is the state of the robot in {W} . The mobile robot has a differential driving mechanism using two wheels and the kinematic equation of the wheeled mobile robot can be given by

$$x_r=v_r\cos\theta_r \quad (2)$$
$$y_r=v_r\sin\theta_r \quad (3)$$
$$v_r=a_r \quad (4)$$
$$\theta_r=w_r \quad (5)$$

where a_r is the robot's linear acceleration and v_r and w_r are the linear and angular velocities.(θ_r v_r) are the control inputs of the mobile robot. The line-of-sight angle φ_{ir} which is obtained from the angle made by the line of sight l_{gr} is given by the following equations:

$$cos\varphi_{ir} = \frac{[x_g - x_r]}{\sqrt{(x_g - x_r)^2 + (y_g - y_r)^2}} \quad (6)$$

In this paper we suggest virtual planning method which is developed by FethiBelkhouche [7] for path planning for service mobile robot in library.

This method is derived directly from the relative equations of motion of robot and dynamic obstacle.

The evolution of the range between the robot and obstacle for virtual planning method is given by following equation:

$$\dot{l}_{ir} = v_i\cos(\theta_i - \varphi_{ir}) - v_r\cos(\theta_r - \varphi_{ir}) \quad (7)$$
$$\dot{l}_{ir}\varphi_{ir} = v_i\sin(\theta_i - \varphi_{ir}) - v_r\sin(\theta_r - \varphi_{ir}) \quad (8)$$

Above equation shows the tangential component of the relative velocity. We can see the proof of the equations for the tangential and the normal components of the relative velocity from the reference [7].

A negative sign of $\dot{l}_{ir}$ indicates that the robot is approaching from obstacle *D.* If a zero rate, range implies constant distance between the robot and obstacle. The system presents a nice and simple model that allows real time representation of the relative motion between robot and obstacle

2.1 Mobile Obstacle Kinematics

The kinematic configurations of the mobile obstacle $D_i(x_i, y_i)$ are given by:

$$\dot{x}_i = v_i cos\theta_i \quad (9)$$
$$\dot{y}_i = v_i sin\theta_i \quad (10)$$
$$\dot{\theta}_i = \omega_i \quad (11)$$

The distance between the robot *R* and the obstacle *D is* given by:

$$l_{ir}=\sqrt{(y_i - y_r)^2 + (x_i - x_r)^2} \quad (12)$$

Where l_{ir} is distance between robot *R* and obstacle *D.*

The line-of-sight angle φ_{ir} is the angle which is made by the line of sight l_{ir} and it is given by

$$cos\varphi_{ir} = \frac{[D_x - R_x]}{\sqrt{(D_x - R_x)^2 + (D_y - R_y)^2}} \quad (13)$$
$$sin\varphi_{ir} = \frac{[D_y - R_y]}{\sqrt{(D_x - R_x)^2 + (D_y - R_y)^2}} \quad (14)$$

2.2 Navigation Process

Kinematic-based linear navigation laws are used to navigate the robot toward the final goal in [6]. A linear navigation law is given by:

$$\theta_r = M\gamma_{gr} + c_1 + c_0 e^{-at} \quad (15)$$

Where θ_r - direction of mobile robot, *M* - navigation parameter, γ_{gr}- angle of line of sight, c_0 andc_1 -direction terms, a - given constant for heading regulation

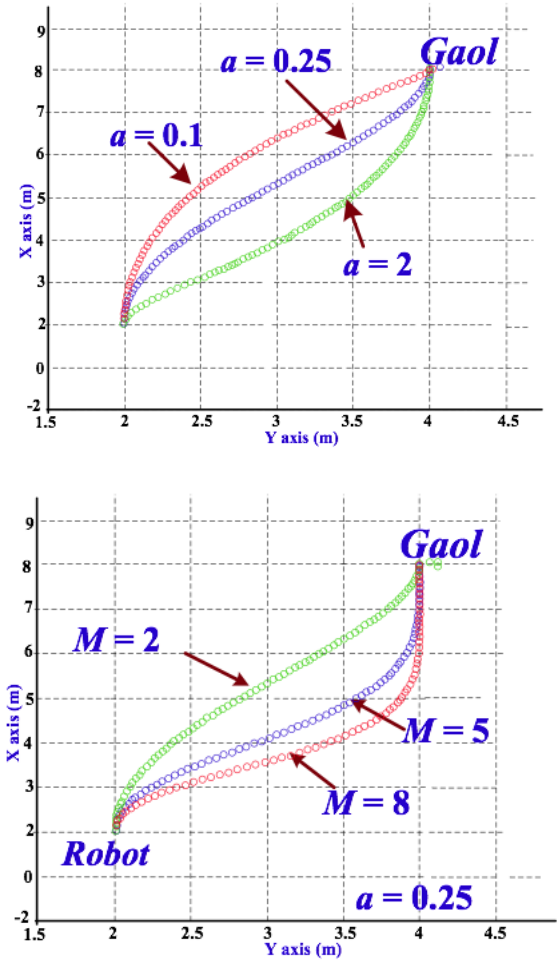
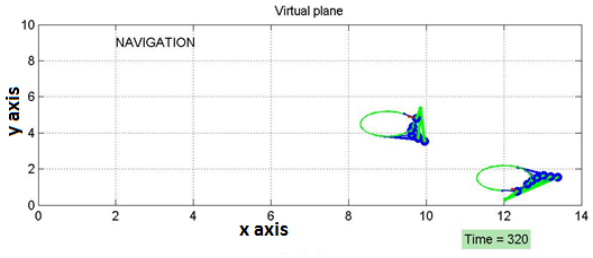
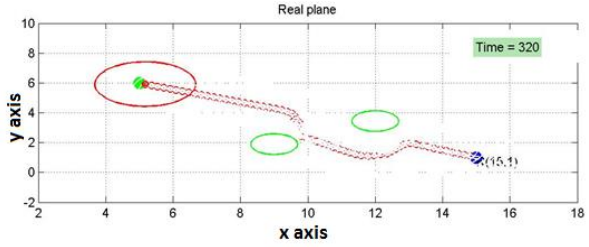


Figure 2. An illustration of the used approach, where the initial configuration is satisfied by the choice of the control law parameters.

Figure 2 shows simulation result of trajectory planning for the mobile robot motion from a start point to the final point. We can see changing navigation parameters gives us different trajectories to follow for mobile robot.



(a)



(b)

Figure 3. Comparison between collision avoidance using the robot-transformed approach versus obstacle-transformed approach.

(a) Real plane
(b) Virtual plane with robot transformed.

Figure 3 shows simulation results for instant path planning while robot moves in a dynamic environment with moving obstacles. From the Figure 4 we can see the mobile robot trajectory after avoiding from a moving obstacle while Figure 3 shows Virtual Plane illustration by considering moving obstacle as static.

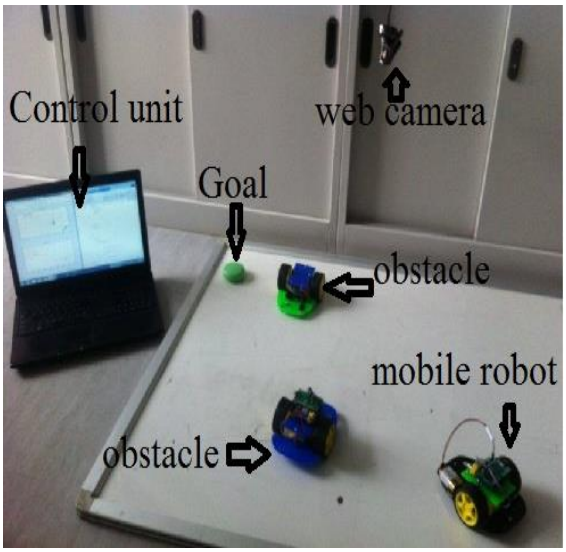


Figure 4. Implementation of control system

3. Implementation of the Kalman Filtering

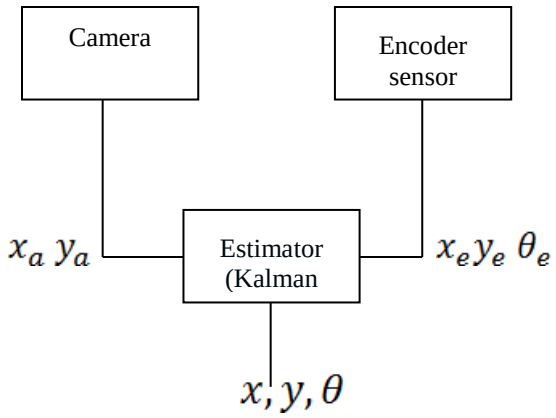


Figure 5. Kalman filter

Overview of Sensor Fusion System θ*e* is angle data and *xe*, *ye* are position data from encoder. After the one integration process for camera data, we get *xa*, *ya* are for the positions. From the position estimator, we get the new estimated position for the mobile robot after fusion of the camera and encoder's information through the Kalman filter, as *x*, *y*, θ.

Using the above error models, we design the indirect feedback Kalman filter[8], (Chui and Chen, 1990), as the state equations of the system.

$$\mathrm{x(k+1) = A(k)x(k) + w(k)} \qquad (10)$$

Where **v**(*k*)is measurement noise. It is assumed that **w**(*k*) and **v**(*k*) are zero-mean Gaussian white noise sequences. **A**(*k*) are system matrices.

4. Conclusion

In this paper we suggested the path planningalgorithm based on concept of virtual space for collision detect and avoid obstacles in dynamic environment. The notion of the virtual plane can be combined with various classical methods for path planning and navigation in dynamic environments. We also proposed quaternion approach for stability attitude of mobile robot and found the quaternion approach is better than rotational matrix approach. In near future, we will try to tackle for the robust control problem even in the presence of noise in sensor such as gyro, GPS etc.

Acknowledgement

This work was supported by project of School of Information and Communications Technology, Mongolian University of Science and Technology.

5. References

[1] S. Takeshi, T. Noriyuki and K. Ryosuke, "Development of an Intelligent Wheelchair with Visual Oral Motion.", 1*6th IEEE International Conference on Robot & Human Interactive Communication*, Jeju, Korea, 2007.

[2] K. Seiichiro and O. Kouhei, "*Semiautonomous Wheelchair Based on Quarry of Environmental Information.*", *IEEE Transactions on Industrial Elelectronics*, Vol.53(4), 2006.

[3] Z. Liu, J. Zhao, L. Zhang, G. Chen, and D. Li "Realization of Mobile Robot Trajectory Tracking Control Based on Interpolation.", *2009 IEEE Internatinal Symposium on Industrial Electronics*, Seoul, Korea, 2009.

[4] F. Belkhouche, T. Jin, and C.H. Sung "Plane collision course detection between moving objects by transformation of coordinates", *17th IEEE International Conference on Control Applications Part of 2008 IEEE Multi-conference on Systems and Control*, San Antonio, Texas, USA, 2008.

[5] W. Ham and T. Khurelbaatar "Computer Graphic Animation for Helicopter Control.", *2009 IEEE Internatinal Symposium on Industrial Electronics*, Seoul, Korea, 2009.

[6] FethiBelkhouche and BoumedieheBelkhouche " *Wheeled mobile robot navigation using proportional navigation"*Advanced Robotics, Vol. 21, No. 3–4, pp. 395–420 (2007)

[7] F. Belkhouche "Reactive Path Planning in a Dynamic Environment.", *IEEE Transactions On Robotics*, Vol.25(4), 2009.

[8] I. Zunaidi, N. Kato, Y. Nomura and H. Matsui "Positioning System for 4-Wheel Mobile Robot: Encoder, Gyro and Accelerometer Data Fusion with Error Model Method.", *CMU. Journal*, Vol. 5(1), 2006

Spatial Keyword Queries using Spark for Big Social Data

Pyoung Woo Yang, Kwang Woo Nam
Kunsan National University
{manner7979, kwnam}@kunsan.ac.kr

Abstract

In this paper, we describe a model for interactive spatial keyword queries for big social data on an Apache Spark system. The geo-tagged social and microblog data have been massively growing with increase of social network service users using smartphones. For such big geo-tagged social data, supporting interactive queries is one of very important functionalities for social analytics and interactive service applications. Spark is a distributed in-memory framework, and recently invented for supporting interactive processing. We propose a preliminary system design for interactive processing of spatial keyword queries. And, this paper shows how to query the big social data by spatial, keyword, and spatial queries.

Keywords: Spatial Keyword Queries, Apache Spark, Distributed System, Hadoop

1. Introduction

As more and more users use social network services on mobile internet, the volume of geo-tagged social and microblog data have been massively growing. A social network service provider stores this big geo-tagged social data into cluster computing system, and analyses them.

Various cluster computing frameworks support to analyze and query such big social data. During the last a decade, MapReduce have been a major framework for processing big data [1][2]. MapReduce is a disk-based framework, thus every intermediate result is stored into disk on iterative processes. Always disk storing strategy assure strong tolerance, but does not provide enough performance for interactive processing. Recently, Spark is an in-memory based framework, which supports Resilient Distributed Datasets (RDDs) to cache intermediate data [3][4]. In recent research [5], Spark is superior to MapReduce except sort algorithm.

In this paper, we introduce an initial design of framework and queries for spatial keyword processing. We extend RDD to support spatial keyword queries and spatial indexes, so called Spatial Keyword RDD(SKRDD).

This paper begins the related work with spatial big data framework. The architecture overview of our initial spatial keyword query framework is shown in section 3, and a model of SKRDD-based spatial keyword queries is described in section 4.

2. Related Work

Over the past a decade, many experimental systems are tried to support indexing and querying spatial data on a distributed Hadoop File System (HDFS) framework. Three notable recent systems are discussed in this section.

Hadoop-GIS[6] was implemented on MapReduce and can be used on Hive framework. For efficient query processing, Hadoop-GIS uses global grid partitioning and r*-tree local indexing.

SpatialHadoop[7] was constructed on MapReduce, and can be used on Pig framework. SpatialHadoop supports sampling-based global partitioning strategy and various spatial indexing including r-tree, quad-tree, and grid index.

GeoSpark[8] are built on Spark framework, and can be access by Spark SQL and other query framework. GeoSpark support Spatial RDD which includes r-tree indexing and grid partitioning.

Hadoop-GIS, SpatialHadoop, and GeoSpark are impressive and comprehensive research systems, which give brilliant key ideas for spatial big data researchers to develop a new system supporting spatial index and queries on cluster computing environment. Although these systems efficiently handle big spatial data and support spatial queries, spatial keyword queries are another dimension which can deals with textual data [9].

3. System Architecture

In Figure 1, we describe the architecture of spatial keyword spark framework. Spatial keyword spark consists of three layers, spatial keyword query API layer, SKRDD layer, and spatial keyword index layer.

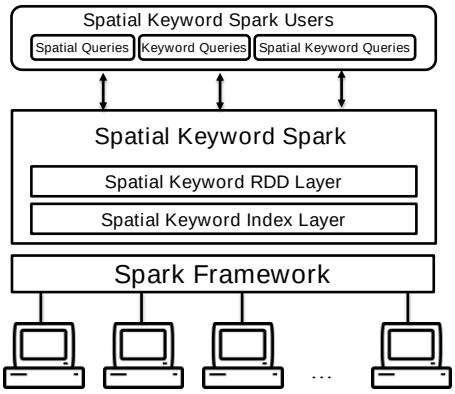


Figure 34. Architecture of Spatial Keyword Spark

Three spatial keyword spark layers are extended from Spark framework, and run on distributed cluster computing environment. We describe more details in the followings.

SKRDD Layer: General RDD does not spatial data type, and queries. Spatial keyword RDD supports spatial data type, spatial keyword data type, and optimized selection of RDD query path.

Spatial Keyword Index Layer: For efficient spatial keyword query processing, spatial keyword indexes can be constructed and called by queries API. This layer includes two kind indexes; a spatial index as like R-tree, and textual index as like Patricia tree.

Spatial Keyword Query API Layer: Users can retrieve and analyze spatial keyword data using spatial keyword query API. Spatial keyword query API includes spatial queries, spatial keyword queries, as well as construction of spatial keyword indexes. Also, using API, users can convert general textual spatial data CSV files into Well Known Text geometry files.

4. Spatial Keyword Queries on Spark

General spatial cluster framework just supports spatial queries, which includes spatial range query, spatial join queries, and k-NN queries. Spatial keyword spark supports not only spatial queries, but also spatial range keyword queries and spatial k-NN keyword queries. We propose a model for all kind queries including new ones.

$T'_{sk} \leftarrow q_s(T_{sk}, r)$

In spatial range query, T_{sk} is restricted by spatial range parameter r which is generally polygon or minimum bounding rectangle.

$T'_{sk} \leftarrow q_{sknn}(T_{sk}, l, k)$

Spatial k-NN query q_{sknn} needs point location l and the number of result rows k.

$T'_{sk} \leftarrow q_{sk}(T_{sk}, r, w)$

When users want to find out some spatial social data with specific keyword w within specific spatial range r.

$T'_{sk} \leftarrow q_{skknn}(T_{sk}, l, w, k)$

Also, they can use spatial keyword query q_{skknn} for retrieving nearest k social data including keyword w from specific location l.

Spatial keyword table T_{sk} can be accessed by spatial range query q_s, spatial k-NN query q_{sknn}, spatial keyword query q_{sk}, and spatial keyword k-NN query q_{skknn}.

5. Discussion and Future Work

In this paper, we proposed a new framework for interactive spatial keyword queries on distributed cluster computing environments. Also, we described the architecture of spatial keyword query framework.
We think our spatial keyword query framework is first challenge for integrating spatial and keyword functionalities.

This paper is a preliminary report for initial design of spatial keyword Spark. We will continue to develop the various indexes, and compare the performance on various and optimized functionalities in next phase.

Acknowledgement

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (2013-R1A1A4A-01013416).

7. References

[1] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Proceedings of OSDI, 2004, pp.137-150.
[2] S. Ghemawat, H. Gobioff, S. Leung, "The Google file system," Proceedings of SOSP, 2003, pp.29-43.

[3] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster Computing with Working Sets," Proceedings of HotCloud, 2010.

[4] M. Zaharia, M. Chowdhury, T. Das, A. Dave, Justin Ma, M. McCauly, M. J. Franklin, S. Shenker, and I. Stoica, "Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing," Proceedings of NSDI, 2012, pp.15-28.

[5] J. Shi, Y. Qiu, U. F. Minhas, L. Jiao, C. Wang, B. Reinwald, and F. Özcan, "Clash of the Titans: MapReduce vs. Spark for Large Scale Data Analytics", PVLDB, Vol.8(13), 2015, pp.2110-2121.

[6] A. Aji, F. Wang, H. Vo, R. Lee, Q. Liu, X. Zhang, and J. H. Saltz, "Hadoop-GIS: A High Performance Spatial Data Warehousing System over MapReduce.", PVLDB, Vol.6(11), 2013, pp.1009-1020.

[7] A. Eldawy and M. F. Mokbel, "SpatialHadoop: A MapReduce framework for spatial data.", Proceedings of ICDE, 2015, pp.1352-1363.

[8] J. Yu, J. Wu, and M. Sarwat, "GeoSpark: a cluster computing framework for processing large-scale spatial data.", Proceedings of SIGSPATIAL/GIS, 2012, pp.70.

[9] X. Cao, G. Cong, C. S. Jensen, and B. Chin Ooi, "Collective spatial keyword querying.", Proceedings of SIGMOD, 2011, pp.373-384.